Artificial Intelligence, Simulation and Society

Petra Ahrweiler *Editor*

Participatory Artificial Intelligence in Public Social Services

From Bias to Fairness in Assessing Beneficiaries

OPEN ACCESS

Springer

Artificial Intelligence, Simulation and Society

This book series brings into its fold key and emerging topics on the interactions between growing artificial intelligence technologies and their social impacts. It addresses various aspects of the relationship between AI, simulation, and society and provides insights into their intersections and stimulates discussions on the opportunities and challenges they present. The series is multi- and transdisciplinary in scope, and dynamic. It invites academic contributed volumes and monographs, but also more popular work suitable for lay readership, and innovatively includes some science fiction to initiate readers into the scope and aims of this novel series.

The specific themes and topics covered under the series are:

- **The ethical and societal implications of AI**: The series delves into the ethical considerations and societal impacts of AI technologies. It explores topics such as privacy, bias, job displacement, and the role of AI in shaping social structures from a social science point of view (sociological, political, economic, cultural, legal).
- **Simulation and modeling of social systems**: The series explores how simulation techniques are used to model and understand complex social systems and create artificial societies in silico. It covers topics such as social network analysis, agent-based modelling (ABM), and the simulation of collective behaviour.
- **AI and social simulation**: The series explores how AI technologies are used in social simulation, for example, modelling intelligent agents in agent architectures of ABM, or calibrating and validating models using intelligent data mining and analysis techniques.
- **AI and simulation in social philosophy**: It looks at how AI and simulation are depicted in social philosophy, for example, the role of AI and simulation in socio-technical evolution, the position of AI and simulation in Western rationalism, philosophical counter-designs of current developments, ontological and epistemological limitations and barriers of AI and simulation.
- **AI, simulation and society in fiction**: The series also innovatively examines the portrayal of AI and simulation in and as fiction, demonstrating how these themes reflect societal fears, aspirations, and ethical dilemmas. The series contains both original fiction and second-order analyses.
- **AI and simulation in entertainment:**It covers simulation techniques, combined with AI, that are used to create virtual worlds and characters that mimic human behaviour. Such simulations are used, for example, in video games, virtual reality experiences, and entertainment applications.
- **AI and simulation in various disciplines**: The series discusses the applications of AI and simulations that are/will be transforming various disciplines and domains such as healthcare (e.g. in medical diagnosis, drug discovery, and patient care), work (e.g. automation, Industry 4.0, workforce dynamics), or education (e.g. virtual reality, personalised learning systems, intelligent tutoring systems). It discusses the potential benefits and challenges of integrating these technologies into the conventional space.
- **AI, simulation, and policy**: The series analyses how AI and simulation techniques can inform the policy cycles. It discusses the use of predictive modelling, analysis of what-if scenarios, and decision support systems in shaping policies in various policy domains such as public policy, technology policy or environmental policy.

Petra Ahrweiler
Editor

Participatory Artificial Intelligence in Public Social Services

From Bias to Fairness in Assessing Beneficiaries

Springer

Editor
Petra Ahrweiler
Department of Technology- and Innovation
Sociology Social Simulation (TISSS Lab)
Institute of Sociology
Johannes Gutenberg-University Mainz
Mainz, Germany

ISSN 3004-9822 ISSN 3004-9830 (electronic)
Artificial Intelligence, Simulation and Society
ISBN 978-3-031-71677-5 ISBN 978-3-031-71678-2 (eBook)
https://doi.org/10.1007/978-3-031-71678-2

This work was supported by German VolkswagenStiftung (98560)

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book *Artificial Intelligence for Assessing People in Public Social Services: From Bias to Fairness: Towards Contextualised, Culture-Sensitive, Responsive and Participatory AI* results from interdisciplinary research across the globe.

By trying to *close the gap between expert-driven technology development and society,* social scientists have worked with computer scientists to find out about systems of artificial intelligence (AI) used for selecting beneficiaries from non-beneficiaries in welfare provision decisions, a process called 'social assessment'. Research used participatory methods that involved stakeholders from policy, industry, law and many other societal domains including civil society, and here especially vulnerable groups.

Values matter! Social services are supposed to ensure the fairness of life chances: They are there to alleviate, balance and compensate for difficult life circumstances, to mediate inequalities and to remedy injustices. Social justice decisions are value decisions about who deserves what and why. In AI-based social assessment for public service provisions, technology has to deal with potential recipients in a dynamic societal value framework distinguishing between legal/fraudulent, deserving/non-deserving, needy/non-needy, high-performing/low-performing, and desirable/non-desirable.

Context matters! It is very important where this assessment takes place. Criteria sets for assessment are highly contingent across the globe. Bias that privileges certain groups while discriminating against others is dependent on cultural context. In AI-based social assessment for public service provisions, technology has to deal with the role of culture and context and be sensitive to the huge heterogeneity of worldwide fairness concepts.

The book presents results from nine empirical case studies for cultural comparison that collected data on social assessment practices and respective AI use in the welfare systems of Spain, Estonia, Germany, Iran, India, Nigeria, Ukraine, China, and the USA.

This is the first of two volumes authored by the case study partners from those countries and led by the social scientists. Chapters present participatory empirical social research on status, advantages and problems of AI use for social assessment.

The second volume is about participatory modelling and simulation for 'Better AI' in case study countries with computer scientists in the leading role. The two scientific publications are embedded in literary fiction where a non-scientific readership can get acquainted with the topics of research.

Research stems from the international project "Artificial Intelligence for Assessment" (AI FORA) that started in the middle of the COVID-19 crisis and ended during the Russian war against Ukraine, one of our case study partners. The editor is very much indebted to the colleagues who presented their research in this volume dealing with all challenges and difficulties of a large interdisciplinary and international research project.

She is especially grateful for the many constructive and sometimes controversial discussions with fantastic academic colleagues and project partners that helped to shape research in this book: Among others, Nigel Gilbert, Martha Bicket, Sumathi Srinivasalu, Ebin Deni Raj, Albert Sabater, Beatriz Lopez, Roger del Campepadros, Emmanuel Ejim-Eze, Hassan Bashiri, Hui Li, Erik Johnston, Margaret Hinrichs, Chelsea Dickson, Steven Popper, Elina Treyger, Jirka Taylor, Jesús Siqueiros Garcia, Martin Neumann, Elisabeth Spaeth, Blanca Luque Capellas, David Wurster, Jennifer Abe, Gerhard Kruip, George Kampis, Elisabeth André, Ruben Schlagowski, Katharina Weitz, Tome Sandevski, Aoibheann Gibbons, Markus Knauff, Iris Lorscheid, Zsolt Juranyi, Massimo Rusconi, Hana Fehrenbach, Andrew Chan, Dario Brockschmidt and Christian Haenel.

Special thanks for great contributions to the deep questions of the book, open-mindedness and generous hospitality go to the 'Safe Spacers', i.e. the monastic leads and coordinators of the project's Safe Spaces for technology co-design with vulnerable groups, M. Maire Hickey OSB, Sr. Josephine Yator OSB, Sr. Jeanne Bott OSB, Fr. Cyprian Consiglio OSB Cam, Fr. Dorathick OSB Cam, P. Manel Gasch OSB, Br. Johannes Tebbe OSB, Br. Elija Pott OSB, Br. Josef van Scharrel OSB, Br. Simon Griskiewitz OSB and P. Jeremias Marseille OSB Cam together with their communities.

Acknowledgements All research presented in the volume has been funded by the German VolkswagenStiftung under grant agreement number 98 560. The editor also gratefully acknowledges additional funding for open access publication by the same funder, and matching open access funding by Johannes Gutenberg University, Mainz, Germany.

Mainz, Germany Petra Ahrweiler

Contents

Contributors

Jennifer Abe Psychology Applied Research Center (PARC), Bellarmine College of Liberal Arts, Los Angeles, CA, USA

Petra Ahrweiler Department of Technology- and Innovation Sociology, Social Simulation (TISSS Lab), Institute of Sociology, Johannes Gutenberg-University Mainz, Mainz, Germany

Hassan Bashiri Department of Computer Science, Hamedan University of Technology, Hamedan, Iran

Roger Campdepadrós Facultat de Ciències Econòmiques i Empresarials, Departament d'Empresa, Edifici Econòmiques, C/ de la Universitat de Girona, Girona, Spain

Blanca Luque Capellas Department of Technology- and Innovation Sociology, Social Simulation (TISSS Lab), Institute of Sociology, Johannes Gutenberg-University Mainz, Mainz, Germany

Emmanuel Ejim-Eze Department of Science Policy and Innovation Studies, National Centre for Technology Management, Head Quarters, Obafemi Awolowo University, Ile-Ife, Nigeria

Margaret Hinrichs Global Futures Laboratory, School of Complex Adaptive Systems, College of Global Futures, Arizona State University, Tempe, AZ, USA

Li Hui Shanghai Institute for Science of Science (SISS), Shanghai, PR China

Ashly Ann Jo Department of Anthropology, University of Madras, Chepauk, Chennai, India

Erik W. Johnston Global Futures Laboratory, School of Complex Adaptive Systems, College of Global Futures, Arizona State University, Tempe, AZ, USA

Oleksandr Khyzhniak Centre for International Cooperation, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Gerhard Kruip Department of Social Ethics, Faculty of Catholic Theology, Johannes Gutenberg-University Mainz, Mainz, Germany

Beatriz López Exit Grup, University of Girona, Carrer Universitat de Girona, Girona, Spain

Maris Männiste Faculty of Social Sciences, Institute of Social Studies, University of Tartu, Tartu, Estonia

Muniraj Mathaiyan Department of Anthropology, University of Madras, Chepauk, Chennai, India

Martin Neumann Faculty of Humanities, Department of Language, Culture, History and Communication, University of Southern Denmark, Slagelse, Denmark

Gunanithi Perumal Department of Anthropology, University of Madras, Chepauk, Chennai, India

Wu Qi Shanghai Institute for Science of Science (SISS), Shanghai, PR China

Ebin Deni Raj Indian Institute of Information Technology, Kottayam, Kerala, India

Caity Roe Global Futures Laboratory, School of Complex Adaptive Systems, College of Global Futures, Arizona State University, Tempe, AZ, USA

Albert Sabater Facultat de Ciències Econòmiques i Empresarials, Departament d'Empresa, Edifici Econòmiques, C/ de la Universitat de Girona, Girona, Spain

Cristina Sánchez Facultat de Ciències Econòmiques i Empresarials, Departament d'Empresa, Edifici Econòmiques, C/ de la Universitat de Girona, Girona, Spain

Manjubarkavi Selladurai Department of Anthropology, University of Madras, Chepauk, Chennai, India

Shobana Sharma Department of Anthropology, University of Madras, Chepauk, Chennai, India

Jesús M. Siqueiros García IIMAS Unidad Mérida, Universidad Nacional Autónoma de México, Ucú, Yucatán, México

Mihkel Solvak Faculty of Social Sciences, Johan Skytte Institute of Political Studies, University of Tartu, Tartu, Estonia

Elisabeth Späth TISSS Lab, Institute of Sociology, Johannes Gutenberg-University Mainz, Mainz, Germany

Sumathi Srinivasalu Department of Anthropology, University of Madras, Chepauk, Chennai, India

Avo Trumm Faculty of Social Sciences, Institute of Social Studies, University of Tartu, Tartu, Estonia

Triin Vihalemm Faculty of Social Sciences, Institute of Social Studies, University of Tartu, Tartu, Estonia

Jieshu Wang Global Futures Laboratory, School of Complex Adaptive Systems, College of Global Futures, Arizona State University, Tempe, AZ, USA

David Wurster Department of Technology- and Innovation Sociology, Social Simulation (TISSS Lab), Institute of Sociology, Johannes Gutenberg-University Mainz, Mainz, Germany

Chapter 1
Using a Case Study Approach for Investigating the Status Quo and Future Options of AI-Based Social Assessment in Public Service Provision

Petra Ahrweiler, Jennifer Abe, and Martin Neumann

Abstract The chapter features the research project 'Artificial Intelligence for Assessment' (AI FORA). AI FORA's results will be presented in two volumes where this is the first one on the project's empirical research. After a general introduction to the project, its topic, and its approach, two material sections follow, because their topics are central for AI FORA's work. Section "The Pervasive Practice of Assessment and AI" will discuss the pervasive practice of social assessment in our societies which is more and more delegated to AI. Section "The Role of Culture and Context" will present existing cultural comparison approaches and evaluate their capacity to address the role of culture and context for AI-based social assessment in social service provision. Finally, the chapter will introduce the contributions of this book: Each chapter describes a unique cultural representation of context-specific social assessment practices to use AI for public social service provision in different national welfare systems.

P. Ahrweiler (✉)
Department of Technology- and Innovation Sociology, Social Simulation (TISSS Lab), Institute of Sociology, Johannes Gutenberg-University Mainz, Mainz, Germany
e-mail: Petra.ahrweiler@uni-mainz.de

J. Abe
Psychology Applied Research Center (PARC), Bellarmine College of Liberal Arts, Los Angeles, CA, USA
e-mail: jennifer.abe@lmu.edu

M. Neumann
Faculty of Humanities, Department of Language, Culture, History and Communication, University of Southern Denmark, Slagelse, Denmark
e-mail: martneum@sdu.dk

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_1

Artificial Intelligence for Assessment: AI Assessing People—People Assessing AI

Worldwide, national welfare systems are challenged by scarce public resources, increasing citizen demands for state support, and growing population sizes. Public social services address people's vital needs from cradle to grave, trying to alleviate poverty and inequalities and ensure fair living conditions. Currently, urgent problems and pressures in social service provision are profoundly challenging the survival of poor individuals, marginalised minorities and vulnerable populations, and the stability of many countries due to scarce public resources, economic and financial crises, health threats, and inequalities of life chances in different regions. Most people use social services at some point in their lives (Corlet Walker et al., 2021). How to ensure a fair distribution of taxpayers' money is therefore a recurrent policy issue that depends on a society's ideas of social justice and fairness.

While social services can counteract discriminatory practices thus creating equal opportunities, the selection mechanisms of social assessment that decide on how and to whom these services are provided and distributed are *per se* discriminatory. Since public social service provision is about distributing scarce resources, not every applicant can get everything: Decisions to allocate services to applicants with certain attributes will necessarily imply the consequence of not having it allocated to others.

This makes the practice of social service provision at the same time anti-discriminatory and discriminatory with a variable trade-off between these two. On the discriminatory side, service recipients must fulfil certain criteria to be considered as beneficiaries. There are winners benefiting from discriminatory practices, and there are those who will suffer most from them and potential system failure. The question of social assessment, i.e., who gets what from the state, concerns everybody, whether a policymaker hoping for efficiency and objectivity in allocation, a recipient hoping for support and wellbeing, a service provider, a taxpayer, or a member of a vulnerable group.

Criteria sets, however, are highly contingent across the globe (Alesina & Angeletos, 2005; Fleurbaey, 2008; Taylor-Gooby & Martin, 2010). Though bias due to value judgements concerning fair distribution is prevalent in every selection system because they all work with inclusion/exclusion dualities as explained above, the question where this bias that privileges certain groups while discriminating against others is exactly located is answered differently across the globe dependent on cultural context: Who is considered as eligible, needy, and deserving to be a beneficiary is subject to cultural interpretation, social change, and multi-actor negotiation.

In many countries, public administrations are increasingly using Artificial Intelligence (AI) technologies to decide on the provision of public social services such as unemployment benefits, pension entitlements, kindergarten places, and social assistance to their citizens, hoping to achieve greater efficiency and objectivity (Angwin et al., 2022; Eubanks, 2018). This can concern both present and future, i.e. assessing whether a certain individual, e.g. an applicant for unemployment

benefits, is currently eligible for the service applied for, or assessing risks such as the probability that a certain individual will be likely to get into a situation of applying for a certain social service in the future. In the first case, assessment will lead to providing or not providing the social service in question, e.g. certain unemployment benefits; in the second case, it might lead to prevention and intervention offers to reduce risks of future service needs, e.g. in the case of unemployment suggesting prophylactic education and training measures. However, justifying preemptive intervention on the individual level with statistical predictive profiling based on correlations is difficult: Every individual can belong to the statistical minority group, and any intervention decision can belong to the minority of wrong decisions.

For both purposes, data profiles of citizens are analysed and assessed, and profiles automatically checked and scored. This is to determine whether their owners are eligible to receive support from the state, what type and quantity of service to allocate, whether individuals have a risk of getting needy in the future, to detect existing fraud and risk groups for potential future fraud (Dubois et al., 2018) and other applications. Whether the introduction of AI into social assessment makes things better or worse is of interest to everyone and makes everyone a potential stakeholder in determining the design of social assessment innovations.

Technology production is challenged to improve 'bad AI'. AI-based social assessment systems, because they are based on machine learning from historical data, are accused of perpetuating old and introducing new bias and discrimination, often to the detriment of the most vulnerable groups in society. Although types and degrees of AI implementation vary between countries, delegating decisions about the distribution of scarce resources to machines leads everywhere to important questions about ethics, justice, quality, responsibility, accountability, and transparency of welfare decisions triggering reflection about the values behind them. The public discourse on AI-based social assessment for public service provisions often is highly controversial and emotional: Societal core values are affected and at stake.

To project, design, and implement 'better AI' implies data-driven scenario analysis of 'techno-futures' (Grunwald, 2013) though the appraisal of potential social futures is a huge research challenge connected to complexity. Realistic models based on the data and insights gained by research presented in this volume are required that can represent the interaction dynamics between society and technology for anticipating, monitoring, evaluating, and improving the societal impact of AI assessment and prognosis technologies applied to human behaviour. Modelling has to consider values and context, and it needs to be participatory and inclusive: Scenario simulations modelling future societies can test and experiment with AI in use and desired options.

The AI FORA project[1] (AI FORA is the acronym of '**A**rtificial **I**ntelligence **for** **A**ssessment') investigates the issues above providing case-specific answers and solutions to the question how contextualised, value-sensitive, responsive, and

[1] Research project funded by German Volkswagen Foundation 2021–2024 under grant agreement number 98 560.

dynamic AI systems can be co-designed starting from existing systems that are perceived as problematic. Research needs a participatory reconstruction and review of existing systems to allow for a participatory anticipation, projection, and realisation of the desired systems. The AI FORA project is following this sequence for a range of case study welfare systems, chosen to maximise their heterogeneity. Not only do societies differ in the values and norms they facilitate, foster, and propagate, but also political, economic, and societal dependencies and pressures vary between societies. Perceptions, attitudes, discussions, and acceptance of AI use for public policy vary between these countries, as do the types and degrees of AI implementation, with reference to norms and values in use, but also related to technology status, economic models, civil society sentiments, and legislative, executive, and judicial characteristics.

Research is often only available to experts, shaped by their interests, and framed by their language (Humm & Schrögel, 2020). This acts as a threshold particularly excluding the general public, especially those people particularly affected, i.e. in need of better ways of social benefit provision. Thus, AI FORA features inclusive formats that are participatory and accessible to reach its target audiences.

This book on AI FORA's case studies was preceded by a novel published in the same series called 'Angels and other Cows' (Ahrweiler, 2024), which blends genres such as sci-fi, romance, adventure, mystery, and comedy. The novel introduces the research topic making available issues of AI use in the public sector to a broad readership and attracting also non-scientists to academic research. The two scientific books on empirical results and results of modelling will be embedded in dissemination to the general public: Final results and the impact of AI FORA research will be evaluated by a second novel concluding its dedicated strategy of inclusive science communication (Conceicao et al., 2020; Gascoigne et al., 2020; Metcalfe et al., 2022).

Results are particularly relevant for decision-makers. In AI FORA, the policy community, i.e. people working in all phases of the policy cycle (problem definition, policy formation, policy development, policy implementation, policy enforcement, policy evaluation) is both, participant and client of the research and co-design process. All AI FORA case studies actively involve national and local policy makers, policy analysts, and local decision-makers as well as staff in public administration to co-create the research and development process from problem definition via data collection and analysis up to modelling and scenario evaluation. Special formats have been created to communicate science successfully to policy makers, policy analysts, and public administrators building on prior research about target-specific requirements (Ahrweiler et al., 2019; Gilbert et al., 2018).

Results of the modelling and simulation work of AI FORA will be published in the second volume of this series (see for an outlook Chap. 13). The project applies a complex networks approach of social innovation to technology co-design using new participative formats and methodologies (see Chap. 2). This, for example, addresses the main challenge for inclusion: Creating a space based on trust that enables stakeholders to communicate without being preconfigured, discriminated, or constrained by the environment in which their encounters take place. AI FORA

works with the so-called 'Safe Spaces' concept as a specific approach to stakeholder communication in technology co-design. Furthermore, an 'Ethical Observatory' (see Chap. 3) has been developed to continuously reflect on the normative aspects of AI FORA research and on issues of ethical research design.

Context, especially in terms of the cultural and social embeddedness of values and attitudes, is key to understanding the use of AI to perform practices of social assessment. The case studies presented in this volume report on the empirical social assessment practices in their countries with different value reference frameworks, investigating the context-dependent service provision selectivity of their welfare system in order to analyse the requirements, constraints, and consequences of the use of AI to support or replace conventional practices.

The country case study chapters of this volume are embedded in an introductory section consisting of three conceptual chapters: On technology-based social assessment and cultural comparison (see this chapter with its following sections), on inclusive technology co-design and AI governance (see the next chapter), and on normative as well as ethical issues (Chap. 3). A concluding chapter summarises the book's results and provides an outlook to the next volume.

The remaining chapter is organised as follows: After an introduction to the pervasive practice of social assessment in our societies which is more and more delegated to AI, the chapter will discuss cultural comparison approaches concerning their capacity to address the role of culture and context for AI-based social assessment in social service provision. Finally, the chapter will introduce the contributions of this book: Each chapter describes a unique cultural representation of context-specific social assessment practices to use (AI) technologies for public social service provision in welfare systems.

The Pervasive Practice of Assessment and AI

The practice of assessment is socially pervasive and constitutive (Bowker & Star, 2000): The fundamental importance of valuation and evaluation for sociality in general is an essential component of the constitution of social order (Krueger & Reinhart, 2017; Cefaï et al., 2015; Muniesa & Helgesson, 2013). The practice of evaluation in epistemic cultures (Knorr-Cetina, 2002) is closely linked to typologisations and categorisations (Lamont, 2012), in which technological artefacts, intended or unintended, also function as differentiating entities: 'In modern society, technology is a powerful code that shapes, celebrates and legitimises the patterns of interpretation and evaluation of modern man' (Hoerning, 1989: 100; *own translation*).

It is a *conditio humana* that humans categorise the social and physical world to sustain their existence and even a universal principle of biological existence that cognition is shaped by what Berthoz denotes as 'simplexity' (Berthoz, 2012), that is: reducing the world's complexity to a few numbers of categories which allow for orientation and action in the world. In human societies, Fourcade (2016)

differentiates between judgements *nominal* (oriented to essence), *cardinal* (oriented to quantities), and *ordinal* (oriented to relative positions) classificatory judgements. Thus, categorising and, in consequence, social assessment is ubiquitous and inevitable. With the spread of neo-liberal market fundamentalism, the practice of quantitative, numerical evaluation has gained increasing significance, which dramatically reduces the individual's chances for defining an identity which provides a conception of self-worth (Lamont, 2012). For this reason, Lamont (2012) makes a normative claim for the co-existence of multiple matrices for evaluation to increase social resilience. It follows for the context of AI FORA research that it is important to keep the difference between assessment and evaluation in mind. While both concepts rely on collecting and somehow measuring data on performance (either on a nominal, cardinal, or ordinal scale), assessment is diagnostic whereas evaluation is based on the concept of values. Thus, assessment is the process of categorising the assessed subject. In contrast, evaluation involves judging the quality of a performance with reference to a standard (Parker et al., 2001) and deciding whether the standard is met or not.

In the differentiation of human beings through technology, AI systems are characterised by their particular capacities for 'statistical profiling' (Fourcade, 2016). These capacities currently lead to a great invasiveness and dominance of AI technologies in decision-making contexts (Newell & Marabelli, 2015; Zarsky, 2015), which cannot be analysed in isolation, but as, for instance, AI-in-society, which in turn calls for legal regulations (Tzimas, 2021).

AI-based social assessment is often seen as a continuation of the concept of bureaucratic governance (Peeters & Schuilenburg, 2018) as it has been described by Max Weber (1922) which is pertinent for the process of—multiple—modernisation (Eisenstadt, 2000). The technology of AI-based assessment systems fulfils the characteristics of bureaucratic governance of division of labour between technical experts, rule-based operations, and management of information (Muellerleile & Robertson, 2018).

At the first sight, the argument that AI-based technologies reinforce bureaucratic governance seems to be counter-intuitive: At the onset of the rise of the internet and modern information technologies, these technologies have often been perceived as emancipatory technological developments (Benkler, 2006). A free flow of information seemed to foster openness and transparency, creating the vision of bottom-up network societies and flattening hierarchies (Kreiss et al., 2011), thereby weakening hierarchical bureaucratic structures. However, with or without AI-based technology, social assessment remains a control instrument. Assessing individuals implies surveillance and subsequently assigning individuals to a categorical scheme. Such an assignment produces social sorting (Lyon, 2003) of a population into pre-defined categories. This may be the case in government as well as private companies, for instance involved in marketing. Furthermore, assessing risk scores ranging from recidivism to mortality involves the calculation of, typically probabilistic, numbers. Governance by numbers is a core element of bureaucratic transformation of governance (Porter, 1995; Hacking, 1990). Research on 'AI-in-Society' is widespread in many domains such as the development of algorithms or ethical assessment of AI

use (Dignum, 2018; Mantelero, 2018; Riterich, 2018). Bias, discrimination, and the fairness or unfairness of algorithmic decision-making raise ethical concerns. Autonomous decision-making by machines leads to the problem of accountability for such possible bias or discrimination (Pasquale, 2015). Leaving important aspects of individual lives to the rule of AI systems risks 'paving the way to a new feudal order' (Citron & Pasquale, 2014: 19).

The growing use of AI technologies for such assessment practices calls for Science and Technology Studies with a focus on its social implications (Fourcade & Healy, 2017). Techniques of human differentiation—human differentiation through technology: Methodologically, these changes of perspective entail a postulate of symmetry between explanans and explanandum, between technology and society as a concept to be explained (cf. Callon & Latour, 1992), require a 'sociology of translation' (Callon, 1986), and also demand a certain impartiality in the autonomous evaluation by the social sciences (Haraway, 1995).

AI assessing people, people assessing AI: The following paragraphs discuss fairness and trustworthiness issues of machine assessments. The problem of trustworthiness addresses the question of whether or to what degree the results of an algorithmic assessment procedure can be trusted. Algorithms provide exact calculations without errors. Thus, algorithmic decision-making implies the promise of higher accuracy than assessment and decision-making by humans. Big data and algorithms are perceived as having high potential in psychology (Adjerid & Kelley, 2018): Computer-based personality judgements are seemingly more accurate than those made by humans (Youyou et al., 2015). However, this claim has been rarely tested.

The quality of algorithmic decision-making depends on the multiple stages of the algorithmic value chain (Danks & London, 2017). For instance, an examination of the accuracy of a software for assessing the risk of offender recidivism has shown that the software predictions had only modest quality (Dressel & Farid, 2018). The software calculates the risk of a criminal offender's recidivism within the next 2 years based on 137 features of an individual and his or her criminal record. However, the neutrality of the software has been questioned (Angwin et al., 2016), causing a public debate in the USA on algorithmic fairness and discrimination (Flores et al., 2016; Kleinberg et al., 2016).

For this reason, the more fundamental question has been investigated how accurate the predictions are which are made by this algorithm. The study compares the false positives and false negatives, i.e. wrongly predicting recidivism and wrongly predicting no recidivism, of the software with assessments made by humans. The test persons for the study had no prior experience in criminal justice. With information of only 7 features given to the humans no statistically significant differences in the accuracy of the predictions made by the software and the predictions made by non-expert humans could be found (Dressel & Farid, 2018).

Fairness goes beyond the mere question of accuracy and trustworthiness: Even a perfectly accurate computation can be unfair. The critical examination of assessment software mostly concentrates on the issue of discrimination. While certainly discrimination is unfair, the concept of fairness or unfairness of AI-based social

assessment is broader. Discrimination does not equal unfairness. Besides a normative definition of fairness, a descriptive approach of what humans perceive as fairness reveals a multidimensional and context-dependent concept. An empirical investigation revealed at least 8 distinct elements in judging whether an assessment is perceived as fair (Grgić-Hlača et al., 2018). First, evidence for any judgement needs to be *reliably* assessable. Furthermore, the evidence in support of a judgement needs to be *relevant* for the case in question.

While these two conditions need to be obviously fulfilled for a fair judgement, it is also perceived as unfair to include information in an assessment that is regarded as falling under the right of the protection of *privacy*. Furthermore, justification of a judgement should be restricted to those circumstances that can be willingly influenced the individual. This is denoted as *volitionality*. For instance, it is perceived as unfair to include in a judgement if their father or mother had been arrested as this cannot be influenced by an individual. Also, judgements about fairness are influenced by the conviction that a feature need to have a *causal influence* on the behaviour of the subject under consideration. Fairness judgements are also influenced by the danger that the subject of assessment might be trapped in a *vicious cycle* through the assessment. For instance, it might be considered as unfair that in case of doubt someone is arrested rather than condemned to a suspended sentence, because friends of the assessed subject have been arrested as the arrest might cause a vicious cycle. Moreover, including a feature in an assessment that would generate disadvantages for a minority or otherwise sensitive group is typically perceived as unfair. Thus, perception of unfairness includes considerations about *causes of disparity in outcomes*.

Finally, it is considered as unfair if assessed features are correlated to *group membership*. The 8 dimensions might not be a complete list but are sufficient to explain fairness judgements in the survey (Grgić-Hlača et al., 2018).

Thus, perception of unfairness includes considerations about *causes of disparity in outcomes*. Finally, it is considered as unfair if assessed features are correlated to *group membership*. The 8 dimensions might not be a complete list but are sufficient to explain fairness judgements in the survey (Grgić-Hlača et al., 2018).

The survey shows that these dimensions are not perceived as unequivocally important but nevertheless agreed among the sample of different demographic characteristics and value orientations. These dimensions can be violated by software distinctly and need to be taken into account in development of fair assessment software. Regarding AI assessment tools, unfairness has been subject to a court decision in The Netherlands which stopped the application of a welfare fraud detection system that used personal data not complying with the right of privacy following the European Convention on Human Rights (Van Bekkum & Borgesius, 2021).

While the literature has not reached a consensus on the definition of algorithmic fairness (Srivastava et al., 2019), it is agreed that the increasing use of predictive assessment systems is associated with potential risks (Mitchell et al., 2021; Koebis et al., 2022). Such risks can result from bias (Pagano et al., 2023) or noise (Kahneman et al., 2021). An analysis (Binns, 2018) suggests that risks are related to the way how certain affected groups are represented in the system, in particular a biased

representation of affected groups such as religious or sexual minorities (Suresh & Guttag, 2019; Mehrabi et al., 2021), whereas Beigang (2022) concludes that unfairness might be due to factors outside the specification of the algorithm. Often, also the opacity of the systems provides a problem (Odilla, 2023). An overview of the literature on algorithmic unfairness is provided by Starke et al. (2022).

A taxonomy (Shelby et al., 2023) identifies *representation harm*, i.e. societal beliefs and unjust hierarchies are represented in the system, *allocative harm* which denotes how these representations affect the decisions made by the system, *quality-of-service harm* meaning that the choice to design the system for specific users might discriminate other users, *interpersonal harm* which may arise when social relations are affected by technological affordances, and *social systems harm*, when a technological system destabilises a social system and increases inequality. Thus, there is a strong risk that technology reflects and reinforces biases and discrimination of the society at large (Edler Duarte, 2021). Finally, algorithmic decision-making is challenged by the fact that law and administrative regulation are situated in dynamic, ever changing social environments and need to balance between competing demands. However, flexible adaptation to varying circumstances provides a challenge for rules codified in algorithms once the rules are hardwired in the system (Casey & Niblet, 2017). While it could be argued that machine learning might detect patterns, for instance in court decisions, which interpret and thereby develop the codified law, this comes at the cost that it both freezes the standards of the time the implementation was encoded and that the further adoption of legal interpretations, for instance due to changing social values, will come to a halt.

How about bias and discrimination? At first sight, computers seem to be efficient and objective following the standards of formal rationality, because they undertake complex calculations on raw data far more efficiently than humans. Algorithmic procedures objectively follow clearly defined rules that are characteristic of formal rationality. However, already the concept of 'raw data' to be processed by computers is illusory (Gitelman, 2013). For detecting sources of algorithmic bias, a scheme has been developed by Odilla (2023) to differentiate between infrastructural, individual, and institutional sources of unfairness in the development of AI predictive tools.

Danks and London (2017) put forward a perspective which attempts to take the whole lifecycle, and, in particular, the use of such systems, into account. Their scheme has been extended by Silva and Kenney (2018). It provides a kind of algorithmic 'value chain' (Silva & Kenney, 2018) concerning the different development and application stages of algorithms.

Following the model as depicted in Fig. 1.1, sources of biases can be identified at every stage along this value chain. At the stage of the input data, Danks and London (2017) differentiate between training data bias and algorithmic focus bias. Training data bias is the most obvious source of flaws (Barocas et al., 2017). Algorithmic focus bias denotes the fact that all data come in categories, and bias may occur by either the inclusion or exclusion of information in these categories (Danks & London, 2017). Sources of bias at the stage of algorithms are denoted as algorithmic processing bias, for instance by the weighting of variables.

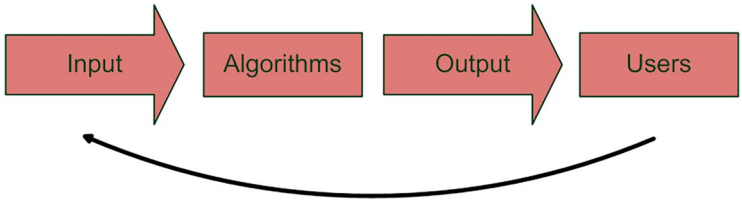

Fig. 1.1 Algorithmic value chain following Silva and Kenney (2018)

At the output stage, several sources of bias are distinguished: misinterpretation bias may occur, for instance, in the process of transforming stochastic information into a deterministic decision. Automation bias happens when the users of the software believe that the output of an algorithm, which is typically stochastic, is objectively true. Non-transparency bias denotes the fact that the result of a neural network cannot be explained even by the programmer (Kuang, 2017). This points to a lack of transparency of assessment software. In the absence of a reason for a decision, the people who are subjects of the decision cannot identify whether they have been potentially treated in a discriminatory manner. Finally, at the stage of the users, consumer bias and feedback loop bias are identified. These are not built into the algorithms. Rather, they occur through the use of digital platforms, i.e. reflect the bias of the users. However, online platforms and the use of the internet generate more data.

This provides input for learning systems and, in consequence, feedback between the algorithm and the users. It is important to note that every stage has to be taken into account with equal weight. While algorithms may produce incorrect results in comparison with some empirical data, bias is not something purely technical in the software development process. Algorithms produce bias when they are applied, which goes along with a violation of values.

Values, however, are inherently a social concept: They are guidelines for social conduct defining what is socially accepted as good or desirable. For this reason, bias at all stages of the algorithmic value chain implies discrimination generated or enforced by AI-based social assessment systems.

In a sociological perspective, technology is to be analysed as a reflection of the social value system: ‘Algorithms do not make judgments they are the products and the tools of human judgments’ (Burk, 2019).

While assessing AI-based social assessment, it has to be noted, though, that most decision-making is not entrusted completely to algorithms (ADM: automated decision-making). However, often it remains opaque where and how algorithms come into play (Barocas et al., 2017; Silva & Kenney, 2018). Therefore, AI governance is in need of ethical guidelines. For developing such guidelines, the full lifecycle of algorithmic systems has to be taken into account. Dignum (2018) distinguished between three dimensions in which ethical reason is relevant: Ethics-by-design, which addresses the integration of ethical reasoning in artificial autonomous systems; ethics-in-design addresses methods for analysis of ethical

implications of AI systems; whereas ethics-for-design regulates the codes of conduct for ensuring integrity of developers and users of AI systems.

It has been suggested in the context of ethics-by-design to stop a system before it can become destructive (Orseau & Armstrong, 2016) or to develop a moral Turing test (Wallach & Allen, 2008). However, these suggestions are difficult to realise and might not keep up with the complexity of the target. In the context of value-sensitive design (Aldewereld et al., 2014) and evaluation (Diakopoulos, 2015) of autonomous systems, Rahwan (2018) instead argues for the need of a new algorithmic social contract to prevent the danger of losing democratic control over algorithmic decision-making by extending the concept of human-in-the-loop (HITL) (Sheridan, 2006). HITL instantiates human supervisory control of AI systems for identifying misbehaviour and establishing an accountable entity. Due to their wider range, in the case of AI-based social assessment systems human supervision additionally has to account for a trade-off between different values and how to agree on distribution of costs and benefits to different stakeholders in a new social contract. This is what Rahwan (2018) denotes as society-in-the-loop. For initiating public feedback on regulations and legislations, an articulating and negotiating between (different) values is needed as well as monitoring compliance. Arnold and Scheutz (2018) suggest that AI systems should be tested in advance in a simulated environment that contains an ethical scenario-generating mechanism. The objective of AI FORA can be read as realising the proposals made by Rahwan (2018) and Arnold and Scheutz (2018): Building a co-creation lab for initiating public feedback on instantiating regulations of AI-based social assessment systems that is informed by a scenario-generating simulation.

A value-sensitive approach to AI-based social assessment needs to be participatory with the society in the loop and also context-specific: Context, especially in terms of the cultural and social embeddedness of values and attitudes, is key to understanding the use of AI to perform practices of social assessment. Discriminatory practices to distinguish beneficiaries from non-beneficiaries depend on the beliefs, norms, and values in place as reference frameworks for public resource distribution in different societal contexts. There is no approach to social assessment that would be perceived as fair everywhere.

Fairness concepts vary across national welfare systems depending on culture, religious tradition, and belief system. Furthermore, those heterogeneous value systems are in constant flux and impact on the margin of discretion (Alesina & Angeletos, 2005; Venkatapuram et al., 2017; Hank & Erlinghagen, 2009; Picot & Tassinari, 2017; Schaefer et al., 2015) that public administrations have while following legal and policy frameworks on social assessment.

The next section is dedicated to AI FORA's central assumption, i.e. the importance of cultural values and social context for AI-based social assessment. Again, we use the ubiquitous feature of assessment to approach this: In order to use 'context' as an explanatory variable for observed differences, we need to *identify* 'context' in the first instance, *categorise* different contexts (Bowker & Star, 1999; Navis & Glynn, 2010), *compare* (Epple et al., 2020; Deville et al., 2016; Heintz, 2016; Steinmetz, 2019), and *valuate* (Karpik, 2010; Kornberger et al., 2015; Lamont,

2012) the various contributions of context. Comparative Studies and Valuation Studies share the underlying mode of differentiation (Heintz, 2021).

The Role of Culture and Context

Context is important. AI-based social assessment systems need to be situated within their distinctive and varied cultural contexts—contexts that represent different languages, values, worldviews, religions, and racial/ethnic groups, located within and across nation-state boundaries. This section starts by reviewing major theoretical approaches to cross-cultural measurement and the emerging literature describing strategies for accounting for different cultural contexts in AI applications around the world. It will then discuss their limitations for capturing the culturally situated nature of participatory AI projects at a local level. The AI FORA approach for the latter can be described by an adaptation of the so-called 'culture cube' model, which will be introduced at the end of this section.

Can social justice be processed by technology in different value contexts worldwide aiming at global fairness? 'Fairness' is a global issue: Everywhere, fairness, i.e. social justice, is seen as something worthy to be achieved. However, though the general goal is shared, interpretations, meaning, and ways differ. While AI ethics may identify *if* and *what* constructs, such as fairness, trust/trustworthiness, and responsible use, are universally valued and endorsed (see section "The Pervasive Practice of Assessment and AI"), consideration of *how* such ideals are construed and interpreted within different cultural contexts is another matter altogether.

That is, culture represents the 'shared symbols and meanings that people create in the process of social interaction which orient people in their ways of feeling, thinking, and being in the world' (Carpenter-Song et al., 2007: 1362) serving as a primary filter through which people engage with the world. Perceptions of AI emerge from these shared interactions, which are part of the 'meaning-centred' definition of culture. This process-oriented approach to defining culture emphasises how it is both *located within* and *emerges from* social interactions, so that culture is always being shaped by people and shaping people, a process that is inherently continuously dynamic, fluid, and relational. Can a machine map the tremendous diversity of ideas regarding social justice found in different cultural contexts? And perhaps even dynamically address social reform processes that seek to reduce discrimination and bias in individual societies? A key question that emerges in the AI FORA project, then, is how does culture affect how ethical dilemmas related to AI use in social service delivery are identified, framed, and addressed? While there may be underlying universal principles, such as fairness and trustworthiness that shape how AI ethical issues are framed, these principles may not be valued and/or construed in the same way in different cultural contexts.

Wong (2020) describes how differences in cultural values represent challenges to global ethics and a human rights approach to AI governance. He asserts that: '…normative standards for the *global* ethics and governance of AI technologies should be

viewed not as a pre-determined endpoint but as an on-going process of negotiation and construction, and thus ought to be *open* and *responsive* to cultural values. The requirement of openness and responsiveness demands scholars and practitioners of AI ethics and governance to think *together with* the cultural others when deciding whether specific normative standards are appropriate for the evaluation of AI technologies and why they are appropriate in a cross-cultural setting' (p. 713). Wong's human rights approach to culture and global ethics in AI governance also reflects a process-oriented, relational approach to construing culture. The challenge of assessing participatory AI across cultural contexts, however, is that most definitions of culture used within cross-cultural measurement are not process-oriented and dynamic but tend to treat culture as a relatively fixed feature that can be compared across cultures, shifting slowly over time. From this 'behaviour-based' view of culture, culture is seen as located within and reflected through the values, customs, and behaviours of people within a defined context (Carpenter-Song et al., 2007). When cross-cultural comparison is the primary focus and research aim, this treatment of culture as fixed and behaviour-based enables the comparison of countries across different domains and across different 'cultural dimensions'. These cross-cultural approaches to measurement are described in the next section.

A Review of Cross-Cultural Measurement Approaches

How are cultures similar and different to each other, and how do these variations affect human behaviour, including the perception and approach to AI ethics? One major issue in cross-cultural psychological research is that 'culture' is often equated with 'nation' or 'ethnic group' so that research may be less 'cross-cultural' than 'cross-national' in nature (Georgas et al., 2004).

From the perspective of culture as relatively fixed, culture is a 'societal concept and a group phenomenon [that] consists of values, norms, beliefs, attitudes, and more, forming patterns that distinguish one group of people from another, be it a country, a region, an ethnicity, or some other group' (Kaasa, 2021: 339; see also Hofstede, 2001; Schwartz, 2008). Using this concept, three different major approaches to measuring culture are described below, reflecting the theoretical frameworks of Hofstede (1980, 2001), Schwartz (2008), and Inglehart (1997).

Hofstede's Cultural Dimensions

The *dimensional* concept of culture innovated by Hofstede dominates the field of cross-cultural psychology and international business management (Beugelsdijk & Welzel, 2018). In his groundbreaking research, Hofstede (1980, 2001) identified four dimensions of cultural values using factor analysis for a survey conducted with a global sample of over 116,000 IBM employees working in 40–72 countries

between 1967–1973 (Kirkman et al., 2006): *Power distance, uncertainty avoidance, individualism, and masculinity*.[2]

Although his items were originally designed to survey employee attitudes across different cultural contexts, Hofstede identified cultural dimensions to enable cross-cultural comparisons. He aggregated individual scores to create a mean for each country and examined the relationship between country-level scores for different nations with national social and economic indicators (Georgas et al., 2004). Later, Hofstede added two cultural dimensions, using data from the World Value Survey, which he labelled *Long-term vs. Short-term orientation* and *Indulgence vs. Restraint.* Cultures identifying as high on long-term orientation are viewed as 'more future oriented and easily accepted delayed gratification of individual effort' while cultures with short-term orientation are more likely to have an emphasis on the 'here and now' (Beugelsdijk & Welzel, 2018). Indulgence reflects the degree to which emotional expression is encouraged and enjoyment of momentary pleasures is favourably viewed versus cultures which are characterised as tending to suppress emotions and with a preference for following codes of conduct and discipline (Hofstede et al., 2010; Minkov & Hofstede, 2012).

The most widely developed of Hofstede's original cultural dimensions, *Individualism-Collectivism*, examines 'the extent to which people are autonomous individuals or embedded in their groups' (Triandis & Gelfand, 2012: 499), and originally associated individualism with *independence, inner-directedness,* and *personal* interests and collectivism with *emotional dependence, family, and group interests* (Hofstede, 2001). Individualists have a strong focus on their own concerns, are motivated by their own preferences, and have loose social ties, while collectivists see themselves as part of a group, and are oriented towards harmony and fulfilling group needs (He & Lee, 2020). Research in psychology continues to demonstrate the importance of this dimension for understanding cross-cultural differences in self-construal, motivation, and emotion (Markus & Kitayama, 1991), and even basic perceptual processes and thinking styles (Masuda & Nisbett, 2006; Chen et al., 2016; Koo et al., 2018).

At the same time, a meta-analysis indicated serious challenges about the limited conceptions of culture inherent in these measurement approaches (Oyserman et al., 2002; Miller, 2002). Another cultural dimension is *Power Distance*, defined as 'the extent to which the less powerful members of society expect and accept that power is distributed unequally' (Hofstede, 2001). While this concept was originally measured as part of the individualism-collectivism dimension, it was split off into its own category based on theoretical considerations (Hofstede, 1980; Kaasa, 2021), although empirically, individualism and power distance represent two ends of a single dimension (Beugelsdijk & Welzel, 2018). Power distance contrasts authority, autocratic decision-making, obedience, and fear of disagreement, to consultative decision-making, participation, and willingness to disagree (Kaasa, 2021).

[2] After applying a criterion for a minimum of 50 persons per country, the 72 countries in the sample were reduced in number to 40 countries (see Beugelsdijk & Welzel, 2018).

The last dimension, *Uncertainty Avoidance*, is related to the 'extent to which members of a society feel threatened by uncertainty and ambiguity' (Hofstede, 2001). On one end, it is associated with a desire for *order, rules, traditions, and security,* as well as a *resistance to change* and *intolerance* and on the other end, with *willingness to take risks, (desire for) as few rules as possible,* and *tolerance towards outgroups*. The dimension of *masculinity* originally contrasted masculinity as an achievement orientation (e.g., *achievement, self-realisation, assertiveness, competitiveness,* and *ambitiousness*) with femininity as a more interpersonal orientation (e.g., *relationships, affectionateness, compassion, benevolence, environment, physical conditions, employment security)*. Subsequent research shows, however, that the masculinity cultural dimension is better understood primarily as high and low levels of *achievement orientation*, since the interpersonal, femininity dimension appears more aligned with the collectivism dimension of individualism-collectivism than reflective of gender egalitarianism (Maleki & de Jong, 2014; Minkov & Kaasa, 2021). Relatively speaking, there is somehow less research on the masculinity dimension compared with Hofstede's other three original dimensions (He & Lee, 2020), especially individualism-collectivism, which has evolved into a multidimensional and multi-level construct (Triandis et al., 1988; Oyserman et al., 2002; Singelis et al., 1995).

Schwartz's Personal Values Inventory

Schwartz (1992) focused his work on identifying universal cultural values using large datasets of 38 samples of teachers and 35 samples of university students from 38 nations, mostly in Europe, to empirically measure and define various elements of culture (Kaasa, 2021; Beugelsdijk & Welzel, 2018). Schwartz (2012) identified ten personal values that guide attitudes and motivation, organised further in terms of how these values regulate personal interests and characteristics (e.g., *Achievement*, *Stimulation, Hedonism, Power, Self-Direction)* or relationships with others (e.g., *Benevolence, Conformity, Tradition*), with two 'boundary values' that regulate the boundary between personal and social interests and goals (e.g., *Security* and *Universalism*). Individual scores were aggregated to create country-level scores, then, using multidimensional scaling Schwartz identified seven different value clusters grouping different nations based on their country scores. These value clusters were configured into three dimensions: *hierarchy vs. egalitarianism, mastery vs. harmony,* and *embeddedness vs. autonomy.*

These dimensions have been viewed as conceptually overlapping with Hofstede's dimensions (Nardon & Steers, 2009; Maleki & de Jong, 2014): Embeddedness vs. autonomy shares similarities with individualism-collectivism; hierarchy vs. egalitarianism is similar to power distance and mastery-harmony has some overlap with masculinity-femininity, although this dimension is framed more as 'relationship with the environment' (Nardon & Steers, 2009). Schwartz (2004) notes that while there are conceptual overlaps between his identification of cultural values with Hofstede's cultural dimensions, they are not the same.

Inglehart & Welzel's Cultural Map

Inglehart (1971, 1990, 1997) was a political scientist who developed the European Values Studies, which became the World Values Survey, a global survey that systematically examines social and cultural change in over 100 countries. He documented generational changes in cultural orientations in Western countries from an emphasis on *existential security* or materialist values to *expressive freedom* or post-material values (Beugelsdijk & Welzel, 2018). Similar to Maslow's hierarchy of human needs, security and freedom are both needed for societal thriving, but security is prioritised when people do not have their basic socioeconomic needs met and are oriented towards addressing threats. However, as soon as safety is established, freedom is prioritised to enable 'ingenuity, creativity, and recreational pleasure' (Beugelsdijk & Welzel, 2018). Inglehart and Welzel (2005) examined these generational shifts in priorities from 'survival' to 'emancipative' values into a revised theory of modernisation, and later into an 'evolutionary theory of emancipation' (Welzel, 2013). Their approach has been particularly influential in sociology and political science in describing cultural change as generational shifts in behaviour, changes that tend to happen between generations rather than within generations because 'people tend to stick more strongly to their once adopted values as they age' (Beugelsdijk & Welzel, 2018: 1470).

Inglehart and Welzel (2005) revised Inglehart's work to provide a world map of cultures with two dimensions, traditional vs. secular world values (y axis) and survival vs. self-expression values (x axis). The authors place countries into the four resulting sectors based on their World Values Survey scores, with economic development, as well as religious and cultural heritage used to place countries into clusters that reflect their shared values. Traditional values emphasise the importance of religion, family ties, and deference to authority, while secular-rational values tend to minimise the influence of religion, traditional family values, and authority. The maps are updated regularly to examine cultural change across generational cohorts.

Beugelsdijk and Welzel (2018) synthesised Inglehart's dynamic concept of culture with Hofstede's dimensional approach, as an empirical response to criticisms that Inglehart's approach tends to be dimensionally reductionistic, while Hofstede's dimensional approach neglects cultural dynamics. They used data from the European Values Studies and World Values Surveys to examine 5 generational cohorts born between 1900–1999 in 110 countries, representing over 495,000 individuals.

They re-examined Hofstede's dimensions and empirically identified three dimensions that were mapped and labelled as (a) individualism-collectivism (combining low power distance and collectivism); (b) duty vs. joy (renaming restraint vs. indulgence); and (c) trust vs. distrust (renaming uncertainty avoidance, with high levels as trust and low levels as distrust). They found intergenerational value shifts on their Hofstede-inspired dimensions towards more individualism and joy/indulgence) which were explained by economic development, generational effects, and country-specific factors (e.g., geographic location and political context). They also found that, while there was evidence for value shifts, national cultural differences and relative country rankings were fairly stable over time.

AI FORA and Measuring the Influence of Culture at the Level of the Nation State

AI FORA investigates AI-based public service provision of national welfare systems within nine country case studies. It is the nation state that, using its common legal and administrative framework, allocates services to its citizens. Welfare selectivity is, i.e. who gets what from the state. Therefore, for measuring the influence of culture, the nation state as unit of analysis for AI FORA's country-level data seems to be the appropriate framework. It is the state that provides the container (Hofstede, 2003) for negotiating culturally specific administrative practices in social welfare provisions (Anderson, 1983; Triebe, 2014; Druckman & Nelson, 2003). Comparing at such granular level, we see differences in social service provision that are clearly related to culture, which spring to the eye immediately: For example, in India, the allocation of social services (Public Distribution System PDS) is related to religious caste membership, while in China, the allocation is related to desired citizen behaviour (Social Credit System). These are mostly cultural differences which are restricted to the nation state being manifested in national welfare policies. Such high-level cultural imprints were, for example, discussed at AI FORA stakeholder workshops with policy representatives (policymakers, senior civil servants, social security services and NGO) of different countries reflecting on their national welfare strategies, normative issues, and context-dependency regarding AI-based social assessment applications. Opinions and attitudes varied among nations and their domestic public discourses. Fairness concepts implemented mirrored fundamental cultural issues translated to fairness concepts of national public welfare policy in each case study country: They resulted in heterogeneous categorisation systems for social assessment in administrative practices of state agencies allocating social services.

Thus, to capture as much cultural heterogeneity as possible from a nation-state perspective, the Inglehart/Welzel cultural map based on the seventh wave of the World Values Survey 2021 (WVS Database, https://www.worldvaluessurvey.org/WVSContents.jsp, last accessed 27.05.2024) was consulted supported by a factor analysis from the Hofstede dimensions (Hofstede et al., 2010). Selecting countries from cultural clusters that were as heterogeneous as possible led to the set of case studies shown in Fig. 1.2: Estonia, Spain, USA, India, Germany, China, Ukraine, Iran, and Nigeria (Mexico was selected but could not be realised).

AI FORA's country case studies differ in the Inglehart/Welzel coordinate system of *Survival vs. Self-Expression Values* and *Traditional vs. Secular-Rational Values*, in the six Hofstede dimensions as explained above, and in three more general categories (*level of digitisation, government form, and country size*) that can be assumed to have an effect on the implementation of and the discourse on AI-based social assessment (Erumban & de Jong, 2006). In AI FORA, the cross-cultural comparison frameworks with their focus on the nation state were mainly used for selecting case study countries and motivating this choice. However, to check for the assumed cultural differences on a high granular level and enable a generic comparison

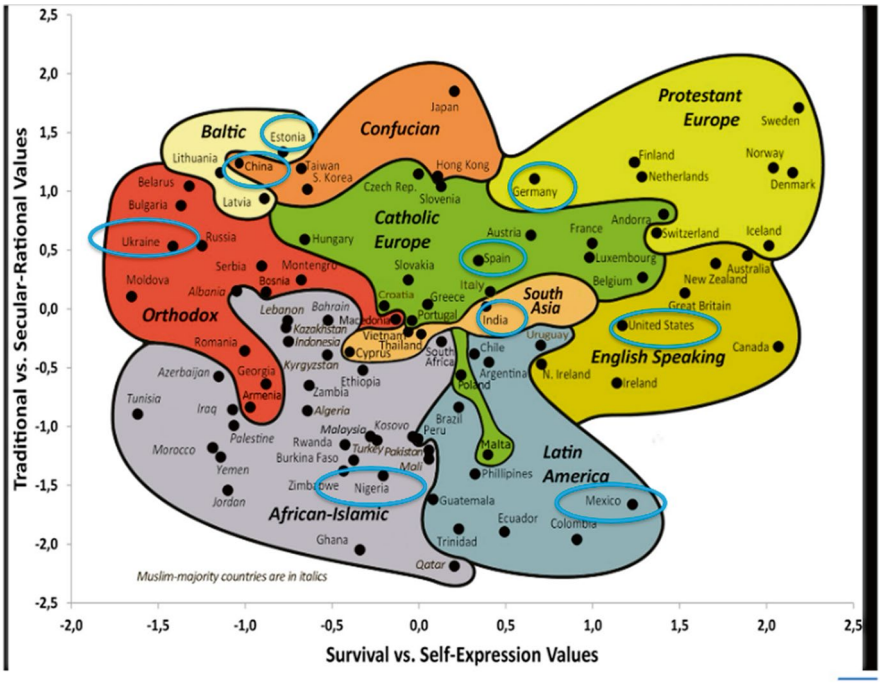

Fig. 1.2 Inglehart-Welzel Cultural Map (Wave 7, 2017-2021); blue circles = AI FORA case studies

between case studies, the specific methodologies used in AI FORA provided further opportunities for the cross-cultural measurement frameworks.

For example, Hofstede's six-dimensional framework can be implemented by an agent-based model (ABM) of cultural-dependent decision-making (Van Damme et al., 2023; Hofstede et al., 2021); and the ABM can be translated into a serious game for human players (Ahrweiler et al., 2023) to be played in different national contexts. Games can be used not only to see whether the same game played in one national context will produce different outcomes in another context, but also whether the outcome is the one suggested by the Hofstede dimensions when the decision-making situations in the game are value-coded according to the cultural comparison framework. Results from the AI FORA project concerning this comparison exercise between case studies are reported in Chap. 13 (Conclusions). Of course, this approach also reveals much heterogeneity. Social service provision is a problem context that everybody is exposed to, with each of us deciding how to behave when faced with specific situations. Some of the decision-making is related to the bigger societal footprint as cultural background, but there is also a huge variety of all kinds of behaviours within a culture due to individual or group characteristics, for example, very fundamentally, being committed to societally-shared norms and values, or defecting them, but also, because there are different cultures within a state.

The cross-cultural comparison models presented above use different approaches to responding to the challenge of linking a societal level cultural context, measured as country-level indices, to psychological variables at the individual level (Georgas et al., 2004). Social cognition processes at the individual level provide key parallels to AI processes, describing how people 'select, interpret, remember, and use social information to make judgments and decisions' (Aronson et al., 2023). That is, in order to examine AI ethical concepts such as fairness and trustworthiness within a cultural context, it is not enough to describe 'dimensions of nations', but also requires an examination of psychological processes at a more local level. The analyses conducted by Beugelsdijk and Welzel (2018) were completed at the country level, and they note that 'dimensions of cultural variation found across nations tend to be robust in their configuration, stable over time, and strongly linked to other characteristics that describe a society's aggregate reality' (p. 1498). Nonetheless, an 'ecological fallacy' occurs when inferences about individual behaviour are deduced from group scores, so that country scores cannot be used to infer anything about the cultural values or orientations of individuals at a more local level (Winzar, 2015).

The Culture Cube: Articulating Culture Within Local Contexts Advocating for a 'Mini-Ethnography Approach'

The 'culture cube' model was originally developed for identifying and articulating the cultural underpinnings of programs across different communities participating in a large initiative to reduce mental health disparities in the state of California (Abe et al., 2018). This initiative shared key similarities to the AI FORA project in that explicit attention to how culture and context influenced and affected the design, implementation, and impacts of diverse community-based approaches to the prevention and early intervention for mental health issues was an important priority. The basic elements of this conceptual model for providing qualitative, locally generated descriptions enable that cultural and contextual factors can be recognised within each AI case study.

Culture emerges out of interpersonal realities and dynamic relational processes that are themselves embedded within historical, social, political, and economic contexts (Carpenter-Song et al., 2007; Garneau & Pepin, 2015; Gregory et al., 2010). This dynamic and nested understanding of culture underscores how processes of social cognition, reflecting particular moments in time, place, and with specific people, shape our construals and experiences of culture. Gallimore and Goldenberg's (2001) concept of *activity settings* represents the everyday activities through which 'the less visible cultural models for living are expressed and manifested' (Abe et al., 2018: 124; Gallimore et al., 1993; Sarason, 1972). These cultural models are socialised through everyday life and represent nonconscious schemas that help shape and give meaning to experience. Gallimore's framework helps to underscore that culture can be viewed within the architecture of how programs are designed and

in how they operate and also serves as a reminder that cultural assumptions, beliefs, and values are often not consciously recognised or explicitly described. Kleinman's *explanatory models of illness* framework examines the cultural assumptions regarding the meaning of illness from multiple perspectives using key questions (Kleinman, 1981, 1988; Weiss, 1997).

The original culture cube uses elements from both of these frameworks to identify the explicit everyday architecture (activity settings) through descriptions of three specific characteristics, *project, place*, and *person*s (3 Ps); these characteristics 'contain' values and assumptions through which key questions regarding the *conceptualisation of the problem, perceived causes*, and expected *consequences/ changes* (3 Cs) resulting from the intervention can make culture explicit and visible (explanatory models). The 3 Cs are reinterpreted as they apply to the AI FORA context and are labelled as *culture, challenges,* and *changes*. The 3 Ps of *program, place,* and *persons* represent the visible dimensions of the culture cube, in that they can be observed and documented without much inference. Looking at the AI FORA case studies, the description of *program* refers to *the type of AI system used and the focus of AI application efforts*. In each participating country, the specific use of AI technology for providing social services to vulnerable populations varies. This *program* description enables an understanding of what how each participating country defines its specific *program* of social service provision within the AI FORA project. For *place*, the culture cube focuses on *where the program takes place, both in terms of its geographical location and organisational position*. The inclusion of place helps to anchor the program within a specific geographical, cultural, and systemic/ organisational context. The third visible dimension is *persons*, which includes *all participants in the project*, including which *population is regarded as vulnerable and how their needs are defined, who is addressing their needs*, and *whom else is involved in the process*. For the AI FORA project, each case study includes an actor network map to enable an understanding of this network. In the culture cube model, *persons* are not incidental to the delivery of the services, but are seen as intrinsically important to understanding the distinctive features of the program.

The *activity settings* framework from which these 3Ps are drawn underscores that the basic unit of analysis for the AI FORA project is multidimensional, consisting not just of programs in an abstract sense, but as anchored in the persons and places in which they are embedded. The culture cube's contributions reside in combining the visible dimensions of a project, with the invisible assumptions, values, and expectations inherent within them, reinterpreted as the 3 Cs of *culture, challenges,* and *changes* within the AI FORA project. Specifically, *culture* refers to the *assumptions, values, and norms related to AI use, especially related to perceptions of AI fairness and equity, as defined within the specific cultural context*; *challenges* focus on the sources of challenges in *implementing AI in ways that address the needs of vulnerable populations in ways that are participatory, reduce inequities, and increase fairness*; and *changes* as the *AI-use outcomes and systems changes desired, drawn from the perspectives of multiple stakeholders, including those from vulnerable populations*. For example, for *culture,* is the use of AI in decision-making seen as more fair or less fair than human decision-making? Should humans or AI

have the 'final say' in decisions that affect public social service delivery in order for the distribution of resources or goods to be viewed as most fair? For *challenges*, what are the particular expressions of obstacles in the efforts of countries to apply participatory AI for addressing the needs of vulnerable populations? Are the voices of vulnerable populations valued? Do perceptions of corruption, neglect, threat, and/or ineffectiveness affect perceptions of AI use? And, for *changes,* what would a fair, ethical AI application within a given project even look like from the perspective of different stakeholders? How discrepant is the ideal vision from these differing perspectives?

These dimensions are not immediately evident, even with extensive descriptions of a project, unless explicitly lifted up and considered. Yet, these invisible dimensions get at the heart of how social justice is processed by technology in different value contexts worldwide aiming at global fairness.

The case studies in this volume include attention to both visible and invisible dimensions of the implementation of AI to public social service within different countries. These countries represent the range of different value zones contained within Inglehart and Welzel's (2005) cultural map and can also be described at the macro-level, in terms of Hofstede's cultural dimensions. Yet, the way in which culture and context are described or measured in order to inform a discussion of AI ethics, especially within a participatory AI framework, must be both *local*, to honour the participatory, community-informed AI approaches, and *holistic*, to represent culture as multifaceted and dynamic, emerging from people's relational bonds and expressed through their place-based activities.

The Case Studies[3]

To achieve the required heterogeneity of national welfare systems from a broad variety of cultural contexts, we used the level of digitisation, the government form, and the country size as general distinguishers as well as the cultural clusters of the seventh wave of the World Values Survey 2021 supported by a factor analysis from the Hofstede dimensions. The aim of including at least one country per cluster of the Inglehart/Welzel map could be achieved with the exception of Mexico. Authors were asked to present the empirical results of their case study on AI-based social assessment in public service provision of their country focussing on the role of cultural values following the project's central assumption that these are important to consider for appropriate AI use, in particular with regard to potential culture-related bias and discrimination issue and the situation of vulnerable groups. The task of each chapter was to contribute empirical knowledge and recommendations for 'better AI' for addressing AI FORA's central objective to co-produce contextualised,

[3] The authors of the chapters contributed the abstracts for this section.

participatory, value-sensitive, responsive, and dynamic AI for social assessment in public service distribution.

Not all case studies chose the broader features of their national welfare system (India, Spain, Ukraine, Nigeria, Iran), mostly with a geographical focus, as their unit of analysis, but some concentrated on a particular social service area such as public unemployment benefits (Estonia), public systems of care (USA, China), or state benefits for asylum-seeking refugees (Germany). Thus, all chapters start with introducing their specific focus and their country-specific research questions related to AI FORA's interest in the influence of values, culture, and context on (AI-based) social assessment for public service provision. Chapters then provide an overview on the actors and stakeholders involved and on the current social assessment practices for providing public services.

They identify why context and culture is important for social assessment of beneficiaries in their country, and in which parts AI or data analytics technologies are implemented. Following up on these structural descriptions, most chapters present results from qualitative social research that cover existing and desired systems of AI-based social assessment in the case study countries: Problems and barriers of the current assessment systems are identified with a specific focus on vulnerable groups; and desired systems as projected by stakeholders in case study countries are sketched out.

Empirical results mostly stem from using the AI FORA methodologies of inclusive technology co-design as presented in Chap. 2. Data and results come from literature reviews, document analyses, interviews, focus groups, world cafés, Participatory Systems Mapping, gamification, surveys, questionnaires, participatory workshops, media discourse analysis, and others. Chapters end with a discussion of results and an outlook to the modelling and simulation activities in the next step of research, which will be presented in the follow-up volume. The next section shortly introduces the case study chapters[4].

Spain

Chapter 4 by **Albert Sabater, Beatriz López, Roger Campdepadrós, and Cristina Sánchez** is about 'Participatory Action Research for AI in Social Services: An Example of Local Practices from Catalonia'. After reviewing some experiences of AI use in social services in Catalonia due to its key role in developing the first AI strategy in Spain, the work adopts an ethics-from-below perspective through Participatory Action Research (PAR), engaging social workers, policymakers,

[4] It has to be noted that Iran, Nigeria, and the Ukraine did not belong to the original set of funded case studies, because they joined AI FORA later. Work in these case studies was self-funded. The dedication of the respective researchers Hassan Bashiri, Emmanuel Ejim-Eze, and Oleksandr Khyzhniak was highly appreciated by the AI FORA team. We especially acknowledge the Ukrainian contribution in times of war.

technologists, developers, and social scientists to assess AI-based social services in Catalonia. Two key meetings under the umbrella of the AI FORA project illuminated the discourse, highlighting the importance of continuous, inclusive evaluation to maintain community relevance and integrity.

Methodologically, research employs focus groups for a comprehensive exploration of the advantages and disadvantages of AI in social services, and a World Cafe to gain expertise and perspective around four central themes: (1) data-driven social services for resource allocation, (2) predictive analytics and early intervention, (3) evaluation and continuous improvement, and (4) stakeholder collaboration. The discussion highlights the necessity of continuous ethical reassessment and updates within organisations to maintain integrity, especially in the context of AI in social services, emphasising the importance of aligning with evolving societal norms and technological advancements. Additionally, it emphasises the role of organisational culture and digital literacy in adopting AI, advocating for a balanced approach that integrates technological innovation with human empathy and judgement, while addressing the challenges of ensuring ethical AI deployment through proactive, inclusive, and culturally sensitive practices.

Estonia

Chapter 5 by **Triin Vihalemm, Maris Männiste, Avo Trumm, and Mihkel Solvak** is titled 'Specialists and algorithms: Implementation of AI in the delivery of unemployment services in Estonia'. The chapter examines the utilisation of an AI-based tool to evaluate unemployed individuals who receive welfare services from specialists at the Estonian Unemployment Insurance Fund (EUIF). In this case, the machine collaborates with human decision-makers to enhance advising unemployed clients. Specifically, the automated decision support tool provides background information to EUIF consultants by assessing the likely time when clients will find employment. This assessment is based on data related to the current labour market situation within the relevant segment for unemployed individuals, considering factors such as training, residence, and education. By analysing documents and conducting interviews with EUIF consultants, the authors explore various models for sharing decision-making responsibility between humans and machines based on the core values of AI implementation in Estonian society: Effectiveness of information processing and the fairness of decisions made by machines compared to humans.

Germany

Chapter 6 by **Elisabeth Spaeth** deals with 'AI use in the asylum procedure in Germany: Exploring perspectives with refugees and supporters on assessment criteria and beyond'. The chapter takes the AI-based Dialect-/Language recognition

software ('Language Biometrics Assistance System', DIAS), used by the Federal Office for Migration and Refugees (in German: BAMF), as an example for assessing asylum seekers' 'eligibility'. This software should support decision-makers identifying the country and/or region the refugee is coming from, based on their language features, as more than half of those who apply for asylum do not have their passport (anymore) or other supporting documents. Based on document analysis, the chapter presents the most important stages of the asylum procedure, its AI-component as well as political, legal, and technical context. Empirical research conducted based on interactive sessions, such as a world café, was enriched by former exploratory interviews with important stakeholders supporting refugees as well as desktop research to present current discourses. The experienced assessment criteria in the asylum procedure, and beyond, highlighting the experiences of those affected by these assessments, namely refugees, and of those 'guiding' refugees through the different procedures are illustrated and analysed in this chapter. Furthermore, these insights are discussed in the light of the current use of AI for assessment, exploring its implications for fairness through the lens of legitimacy of asylum bureaucracy and agency of refugees.

Iran

Chapter 7 presents 'Social Assessment for Public Service Provision in Iran: Applying AI in the Targeted Subsidies Plan' written by **Hassan Bashiri**. The chapter delves into the intricate relationship between AI and public policy by investigating its application within the Iranian Targeted Subsidies Plan. The research methodology encompasses experimental techniques, involving data collection, document analysis, interviews, questionnaires, and quantitative data analysis.

Research not only underscores the remarkable potential of AI in optimising social services but also highlights challenges such as data access, privacy concerns, and governance issues. Findings reveal that the integration of AI in the TSP has led to substantial financial savings over the past decade. While the TSP initially succeeded in reducing poverty and narrowing the wealth gap, it ultimately fell short of its primary objective due to various factors, including economic instability. AI-driven algorithms have enhanced the accuracy and fairness of household eligibility assessments. Furthermore, the study demonstrates a remarkably high level of public acceptance (87%) and trust (79%) in AI's role within the TSP. In conclusion, the study showcases how AI has become a transformative force in data-driven policy-making within Iran's TSP.

India

Chapter 8 'Social Assessment and Cultural Resistance: The Public Distribution System in Tamil Nadu, India' is contributed by Sumathi Srinivasalu, Manjubarkavi Selladurai, Shobana Sharma, Gunanithi Perumal, Muniraj Mathaiyan, Ashly Ann Jo, and Ebin Deni Raj. The chapter describes and analyses the technology-based social assessment process in the Indian Public Distribution System (PDS) of Tamil Nadu and its impact on vulnerable communities. After introducing the history, characteristics, and policies of the PDS, the chapter presents its current actors, stakeholders, and the processes involved, especially those of socially assessing beneficiaries for receiving rations and the role of technology within. It investigates the performance, characteristics, gaps, and barriers of the existing PDS in the high-performing state Tamil Nadu, the role 'cultural resistance' plays in everyday social assessment practices in the so-called 'Fair Price Shops', the effects of the current practices on vulnerable groups, and how these groups envisage improvements for a more desired system. The Fair Price Shops are supported by a technical infrastructure of biometric databases and distribution algorithms that assign rations to beneficiaries. The chapter uses two data collection approaches to address its research questions: A quantitative household analysis focussing on household demographics, ration utilisation, savings due to rationed items, and overall satisfaction with the PDS; and a Safe Spaces workshop with participants from vulnerable communities to gather qualitative data about the role of culture and context in daily assessment practices. Main results show that the PDS operates within a rich tapestry of cultural and contextual intricacies, profoundly shaping its effectiveness and impact. The use of biometric data and distribution algorithms can benefit from insights into these socio-cultural dynamics: Decision-makers have to take them into account for system improvements.

Nigeria

Chapter 9's author **Emmanuel Ejim-Eze** writes about 'The role of AI in effective social protection delivery—The National Cash Transfer Programme in Nigeria'. Nigerian Social Protection (SP) coverage is limited by booming population, political economy, socio-cultural structure, corruption, infrastructure, etc. Research obtained data from interviews and focus group discussions with SP experts to provide an understanding of context dependencies of actors within the Nigerian SP environment. The research strategy adopted made room for understanding social, economic, political, and environmental pressures that drove the discourse on social assessment of technology use in SP in Nigeria.

Several literature bodies were reviewed and used as sources of secondary data analysis on supply and demand sides of Nigerian SP. Triangulation of data helped to create robust data for research. Findings show that AI and related technologies

were useful in producing high resolution poverty maps for both Nigerian urban and rural areas improving the targeting of social safety net programme (SSNEP) beneficiaries. General impact assessment dominates genuine social assessment of SP and limits development of robust SP interventions in Nigeria. Future use of AI should extend the technology beyond targeting beneficiaries to include other parts of SSNEP delivery chain.

Ukraine

Chapter 10's author, Oleksandr Khyzhniak, describes 'The Potential of Artificial Intelligence in the Assessment of System of Social Integration of Veterans of the Russian-Ukrainian War'. War veterans in modern Ukraine are a particularly socially vulnerable group whose social integration requires new approaches, including the involvement of contemporary information and communication technologies and artificial intelligence (AI). Based on the use of 'Diia', an AI online application that helps Ukrainians receive social services, the study explores the potential of AI in social work with war veterans and the conditions for its effective application. The author identifies three areas of social work with war veterans: social rehabilitation, social integration, and social assistance. Indicators to measure this potential have been developed, which include the ability of the social work system to remove outdated approaches, forms, and types that cannot be digitised; social workers' proficiency in digital technologies; the ability to implement these technologies in service delivery; and maintaining constructive social work through controlled AI application by professional staff (primarily state and municipal services). Our methods included Participatory Systems Mapping and gamification while working with Ukrainian war refugees. Ukraine is becoming a country of veterans due to prolonged war. It has been proved that the digitalisation of veteran services and the use of AI in social work with veterans should be controlled processes with a transparent algorithm and control system.

China

Chapter 11 by **Wu Qi and Li Hui is** about 'Assessment for AI in Social Services: Community Virtual Nursing Home in Shanghai, China'. It assesses the impact of introducing an AI system for increasing the efficiency of service distribution and the quality of life of the elderly in a community virtual nursing home in Shanghai, China. Using data from the Xinhong Street Virtual Nursing Home in Minhang District, the authors employ a difference-in-differences approach to analyse the changes in operational efficiency and quality of life indicators before and after the AI system was implemented. The results show that the AI system significantly improved operational efficiency and the physical health of the elderly. However,

there was no significant improvement in mental health. The study also found that the AI system shifted the distribution of care services from proactive to responsive, leading to a change in care patterns. Overall, the study highlights the benefits of AI systems in improving service efficiency and health outcomes in elderly care, while also raising questions about equitable service distribution and the need for the human touch in care services.

USA

Chapter 12 by **Margaret Hinrichs, Jieshu Wang, Caity Roe, and Erik W. Johnston** is titled 'AI Integration in Mental Health Services: Examining Trends in the US and Peoria, Illinois'. In the USA and globally, public provisioning systems are evolving in two fundamental ways. The first is to reorganise from decentralised services to coordination around systems of care. The second is the widespread integration of AI into multiple social service areas including mental health diagnosis, needs assessment, and service delivery. While AI has displayed tremendous potential across various dimensions of mental health, including prediction, monitoring, diagnosis, treatment, and assessment, the use of AI also introduces new challenges to performance and accountabilities. This chapter explores the use of systems of care in Peoria, Illinois, for coordinating public service provisioning across multiple organisations in service of vulnerable populations. Practitioners identified barriers for the public including logistical, social, cultural, and internal organisational challenges. Lessons from the case motivate a broader exploration of the use of AI in public service provisioning in the USA with a deeper dive into the use of AI in the mental health social service area. Concerns and challenges are included to promote a balanced conversation on the opportunities and accountabilities for using AI in public service provisioning. As the use of AI becomes more widespread, continuous interrogation and reflection are necessary to realise the potential of AI consistent with the values of the public service organisations, to be in service of the publics that benefit from these programs, and to minimise unintended consequences.

Before starting with case study chapters, two further conceptual chapters follow to provide more information on the background of research: The next chapter will introduce AI FORA's approach to inclusive technology co-design and review issues of AI governance.

Acknowledgements This funding acknowledgement should be added for each chapter of the book. The chapters can be downloaded individually which means that the funding notice required by the funder would not be noticable otherwise for people not downloading the whole book but only individual chapters.

References

Abe, J., Grills, C., Ghavami, N., Xiong, G., Davis, C., & Johnson, C. (2018). Making the invisible visible: Identifying and articulating culture in practice-based evidence. *American Journal of Community Psychology, 62*(1–2), 121–134. https://doi.org/10.1002/ajcp.12266

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899–917. https://doi.org/10.1037/amp0000190

Ahrweiler, P. (2024). *Angels and other cows. A celestial adventure into AI worlds, the social good, and unknown connections*. Springer.

Ahrweiler, P., Frank, D., & Gilbert, N. (2019). Co-designing social simulation models for policy advice: Lessons learned from the INFSO-SKIN study. In *2019 Spring simulation conference (SpringSim)*. IEEE. Retrieved May 23, 2024, from https://ieeexplore.ieee.org/document/8732901

Ahrweiler, P., Gilbert, N., Bicket, M., Sabater Coll, A., Luque Capellas, B., Wurster, D., Siqueiros J. M., & E. Spaeth (2023). Gamification and simulation for innovation. In: Elsenbroich, C. and H. Verhagen (Eds) *Advances in social simulation. Proceedings of the 18th social simulation conference*, Glasgow, 4–8 September 2023. Springer Proceedings in Complexity. Springer.

Aldewereld, H., Dignum, V., & hua Tan, Y. (2014). Design for values in software development. In J. van den Jeroen, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of ethics, values, and technological design* (pp. 1–12). Springer.

Alesina, A., & Angeletos, G.-M. (2005). Fairness and redistribution. *American Economic Review, 95*, 960–980.

Anderson, B. (1983). *Imagined communities. Reflections on the origin and spread of nationalism*. Verso.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica. Retrieved May 23, 2024, from www.propublica.org/article/machine-bias-risk-assessmentsin-criminal-sentencing

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*. Auerbach Publications.

Arnold, T., & Scheutz, M. (2018). The big red button is too late: An alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology, 20*(1), 59–69. https://doi.org/10.1007/s10676-018-9447-7

Aronson, E., Wilson, T. D., Sommers, S., Page-Gould, E., & Lewis, N. (2023). *Social psychology* (11th ed.). Pearson Education.

Barocas, S., Bradley, E., Honavar, V., & Provost, F. (2017). *Big data, data science, and civil rights*. ArXiv: https://arxiv.org/abs/1706.03102

Beigang, F. (2022). On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making. *Minds and Machines, 32*(4), 655–682. https://doi.org/10.1007/s11023-022-09615-9

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets*. Yale University Press.

Berthoz, A. (2012). *Simplexity. Simplifying principles for a complex world*. Yale University Press.

Beugelsdijk, S., & Welzel, C. (2018). Dimensions and dynamics of national culture: Synthesizing Hofstede with Inglehart. *Journal of Cross-Cultural Psychology, 49*(10), 1469–1505. https://doi.org/10.1177/0022022118798505

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research, 81*, 149–159. Conference on Fairness, Accountability, and Transparency.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT Press.

Burk, D. (2019). Algorithmic fair use. The University of Chicago Law Review. *Symposium: Personalized Law, 86*(2), 283–308.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the Scallops and the Fishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, action and belief. A new sociology of knowledge?* (pp. 196–232). Routledge.

Callon, M., & Latour, B. (1992). Don't throw the Baby out with the Bath School! A reply to Collins and Yearly. In: Pickering, A. (Hg.): *Science as practice and culture*. University of Chicago Press (pp. 343–368).

Carpenter-Song, E. A., Schwallie, M. N., & Longhofer, J. (2007). Cultural competence reexamined: Critique and directions for the future. *Psychiatric Services, 58*(10), 1362–1365.

Casey, A., & Niblet, A. (2017). The death of rules and standards. *Indiana Law Journal, 92*(4), 1401–1447.

Cefaï, D., Zimmermann, B., Nicolae, S., & Endress, M. (2015). Introduction. Special issue on sociology of valuation and evaluation. *Human Studies, 38*(1), 1–12.

Chen, S. X., Lam, B. C. P., Hui, B. P. H., Ng, J. C. K., Mak, W. W. S., Guan, Y., & Lau, V. C. Y. (2016). Conceptualizing psychological processes in response to globalization: Components, antecedents, and consequences of global orientations. *Journal of Personality and Social Psychology, 110*(2), 302–331. https://doi.org/10.1037/a0039647

Citron, D. K., & Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*, 1–33.

Conceicao, C. P., Åvila, P., Coelho, A. R., & Costa, A. F. (2020). European action plans for science-society relations: Changing buzzwords, changing the agenda. *Minerva, 58*(1), 1–24. https://doi.org/10.1007/s11024-019-09380-7

Corlet Walker, C., Druckman, A., & Jackson, T. (2021). Welfare systems without economic growth: A review of the challenges and next steps for the field. *Ecological Economics, 186*, 107066. https://doi.org/10.1016/j.ecolecon.2021.107066

Danks, D., & London, A. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 4691–4697). IJCAI.

Deville, J., Guggenheim, M., & Hrdličková, Z. (Eds.). (2016). *Practising comparison: Logics, relations, collaborations*. Mattering Press.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism, 3*(3), 398–415.

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*, 1–3.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*, 1–5.

Druckman, J., & Nelson, K. (2003). Framing and deliberation. How citizens' conversations limit elite influence. *American Journal of Political Science, 47*(4), 729–745.

Dubois, V., Paris, M., & Weill, P.-E. (2018). Targeting by numbers. The uses of statistics for monitoring French welfare benefit recipients. In *Creating target publics for welfare policies* (pp. 93–109). Springer.

Edler Duarte, D. (2021). The making of crime predictions: Sociotechnical assemblages and the controversies of governing future crime. *Surveillance and Society, 19*(2), 199–215.

Eisenstadt, S. (2000). Multiple modernities. *Daedalus, 129*(1), 1–30.

Epple, A., Erhardt, W., & Grave, J. (Eds.). (2020). *Practices of comparing: Towards a new understanding of a fundamental human practice*. Bielefeld University Press.

Erumban, A., & de Jong, S. (2006). Cross-country differences in ICT adoption: A consequence of culture? *Journal of World Business, 41*(4), 302–314.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. *St. Martin's Press. Sociologie du travail, 64*(4).

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. Oxford University Press.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Federal Probation, 80*(2), 38–46.

Fourcade, M. (2016). Ordinalization: Lewis A. Coser memorial award for theoretical agenda setting 2014. *Sociological Theory, 34*(3), 175–195.

Fourcade, M., & Healy, K. (2017). Categories all the way down. *Historical Social Research/ Historische Sozialforschung, 42*(1), 286–296.

Gallimore, R., & Goldenberg, C. (2001). Analyzing cultural models and settings to connect minority achievement and school improvement research. *Educational Psychologist, 36*, 45–56.

Gallimore, R., Goldenberg, C. N., & Weisner, T. S. (1993). The social construction and subjective reality of activity settings: Implications for community psychology. *American Journal of Community Psychology, 21*, 537–560.

Garneau, A. B., & Pepin, J. (2015). Cultural competence: A constructivist definition. *Journal of Transcultural Nursing, 26*, 9–15.

Gascoigne, T., Schiele, B., Leach, J., Riedlinger, M., Massarani, L., Lewenstein, B. V., & Broks, P. (Eds.). (2020). *Communicating science: A global perspective*. ANU Press.

Georgas, J., van de Vijver, F. J. R., & Berry, J. W. (2004). The ecocultural framework, ecosocial indices, and psychological variables in cross-cultural research. *Journal of Cross-Cultural Psychology, 35*(1), 74–96. https://doi.org/10.1177/0022022103260459

Gilbert, N., Ahrweiler, P., Barbrook-Johnson, P., Narasimhan, K., & Wilkinson, H. (2018). Computational modelling of public policy: Reflections on practice. *Journal of Artificial Societies and Social Simulation (JASSS), 21*(1), 14. https://doi.org/10.18564/jasss.3669

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron*. MIT Press.

Gregory, A., Skiba, R. J., & Noguera, P. A. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational Researcher, 39*, 59–68.

Grgić-Hlača, N. et al. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk perception. *27th World Wide Web (WWW) Conference in Lyon*, 23–27th of April, 2018 (pp. 903–912).

Grunwald, A. (2013). Techno-visionary sciences: Challenges to policy advice. *Science, Technology and Innovation Studies, 9*(2), 21–38.

Hacking, I. (1990). *The taming of chance*. Cambridge University Press.

Hank, K., & M. Erlinghagen (2009) *Perceptions of job security in Europe's ageing workforce*. Retrieved 23 May, 2024, from https://papers.ssrn.com/abstract=1444357

Haraway, D. (1995). *Die Neuerfindung der Natur. Primaten, Cyborgs und Frauen*. Campus.

He, M., & Lee, J. (2020). Social culture and innovation diffusion: A theoretically founded agent-based model. *Journal of Evolutionary Economics, 30*, 1109–1149. https://doi.org/10.1007/s00191-020-00665-9

Heintz, B. (2016). "Wir leben im Zeitalter der Vergleichung." Perspektiven einer Soziologie des Vergleichs. *Zeitschrift für Soziologie, 39*, 162–181.

Heintz, B. (2021). Kategorisieren, Vergleichen, Bewerten und Quantifizieren im Spiegel sozialer Beobachtungsformate. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 73*(S1), 5–47.

Hoerning, K. H. (1989). Vom Umgang mit den Dingen – Eine techniksoziologische Zuspitzung. In P. Weingart (Ed.), *Technik als sozialer Prozeß* (pp. 90–127). Suhrkamp.

Hofstede, G. (1980). *Culture's consequences: International differences in work related values*. Sage.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Sage.

Hofstede, G. (2003). *Culture's Consequences: Comparing Values, behaviours, institutions, and organizations across nations*. Sage.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind* (3rd ed.). McGraw Hill.

Hofstede, G. J., Franco, E., Damen, F., & Fogliano, V. (2021). *Healthy snacks from mom? An agent-based model of snackification in three countries*. doi:https://doi.org/10.1007/978-3-030-61503-1_41.

Humm, C., & Schrögel, P. (2020). Science for all? Practical recommendations on reaching underserved audiences. *Frontiers in Communication – Science and Environmental Communication*.

Inglehart, R. (1971). The silent revolution in Europe: Intergenerational change in post-industrial societies. *American Political Science Review, 65*, 995–1017.

Inglehart, R. (1990). *Culture shift in advanced industrial society*. Princeton University Press.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton University Press.

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

Kaasa, A. (2021). Merging Hofstede, Schwartz, & Inglehart into a single system. *Journal of Cross-Cultural Psychology, 52*(4), 339–353.

Kahneman, D., Sibony, O., & Sustein, C. R. (2021). *Noise: A flaw in human judgement*. Brown Spark.

Karpik, L. (2010). *Valuing the unique: The economics of singularities*. Princeton University Press.

Kirkman, B. L., Lowe, K. B., & Gibson, C. B. (2006). A quarter century of "culture's consequences": A review of empirical research incorporating Hofstede's cultural values framework. *Journal of International Business Studies, 37*, 285–320.

Kleinman, A. (1981). On illness meanings and clinical interpretation: Not 'rational man,' but a rational approach to man the sufferer/man the healer. *Culture, Medicine and Psychiatry, 5*, 373–377.

Kleinman, A. (1988). *The illness narratives: Suffering, healing, and the human condition*. Basic Books.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016): Inherent trade-offs in the fair determination of risk scores; https://arxiv.org/abs/1609.05807v2

Knorr-Cetina, K. (2002). *Wissenskulturen*. Ein Vergleich naturwissenschaftlicher Wissensformen.

Koebis, N., Starke, C., & Rahwan, I. (2022). The promise and perils of using artificial intelligence to fight corruption. *Nature Machine Intelligence, 4*, 418–424.

Koo, M., Choi, J. A., & Choi, I. (2018). Analytic versus holistic cognition: Constructs and measurement. In J. Spencer-Rodgers & K. Peng (Eds.), *The psychological and cultural foundations of east Asian cognition: Contradiction, change, and holism* (pp. 105–134). Oxford University Press. https://doi.org/10.1093/oso/9780199348541.003.0004

Kornberger, M., Justesen, L., Koed Madsen, A., & Mouritsen, J. (Eds.). (2015). *Making things valuable*. Oxford University Press.

Kreiss, D., Finn, M., & Turner, F. (2011). The limits of peer production: Some reminders from Max Weber for the network society. *New Media & Society, 13*(2), 243–259.

Krueger, A. K., & Reinhart, M. (2017). Wert, Werte und (Be)Wertungen. Eine erste begriffs- und prozesstheoretische Sondierung der aktuellen Soziologie der Bewertung. *Berliner Journal für Soziologie, 26*(3–4), 485–500.

Kuang, C. (November 21, 2017). Can A.I. be taught to explain itself? *The New York Times*. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Lamont, M. (2012). Toward a comparative sociology of valuation and evaluation. *Annual Review of Sociology, 38*, 201–221.

Lyon, D. (Ed.). (2003). *Surveillance as social sorting: Privacy, risk, and digital discrimination*. Routledge.

Maleki, A., & de Jong, M. (2014). A proposal for clustering the dimensions of national culture. *Cross Cultural Research, 48*, 107–143.

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law and Security Review, 34*(4), 754–772.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

Masuda, T., & Nisbett, R. E. (2006). Culture and change blindness. *Cognitive Science, 30*, 381–399.

Mehrabi, M., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computer Survey, 54*(6), 1–35.

Metcalfe, J., Gascoigne, T., Medvecky, F., & Nepote, A. C. (2022). Participatory science communication for transformation. *Journal of Science Communication, 21*(2), E.

Miller, J. (2002). Bringing culture to basic psychology theory—beyond individualism and collectivism: comment on Oyserman et al. *Psychological Bulletin, 128*(1), 97–109. https://doi.org/10.1037/0033-2909.128.1.97

Minkov, M., & Hofstede, G. (2012). Hofstede's fifth dimension: New evidence from the World Values Survey. *Journal of Cross-Cultural Psychology, 41*, 99–108.

Minkov, M., & Kaasa, A. (2021). A test of Hofstede's model of culture following his own approach. *Cross Cultural and Strategic Management, 28*(2). https://doi.org/10.1108/CCSM-05-2020-0120

Mitchell, S., Potash, E., Barocs, S., D'Amur, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions and definitions. *Annual Review of Statistics and Its Applications, 8*, 141–163.

Muellerleile, C., & Robertson, S. (2018). Digital Weberianism: Bureaucracy, information, and techno-rationality of neoliberal Capitalism. *Indiana Journal of Global Legal Studies, 25*(1), 187–216.

Muniesa, F., & Helgesson, C.-F. (2013). Valuation studies and the spectacle of valuation. *Valuation Studies, 1*(2), 119–123.

Nardon, L., & Steers, R. M. (2009). The culture theory jungle: Divergence and convergence in models of national culture. In R. S. Bhagat & R. M. Steers (Eds.), *Cambridge handbook of culture, organizations, and work* (pp. 3–22). Cambridge University Press.

Navis, C., & Glynn, M. A. (2010). How new market categories emerge. Temporal dynamics of legitimacy, identity and entrepreneurship in satellite radio, 1990–2005. *Administrative Science Quarterly, 55*, 439–471.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datafication'. *The Journal of Strategic Information Systems, 24*(1), 3–14.

Odilla, F. (2023). Unfairness in AL anti-corruption tools: Main drivers and consequences. In *The ECPR general conference, 2023*. Carles University.

Orseau, L., & Armstrong, S. (2016). Safely interruptible agents. *Proceedings of the thirty-second uncertainty in artificial intelligence conference*. ACM Digital Library. Retrieved 23 May, 2024, from https://dl.acm.org/doi/10.5555/3020948.3021006#sec-terms

Oyserman, D., Coon, H., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analysis. *Psychological Bulletin, 128*(1), 3–72.

Pagano, T., Loureiro, R., Lisboa, F., Peixoto, R., Guimarães, G., Cruz, G., Araujo, M., Santos, L., Cruz, M., & Oliveira, E. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing, 7*(1), 15.

Parker, P. E., Fleming, P. D., Beyerlein, S., Apple, D., & Krumsieg, K. (2001). Differentiating assessment from evaluation as continuous improvement tools [for engineering education]. *31st Annual frontiers in education conference. Impact on engineering and science education. Conference proceedings (Cat. No.01CH37193)*, T3A-1. doi:https://doi.org/10.1109/FIE.2001.963901.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Peeters, R., & Schuilenburg, M. (2018). Machine Justice: Governing security through the bureaucracy of algorithms. *Information Polity, 23*(3), 267–280.

Picot, G., & Tassinari, A. (2017). All of one kind? Labour market reforms under austerity in Italy and Spain. *Socio-Economic Review, 15*, 461–482.

Porter, T. (1995). Trust in numbers. The pursuit of objectivity in science and public life. .

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14.

Riterich, A. (2018). Big data. Ethical debates. In A. Richertich (Ed.), *The big data agenda: Data ethics and critical data studies* (pp. 33–52). University of Westminster Press.

Sarason, S. (1972). *The creation of settings and the future societies*. Jossey-Bass.

Schaefer, M., Haun, D. B. M., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological Science, 26*, 1252–1260.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–66). Academic Press.

Schwartz, S. H. (2004). Mapping and interpreting cultural differences around the world. In H. Vinken, J. Soeters, & P. Ester (Eds.), *Comparing cultures: Dimensions of culture in a comparative perspective* (pp. 43–73). Brill.

Schwartz, S. H. (2008). *Cultural value orientations: Nature & implications of national differences*. Publishing House of SU, Higher School of Economics.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture, 2*(1). https://doi.org/10.9707/2307-0919.1116

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicolas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scorping a taxonomy for harm reduction. In F. Rossi, S. Das, J. Davies, K. Firth-Butterfield, & A. John (Eds.), *Proceedings of the 2023 AAAI/ACM conference on AI, ethics, and society* (pp. 723–741). AIES.

Sheridan, T. B. (2006). Supervisory control. In *Handbook of human factors and ergonomics* (3rd ed., pp. 1025–1052). Wiley.

Silva, S., & Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon, 55*(1–2), 9–37.

Singelis, T. M., Triandis, H. C., Bhawuk, D. P. S., & Gelfand, M. (1995). Horizontal and vertical dimensions of individualism-collectivism: A theoretical and measurement refinement. *Cross Cultural Research, 29*, 240–275.

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In A. Teredesai & V. Kumar (Eds.), *KDD '19: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2459–2468). Association for Computing Machinery.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data and Society, 9*(2).

Steinmetz, W. (Ed.). (2019). *The force of comparison. A new perspective on modern European history and the contemporary world*. Berghahn Books.

Suresh, H., & Guttag, V. (2019). *A framework for understanding unintended consequences of machine learning*. CoRR abs/1901.10002. Retrieved May 23, 2024, from https://dblp.org/rec/journals/corr/abs-1901-10002.html

Taylor-Gooby, P., & Martin, R. (2010). Fairness, equality and legitimacy: A qualitative comparative study of Germany and the UK. *Social Policy and Administration, 44*, 85–103.

Triandis, H. C., & Gelfand, M. J. (2012). A theory of individualism and collectivism. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 498–520). Sage.

Triandis, H. C., Bontempo, R., Villareal, M. J., Asai, M., & Lucca, N. (1988). Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships. *Journal of Personality and Social Psychology, 54*, 323–338.

Triebe, B. (2014). *Der Nationalstaat als sozialwissenschaftliche Denkkategorie: Eine Analyse des methodologischen Nationalismus*. Tectum Verlag.

Tzimas, T. (2021). *Legal and ethical challenges of artificial intelligence from an international law perspective*. Springer.

Van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security, 23*(4), 323–340.

Van Damme, E., van der Wal, N., Hofstede, G. J., & Brazier, F. (2023). *The influence of national culture on evacuation response behaviour and time: An agent-based approach*. doi:https://doi.org/10.1007/978-3-031-22947-3_4.

Venkatapuram, S., Ehni, H.-J., & Saxena, A. (2017). Equity and healthy ageing. *Bulletin of the World Health Organization, 95*, 791–792.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Weber, M. (1922). *Wirtschaft und Gesellschaft*. Mohr.

Weiss, M. (1997). Explanatory Model Interview Catalogue (EMIC): Framework for comparative study of illness. *Transcultural Psychiatry, 34*, 235–263.

Welzel, C. (2013). *Freedom rising: Human empowerment and the quest for emancipation*. Cambridge University Press.

Winzar, H. (2015). The ecological fallacy: How to spot one and tips on how to use one to your advantage. *Australasian Marketing Journal, 23*(1), 86–92. https://doi.org/10.1016/j.ausmj.2014.12.002

Wong, P.-H. (2020). Cultural differences as excuses? Human rights and cultural values in global ethics and governance of AI. *Philosophy and Technology, 33*, 705–715. https://doi.org/10.1007/s133347-020-00413-8

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Science, 112*(4), 1036–1040.

Zarsky, T. (2015). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology and Human Values, 41*(1), 118–132.

Chapter 2
Inclusive Technology Co-design for Participatory AI

Petra Ahrweiler, Elisabeth Späth, Jesús M. Siqueiros García, Blanca Luque Capellas, and David Wurster

Abstract This chapter reviews existing initiatives to include societal perspectives in AI governance and technology design and introduces the 'Artificial Intelligence for Assessment' (AI FORA) approach applied to AI use in public social service provision. The chapter starts with reviewing contemporary AI governance frameworks which still need to be translated into multi-stakeholder governance and inclusive technology co-design. For this, the emerging field of 'Participatory AI' seems to bear promise. After identifying and discussing the participatory requirements for inclusive technology co-design, especially related to the safe and effective participation of vulnerable groups, the chapter introduces the AI FORA approach. The participatory AI FORA approach starts with the assumption that the gap between technology and society, in this case the disconnect of dynamic cultural values from AI-based social assessment, leads to fairness issues of existing systems. To connect cultural values to technology production for more desirable systems, society, i.e. all societal groups stakeholding in this area of technological innovation, needs to get involved in technology production and policy. The chapter presents the participatory research methods AI FORA employs to achieve inclusive technology co-design around the project's 'Safe Spaces' concept that ensures equitable participation of stakeholders in AI-based social assessment for public service provision. The chapter ends with a reflection on the claims of inclusive technology co-design, the consequences for related science communication in AI, and the impacts on AI policy and governance.

P. Ahrweiler (✉) · B. L. Capellas · D. Wurster
Department of Technology- and Innovation Sociology, Social Simulation (TISSS Lab), Institute of Sociology, Johannes Gutenberg-University Mainz, Mainz, Germany
e-mail: petra.ahrweiler@uni-mainz.de; bluqueca@uni-mainz.de; dwurster@uni-mainz.de

E. Späth
TISSS Lab, Institute of Sociology, Johannes Gutenberg University Mainz, Mainz, Germany
e-mail: espaeth@uni-mainz.de

J. M. Siqueiros García
IIMAS Unidad Mérida, Universidad Nacional Autónoma de México, Ucú, Yucatán, México
e-mail: jmario.siqueiros@iimas.unam.mx

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_2

Introduction

This chapter reviews existing initiatives to include societal perspectives in AI governance and technology design and introduces the 'Artificial Intelligence for Assessment' (AI FORA) approach applied to AI use in public social service provision.

AI governance is a recent phenomenon. The chapter starts with a review of existing initiatives, which have been developed mainly during the past 5 years at the national, European, and international level. Most of these initiatives try to deal with the challenge of bridging the gap between the computer science involved in expert-driven AI governance and 'the user', 'the society', or 'the public'. This follows the demands of non-governmental and non-technical societal actors for participation: 'The often controversial nature of scientific-technological developments has led to attempts on the part of state institutions, policy-makers as well as civil society actors to innovate in new forms of governance, with emphasis on stakeholder and citizen participation' (European Commission, 2007: 43). Gianni and Goujon (2014) distinguish four models of innovation governance: The Standard Model based on the principle of top-down governance of experts and the principle of 'irrationality' of non-experts, the Revised Standard Model that also includes the political mediation especially regarding risks, the Democratic-Inclusive Model where the principle of irrationality is dropped and substituted by the principles of 'industrial biasing' and 'public biasing', and the Co-Constructive Model where the principle of top-down governance is dropped and substituted by the principle of public co-creation. Most of the AI governance frameworks presented below remain at the high level of policy statements rather than being translated and implemented through hands-on technology co-design as postulated by the co-constructive model. Of course, the latter is not a straightforward task, and new territory has to be accessed. The emerging field of 'Participatory AI', which is reviewed in the following chapter, starts to explore this new territory.

Presenting the AI FORA approach, the chapter continues with identifying and discussing the requirements for inclusive technology co-design using as example the topic of this book, AI-based public social service provision. As this is an area concerned with social justice and compensation for the needy, one major, but highly contested requirement is the inclusion and empowerment of vulnerable groups in the innovation process. The chapter reviews existing literature on innovation with vulnerable groups and identifies requirements for their safe and effective participation. One of them is to provide 'Safe Spaces' ensuring equitable interaction for inclusive technology co-design. The chapter discusses existing set-ups and methods for suchlike infrastructures and presents the Safe Spaces framework that was developed for the AI FORA approach and used by the case studies presented in this volume, and reports on experiences and feedback.

The chapter ends with a reflection on the claims of inclusive technology co-design, the consequences for related science communication in AI, and the impacts on AI policy and governance.

AI Governance: Review of Existing Concepts

Current AI governance frameworks have a strong focus on ethics. They provide guidelines to ensure beneficial impacts (cf. Vinuesa et al., 2020) using ethical principles such as trustworthiness, justice, fairness, non-maleficence, and responsibility (cf. Jobin et al., 2019). Initiatives can be identified at the national, European, and international levels, e.g., for the national level the recent statement of Deutscher Ethikrat on Ethics of AI (Deutscher Ethikrat, 2023), for the international realm the IEEE Global Initiative of Ethics of Autonomous and Intelligent Systems (Huang et al., 2023) or UNESCO's 'Recommendation on the Ethics of Artificial Intelligence' (UNESCO, 2022), and as example for the European context 'Regulation laying down harmonized rules on artificial intelligence' (European Commission, 2021). Critics argue that the ethical focus is too principle-oriented thus creating a gap between theory and practice (cf. Bleher & Braun, 2023).

Aiming to fill this gap, the 'European Union's Ethics Guidelines for Trustworthy AI' (European Commission, 2019) and the 'Montreal Declaration for Responsible AI' (University of Montreal, 2018; Díaz-Rodríguez et al., 2023) are important governance frameworks attempting to translate ethical principles into more operational policies, especially for AI-using organisations.

Responsible AI is based on the three principles of auditability, accountability, and liability (cf. Díaz-Rodríguez et al., 2023; Mikalef et al., 2022) focussing on 'how a specific organization is addressing the challenges around AI from both an ethical and legal point of view' (Anagnostou et al., 2022: 1). The more general approach of Trustworthy AI 'contains responsible AI and extends it towards considering other requirements that contribute to the generation of trust in the system' (Díaz-Rodríguez et al., 2023: 14), requiring human agency and oversight, technical robustness, privacy, transparency, societal- and environmental well-being, and non-discrimination. An example of connecting the responsible and trustworthy AI aspects is the Accountability, Responsibility, Transparency (ART) framework (Dignum, 2021). Transparency of opaque AI decision systems is an elementary requirement when people receive decisions about issues of high importance for them, e.g. the provision or rejection of social services according to AI-based social assessment that has been employed by public agencies. The organisation responsible for the decision making needs to reveal details on the performance of systems that are used although this might lead to the so-called 'AI transparency paradox' where even the most truly intentions towards transparency entail unintended and unfavourable effects (cf. Larsson & Heintz, 2020; Hansen et al., 2015). For example, sometimes the information provided to the public is not understandable (Hansen et al., 2015), or it can also happen that making public certain processes leads to mistrust (Tsoukas, 1997, as cited in Hansen et al., 2015; Eisenberg, 2007, as cited in Hansen et al., 2015). Transparent AI starts with information on sources and training data but also includes explanations of the rationale behind decisions within the system. Computational decision systems using AI techniques such as support vector machines, random forests, probabilistic graphical models, reinforcement learning,

or deep learning neural networks (cf. Arrieta et al., 2020) challenge requirements of explainability and interpretability. These challenges are in the interest of users who need to understand the rationale and process behind automated decision making and of experts who need to interpret and confirm the validity of decision outputs (cf. Linardatos et al., 2021: 1).

It is difficult to understand, explain, and interpret what the AI system 'is actually doing'. Explanations (e.g. via the interface) are required that make the systems' behaviours more intelligible and accessible.

Explainable AI (XAI) is defined as a concept for designing AI systems in a way that their processes and decisions can be understood and interpreted by humans (cf. Gunning et al., 2019; Adadi & Berrada, 2018). Focussing on explainability and interpretability, environments for justifying, controlling, improving, and discovering AI systems are created that support humans in understanding AI techniques such as machine learning and emerging patterns as their result (cf. Arrieta et al., 2020). For AI use in social assessment, 'understanding more about system behaviour provides greater visibility over unknown vulnerabilities and flaws (e.g. bias and discrimination)' (Adadi & Berrada, 2018: 52143), and, therefore, can help to improve fairness. However, empirical decision making differs from computational processes, especially in social assessment where values are negotiated and at stake. Though it is important to understand, explain, and interpret what AI systems are doing, it is equally important to be aware of the cognitive, emotional, social, and cultural factors that influence human decision making in the empirical social realm (cf. Schmid & Wrede, 2022). Social assessment is a so-called 'wicked problem', i.e. a problem that is complex and does not have a single correct solution. To avoid that automated decisions are contested, some authors in XAI suggest that AI systems information should be used to co-create decisions with different societal actors (cf. de Bruijn et al., 2022).

As an example for the national level trying to provide hands-on AI governance in risk management, the National Institute of Standards and Technology (NIST), an institution within the Department of Commerce of the United States of America, developed the NIST AI Risk Management Framework which 'is designed to equip organizations and individuals (...) with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time' (https://www.nist.gov/itl/ai-risk-management-framework, accessed 25.09.2023). It includes four main principles: Govern, Map, Measure, and Manage. The first principle, Govern, is about incorporating policies, processes, structures, and corporate culture and connecting 'technical aspects of AI Systems with organizational values' (NIST AI RMF, 2023: 27). The second principle, Map, includes, for example, contextual analysis, a categorisation of the AI system, as well as mapping out targeted usage and goals. These analyses will support the organisation by finding blind spots and risks during the lifecycle of an AI system. The third principle, Measure, uses quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyse, assess, benchmark, and monitor AI risk and related impacts. It includes tracking key performance indicators and metrics for things like social impact and human-AI configurations (NIST AI RMF,

2023: 28). The Manage principle allocates risk resources to risks, mapped out in the Map principle. It 'comprises plans to respond to, recover from, and communicate about incidents' (NIST AI RMF, 2023: 31). At the national level, we can find different commitments with the ethical, transparent, explainable, trustworthy, and responsible AI existing frameworks. However, mostly, attempts remain in the field of goodwill; the next step, which would be to legislate for the development of AI in each country or internationally, is not yet achieved (cf. Hagendorff, 2020; Mittelstadt, 2019; Morley et al., 2020; Mäntymäki et al., 2022). Most European countries have waited for the EU AI Act, which was approved by the European Council of the EU in May 2024 and is expected to be published in June/July 2024: 2 years after its entry into force, this new regulation will apply (European Council, 2024).

This section introduced forms of control, coordination and steering of AI that try to bridge the gap between technology policies driven by experts and governance interests of other groups of society. However, the co-constructive model of governance where the principle of public co-creation is dominant for creating policies is not reached by any of them. Furthermore, the call for public co-construction is not limited to co-creating policies governing technology but extends to co-creating the technology itself.

It is policy itself asking for societal participation in the actual technological innovation process, e.g. the Responsible Research and Innovation (RRI) concept promoted by the European Commission: 'Responsible research and innovation is an approach that anticipates and assesses potential implications and societal expectations with regard to research and innovation, with the aim to foster the design of inclusive and sustainable research and innovation.

Responsible Research and Innovation (RRI) implies that societal actors (researchers, citizens, policy makers, business, third sector organisations, etc.) work together during the whole research and innovation process in order to better align both the process and its outcomes with the values, needs, and expectations of society'. (European Commission, 2017). The call goes for participating in the process of technological innovation from start to implementation. The next section will report on initiatives and concepts that not only ask for societal co-design in policy but in technology production itself.

The Participatory Challenge

The inclusion and involvement of society to solve complex social problems is often attributed to the Scandinavian school of design dated back to the 1970s (Halskov & Hansen, 2015; Kuhn, 1996; Liao & Muller, 2019; Pollini & Caforio, 2021; Steen, 2013; Zamenopoulos & Alexiou, 2018). Participatory action research stemming from different traditions postulates that social actors, especially the vulnerable, must be included in the search for solutions to problems of social concern (Lewin, 1946; Fals-Borda, 1987; Cacari-Stone et al., 2014; Cornish et al., 2023). In a democratic spirit, all participants are legitimate subjects of knowledge production and

should equally be able to participate and contribute. They need to engage in problem definition, data collection, analysis, modelling, and dissemination (Cacari-Stone et al., 2014: 1615). Also, participants must agree about the reach and limitations of the outcomes and their associated risks. In management science, the approach of value co-creation is more focused on business opportunities: 'Co-creation is joint creation and evolution of value with stakeholding individuals, intensified and enacted through platforms of engagement, virtualized and emergent from ecosystems of capabilities, and actualized and embodied in domains of experiences, expanding wealth-welfare-wellbeing' (Ramaswamy & Ozcan, 2014).

Participatory Technology Production

The quest for participatory engagement also applies to the field of technology production and technological innovation. Complex technical innovation issues with high ethical and societal implications for the present and future of our societies such as improving AI use in social service provision cannot be solved and decided by an individual or one subsystem of society alone. Technical solutions require the expertise, participation, and design capacities of all societal groups.

Participatory approaches are becoming more and more common, because they reflect the requirements for collaboration in technology production referenced by concepts such as 'Mode 2' (cf. Gibbons et al., 2010), 'Triple Helix' (cf. Etzkowitz & Leydesdorff, 1998), 'Post-normal science' (c.f. Funtowicz & Ravetz, 1993), and 'Innovation Networks' (cf. Ahrweiler, 2010).

The Need to Include Vulnerable Groups

The particular needs of vulnerable groups are usually not considered in the innovation process, even if efforts are undertaken to involve stakeholders from various backgrounds. Mostly, the 'usual suspects' are considered when it comes to participatory approaches, i.e. professional experts from science, policy, and industry. Exclusion of vulnerable groups can take different forms. For example, a group is excluded because it is invisible to innovators, although such a group is affected by the product developed, usually in unpredictable ways. It may also be the case that design is based on a preconception of the needs of a group. In this case, there is a mismatch between the needs of the vulnerable population and what the expert developers consider to be their needs.

Very often, these groups do not have the social, economic, or political capital to make their voices heard. Even worse is that technological innovation may increase the gap separating them from the benefits of technology, deepening their vulnerable condition (Brown, 2019; Fazelpour & Danks, 2021; Lewis et al., 2012; Pérez-Escolar & Canet, 2022). Vulnerability may be attributed to many different causes,

and these groups may include elders (Bischof & Jarke, 2021; Osborne et al., 2022); migrants (Bustamante Duarte et al., 2018); or historically neglected indigenous communities (Lewis et al., 2012). Technology exclusion can happen in two ways, and one does not exclude the other: One is about access to services through technological devices. In this case, communities may be unable to access and operate technology, and services are simply out of reach. Research on interfaces and user experience is essential here (Benjamin et al., 2020; Bischof & Jarke, 2021). The second is about the technologies for decision making and who will be granted the benefits of services (Caforio et al., 2021; Fazelpour & Danks, 2021). This second aspect is an even more pressing problem with the emergence and incorporation of AI systems and tools by governments.

Participatory and inclusive research with vulnerable and marginalised groups represents a series of specific challenges and particular ethical and methodological considerations, which in these cases often go hand in hand. A major challenge is access to groups in these conditions when they are generally socially invisible (Amann & Sleigh, 2021). In terms of participation, the conditions for equal engagement must be in place.

Furthermore, power asymmetries within and outside the group, which may silence some voices and impose others, must be acknowledged and dealt with. It is important to establish the rules of coexistence and participation, the language to use and the terms of communication.

General Pitfalls of Participatory Approaches

However, participatory approaches in general, and for the development and co-design of technology and innovation, are not without their critics. Of course, including societal groups that are non-scientists and have many different interests, knowledge profiles, and resources, adds a great deal of complexity and risk of failure: Multi-stakeholder contexts need to negotiate heterogeneous interests in conflict-prone discussion arenas with high uncertainty requiring many loops, de-briefings, and societal reflections.

Other important problems are of a methodological and epistemic nature with regard to the criteria for the validity of the knowledge generated and the integration of different forms of knowledge (Durán & Pirtle, 2020). Contributions from different perspectives means different quality criteria for assessing the technology produced. Technical correctness might suffer from competing validity claims.

Furthermore, participatory design and innovation is a practice mostly conceived in the context of countries in the global North, but not that much in the global South, and there are particularities that cannot be ignored such as socio-cultural barriers, power asymmetries within the participating groups, gender differences, rules of interaction, etc. (Hirom et al., 2017). Therefore, chances to benefit from participatory approaches are unevenly distributed. There are certain traps to avoid—pitfalls that can best be summarised under the heading of ‘participation washing’. There is

a risk that participation just masks uneven power distribution: According to Birhane et al. (2022), colonial projects often claimed their legitimacy under the veneer of participation. Instead, participation should empower participants to make a difference and contribute to change.

Participatory AI

In computer science, 'technology co-design' combines software development and the theory of socio-technical systems: It suggests 'that technologies and work practices are best co-designed using participatory methods in the workplace setting, drawing on such common-sense guiding principles as staff being able to access and control the resources they need to do their jobs and insisting that processes should be minimally specified to support adaptive local solutions' (Greenhalgh et al., 2016: 406). According to a review by Birhane et al. (2022: 1), today's AI development is expert-driven, 'characterized currently as technically-focused, representationally imbalanced, and non-participatory'. The relationship between technology production and society is mostly that of designer and user. Especially, AI applications in the public sector are often hidden behind a veil of technical terminology unknown, opaque, and inaccessible to non-experts (AlgorithmWatch, 2020). 'However, the shift away from logic-based AI systems towards more data-driven paradigms such as Deep Learning as well as new infrastructure for capturing and leveraging human-generated data prompted greater demand for 'non-expert' participation in the construction of AI systems' (Birhane et al., 2022: 3).

Birhane et al. (2022) provide an overview on, as they call it, the 'emergent subfield of "Participatory AI"' (Birhane et al., 2022: 4). They also present three case studies where stakeholders were involved in technology development, two on machine translation for African languages and data rights, and one on participatory dataset documentation. The overview shows that there is indeed some work diagnosing participation requirements, specifying the reasons for more stakeholder involvement in AI (e.g. Donia & Shaw, 2021; Iason, 2022) and presenting frameworks, best practice recommendations and guidelines on how to do it (e.g. Lee et al., 2019; Mohamed et al., 2020 and diverse articles in the Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency). Specific pitfalls in participatory approaches to AI innovation are identified:

(a) Especially in data harvesting for AI use in social media with all accompanying issues of privacy and surveillance capitalism, participation is often passive, unintentional, and just 'comes with the technology'. People do not know that they participate. Instead, 'to support the idea of a "participatory condition" of society and technological development requires a degree of agency and intentionality' (Birhane et al., 2022: 3).
(b) Though Birhane et al. (2022) report recent participatory methods in the lifecycle of machine learning including auditing, public consultation such as citizen

juries, joint problem formulation, model cards and datasheets for information disclosures, and even artefact co-development, they voice a suspicion that these methods might merely serve to aid in the refinement of systems from the viewpoint of industry that has mainly asked for public participation in the first place. Instead, the plea goes for participation that ensures empowerment and equity of participants with wider humanitarian or societal benefits.

There are only few empirical examples of real-world case studies of Participatory AI that are published in the academic peer-reviewed literature (cf. van der Veer et al., 2021) and these are quite recent. The majority of references are to arXiv preprints and grey literature. This strengthens the statement that the field of Participatory AI is just slowly emerging. According to Birhane et al. (2022), the reported harms of expert-driven AI systems in many domains have led to a quest for a 'participatory turn' where the designer–user relationship is supposed to change to that of co-designers and co-creators. Among the arXiv articles referenced is promising work on participatory AI design, especially from the perspectives of decolonial studies and relational ethics. Birhane et al. (2022) ask for standards for participatory approaches in AI that include claims for reciprocity, reflexivity, and empowerment. Currently, the composition of stakeholders and the degree of influence they have vary very much between initiatives. Goals, duration, and limitations of the participatory exercise should be clear for participants at all stages of the process. Innovating AI use in public social service provision needs to address a considerable participatory challenge. The previous chapter already pointed to the participatory requirements that stem from cultural diversity and context dependency. These requirements demand participation in the innovation process from a broad range of national welfare systems across cultures. However, the participatory challenge goes deeper than the aggregate level of cultural contexts: It requires one to zoom into the actor level of societies. The next section will introduce the 'Artificial Intelligence for Assessment' approach to inclusive technology co-design developed to address the challenges discussed above and elicit the chances and benefits of participatory approaches in AI innovation.

The AI FORA Approach

It is well-documented that AI technology not only may provide biased information, but also can inadvertently reinforce existing cultural, social, and economic inequalities. Thus, if an AI system is trained on data that reflects unequal access to resources or opportunities, it may further entrench these disparities by providing advantages to dominant groups. This problem is especially important when used in the context of social assessment. 'Problematic applications of machine learning tools in high stakes domains such as criminal justice, healthcare, and hiring have prompted both researchers, civil society, and regulators to increasingly urge greater use participatory methods to mitigate sociotechnical risks not addressed by algorithmic

adjustments or transformations' (Birhane et al., 2022: 3). The question of AI-based social assessment, i.e., who gets what from the state, is a 'high stake domain' as decisions to grant or reject social services often have far-reaching consequences for the individual applicants and demonstrate the value positions of whole societies concerning who is considered as legal, deserving, and needy recipients.

Social assessment for welfare provisions concerns everybody, whether a policy-maker hoping for efficiency and objectivity in allocation, a recipient hoping for support and well-being, a service provider, a taxpayer, or a member of a vulnerable group. Whether the introduction of AI into social assessment makes things better or worse is thus of interest to everyone and makes everyone a potential stakeholder in determining the design of social assessment innovations. An adequate participatory approach needs to involve multiple societal groups co-designing technology for AI-based social assessment. The participatory challenge to ensure a fair distribution of taxpayers' money by technology is to provide a quality space for participation and negotiation on a society's ideas of social justice and fairness, where the diverse voices of all stakeholder communities can impact the shape of technology design. Thus, for realising 'Better AI' in public social service provision, the approach needs to be open, inclusive, participatory, and accessible.

The Innovation Challenge

Social assessment for public social service provision is dependent on cultural context as discussed in the previous chapter. As research in this volume will show, there is a broad spectrum of different cultural value sets leading to different social assessment practices. This can apply to both the entire approach or to inclusion, interpretation, and assessment of individual variables. For example, while most systems evaluate personal characteristics at the time of application for service (e.g. age, employment status, and household composition), in Spain, fairness of assessment means to consider the whole of an applicant's biography, view the current status as a function of it, and evaluate the trajectory rather than their status (cf. Chap. 4 in this volume). Here, fairness means to follow a completely different concept and approach, not looking at average or outlier expressions of variables but at an individual's life course. In India, social service provision is determined by a sophisticated compensation system called 'positive discrimination' which balances personal criteria between constitutional rights and religious caste membership. In China, the social credit system assesses people for their worthiness to receive public goods to the degree they display desirable political, economic, and social behaviour with an educational aim to reward desired and penalise undesired behaviour.

And even in countries that are seemingly similar in value concepts, fairness issues are assessed differently and are subject to turbulent, mostly policy-related change (e.g. concerning granting asylum and further social services for refugees in the countries of the European Union). Moreover, cultural value systems in societies, heterogeneous both in structure and in process, are not static: They are subject to

social change. Societies are developing and, with them, belief and value systems. Accordingly, fairness concepts around social welfare are constantly re-negotiated, which is reflected in societal discourse, in policy reforms, or in new regulations.

Disconnect of Dynamic Cultural Values from AI-Based Social Assessment: Fairness Issues and Consequences for People

Of course, the expectation is that cultural values, i.e. the fairness concepts of different societies, are reflected by and incorporated in the technology that is replacing or assisting human decision making in that cultural context. A disconnect of societal values from technology (see left side of Fig. 2.1) will lead to serious fairness issues breaching cultural routines and practices.

One could argue that AI-based social assessment in different societies would, most probably, be based on 'their' past data and is, therefore, indeed already reflecting 'their' decision making based on 'their' cultural values reproducing 'their' cultural bias and discrimination issues. In this perspective, it seems to be obvious that using a training data set from Ruanda, Africa, would not be helpful for training an algorithm to allocate social services in China, or, in fact anywhere outside Ruanda. However, the few, mostly US-based, companies providing software in this area work with datasets available in their countries. The export of such software as decontextualised technology would mean to apply fairness concepts of IT frontrunner countries to adopter countries is a kind of 'colonial AI'.

AlgorithmWatch (2020) reports on the use of datasets for training algorithms in automated decision making across countries due to missing data availability in the

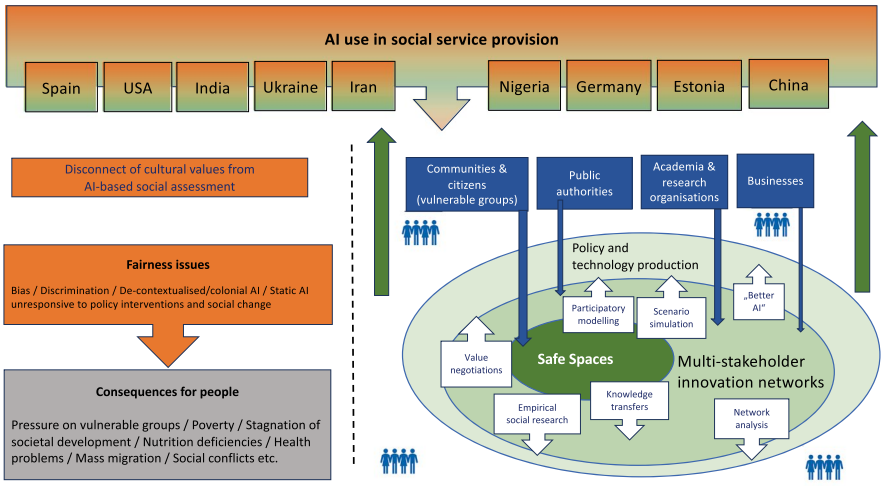

Fig. 2.1 The AI FORA approach

own country. However, responding and respecting different cultural value systems would not allow for such 'de-contextualised' data usage. Fairness issues are literally programmed. In the same way, nobody would come up with the idea that training data from the times of the French Revolution could be used for allocating social goods today. This is not only due to missing availability; it is due to the fact that the obvious misalignment of cultural value systems including ideas of social justice and fairness—these were different times—forbids the very idea. It would simply be inappropriate. However, the assumption that the time lag that is existing now is in any way more acceptable is based on ignorance and misconception about the characteristics and dynamics of social change. Profiling and scoring for social assessment using the current AI techniques is only interesting and useful for stocktaking and as background information for decision making. Not only is data recorded the production of which is based on inputs from a long time ago. The 'time lag' can be sometimes considerable, and the societal value dispositions might have changed slightly or even dramatically in the meantime. Data might be missing for variables which are important now but have not been important in the past and, therefore, not recorded.

The consequences for people are considerable. Reproducing the outcome of the past is not good enough in a societal space where social justice issues, the life chances of vulnerable groups, fundamental ethical principles, and social values are processed, and where value systems that produce fairness concepts are in constant flux; their expressions have high ethical and social implications not only for each individual but for the future of our societies. Policy interventions to increase fairness would run empty, if they were constantly counteracted by algorithms suggesting decisions from the past. The stagnation of societal development could increase pressure on vulnerable groups and make social conflicts become more likely due to misalignment of needs and provisions in important areas such as nutrition, health, or education. Of course, it would be generally possible for AI technology, for example, to evaluate an individual's trajectory through the data space to adapt it to the Spanish case. The cultural specifics of value systems would not be a principal challenge to AI. Furthermore, it would be technologically possible, e.g. by simulation, to create training data responsive to scenarios of social change as potential futures. All in all, it would be hypothetically possible to create contextualised, value-sensitive, responsive, and dynamic AI systems from existing systems that are perceived as problematic. However, this can only be done when there is enough information available on the socio-technical conditions in place—on specifics, commonalities, and differences of existing systems, as well as on the constraints and options for desired scenarios.

(Re-)Connecting Values to Technology: From Existing to Desired Systems

Participants

But who can provide this contextualised data? In the first instance, researchers being familiar with the technical and social characteristics of the systems in place, technology providers who have developed AI systems in use, public authorities that have rolled out the policies to use them, public agencies using them in their daily routines, and, of course, people using their services can provide information on how to improve current systems (see right side of Fig. 2.1).

These are the usual actors of the 'Quadruple Helix' (cf. McAdam & Debackere, 2018): Communities and citizens, public authorities and agencies, academia and research organisations, and businesses. That they are involved is suggested by innovation research and common sense. What about the pitfalls of participation mentioned above? Including societal groups that are non-scientists and have many different interests, knowledge profiles, and resources implies a risk for failure: heterogeneous interests of science, policy, industry, and civil society might escalate conflict rather than lead to co-designed solutions. Is there a credible approach to cooperative and participative network solutions in technology co-design for addressing justice issues around AI-based social service provision while facing plurality of perspectives and, adding to that, complex intercultural context diversity? How can one reconcile the plurality of individual interests and perspectives and mediate intercultural context diversity with high potential for conflict into a co-designed solution?

Innovation as a social phenomenon necessarily criss-crosses numerous functional sub-systems by its very nature (Schumpeter, 1912; Fagerberg et al., 2006): The creation of novelty is provided by the science system as scientific discovery, technological feasibility is contributed by the technological system in practical implementation, commercial realisability is enabled by the economic system; law and the political system provide legitimisation and regulatory certainty. Social adoption and acceptance by the cultural system also belong to innovation as essential characteristics. Innovation is all about stakeholder involvement, cooperation, dynamic networks, and blurring usual boundaries.

Methods

Empirical social research can be used to investigate the actors involved, the societal norms and values that stakeholders use as reference for social assessment routines, the organisational practices and routines in place, and the system's performance. Case-specific detailed system maps about how social assessment routines for distributing social services are conceptualised, organised, and institutionalised will include a policy analysis and a technical analysis for each context. The current state

of the art of technology implementation (databases and systems in place) and its development with reference to international benchmarking needs to be explored.

Methods such as Social Network Analysis can be used to further visualise the structural components and analyse the resource flows (e.g. knowledge, financial capital, social capital, etc.) that are processed in these actor networks (Ahrweiler, 2010). Mapping the structural components of the socio-technical system such as the type of actors involved, resources, technology in place, inputs, outputs, processes, performance, networks, etc. should be complemented by insights into the processes and mechanisms behind them, which requires insights into the behaviours and attitudes of relevant actors (incentives, orientations, norms, values, strategies, intentions, barriers, limitations, visions, options, etc.) and, of course, insights into algorithms and technological processes.

However, it is very important that already the description of 'who is in, and who is out' and 'what is really happening inside' results from a participatory endeavour on a concrete case study level. Participation needs to encompass object formation, problem definition, discussion, and elaboration.

This means that stakeholders as companions of research should be involved from start to finish: They begin by describing and defining what is the case, what is their system, what is their depiction of reality in AI use for public social service provision. Participatory Systems Mapping (Barbrook-Johnson & Penn, 2022) reconstructs structures and processes from the perspective of stakeholders. Researchers familiar with the context and with academic literature about the socio-technical system under study can suggest a first list of actors involved, but it is these actors that should map the system with its processes and identify further actors known by them to have a stake in the area from their knowledge and experience about what is going on at the operational level. This ensures that not only 'the usual suspects' from a research perspective are included in the study but also those who may be 'surprises' to researchers, which will extend learning about multi-stakeholder involvement and engagement.

Interactive and participatory formats at multi-stakeholder workshops can bring forth the culturally shaped and heterogeneous value perspectives of the actor networks involved and host value negotiations. Supporting participative decision making and production of more responsible AI technology adapted to context-specific social value requirements can use a technology co-design approach drawing on the living lab/fab lab movement (cf. Dell'Era & Landoni, 2014).

For each context, the stakeholding societal groups need to 'own' and advise about specific problem solutions to carry them through accepted and supported implementation to successful outcome. This is enabled by a bundle of methodological resources that support joint problem definition and problem solution while respecting high degrees of differentiation for these participatory multi-stakeholder workshop formats. Among them are non-violent communication methods (cf. Rosenberg, 2015), Delphi methods (cf. Brady, 2015; Geist, 2010), citizen juries (cf. Dryzek et al., 2019; Drury et al., 2021), role-playing methods (cf. Meligrana & Andrew, 2003), force-field analyses (cf. Shrivastava et al., 2017), low-tech consultation methods (cf. Tanguy et al., 2023), Participatory Systems Mapping (see above),

scenario and forecast methods (cf. Kosow & Gassner, 2008), cooperation formats such as World Cafés, Fish Bowls, etc. (cf. Löhr et al., 2020; Doll et al., 2018), gamification and simulation (Ahrweiler et al., 2023), expressive group activities to befriend people and set the scene for sessions (cf. Vega-Ramírez et al., 2022), participatory modelling of desired futures and societal scenarios (cf. Voinov & Bousquet, 2010), hackathons, rapid prototyping, and agile programming (cf. Kiev Gama et al., 2023), etc.

Technology co-design is envisaged as a deliberative process in an ecological setting with competing interests constantly negotiating and coordinating the validity of their claims in networks of relationships. Using the results of these 'thick descriptions' (Geertz, 1987) as a blueprint, a comparative analysis across cases can show how and to what degree conventional social assessment processes in different international societies have been replaced or changed by AI.

A comparative analysis can also show where non-AI and AI processes differ, especially with regard to implemented social assessment values, and how societal stakeholders, policy, public discourse, and institutional infrastructures have responded or are still responding in terms of technology assessment. System mapping reveals gaps and barriers apparent from the perspective of stakeholders. It also points to their views about more desirable solutions in both the technical and the social realms. Here, participatory modelling and stakeholder-driven scenario simulations informed by a companion modelling framework (Barreteau et al., 2003; Etienne, 2013) will enable technology co-design towards 'Better AI'.[1]

Innovating AI for Social Service Provision with Vulnerable Groups

To develop ethically and societally responsible AI use for social assessment requires interaction between stakeholders from all kinds of backgrounds, who need 'voice' to communicate at 'eye-level' without being preconfigured, discriminated, and constrained by the environment where these encounters take place. Everybody should be perceived as expert for his/her area of lived experience and should be enabled to bring this to technology co-design. It is important to achieve multi-stakeholder engagement at case study level, bringing together representatives of national governments and administrations, NGOs, civil society organisations, industry, research, and the media to collect data on the values, perspectives, opinions, and attitudes of decision-makers and agenda-setters. Including vulnerable people to innovate in public administration has been acknowledged (e.g., Amann & Sleigh, 2021) as well as mentioned in the Sustainability Development Goals 2020 (Ali G20 2022 'leaving no one behind'), particularly Goal 9, building 'industry, infrastructure, and innovation' as well as Goal 16, building 'effective, accountable and inclusive institutions' (UN. Department of Economic and Social Affairs, n.d.).

[1] Results from the modelling work for 'Better AI' will be presented in the second edited volume on AI FORA to be published in 2025 in this series.

In each national welfare system, there are winners and losers, and sometimes the most vulnerable groups that should benefit most from state interventions fall through the social net (Reeves & Loopstra, 2017). 'Losers' are often marginalised and not sufficiently represented in democratic procedures or political participation activities. However, vulnerable groups, in particular those who have fallen through the social net or are not benefiting from it, can provide valuable information about the injustices, failures, and flaws of existing social assessment systems (DiSalvo et al., 2012; Zamenopoulos & Alexiou, 2018). To improve such systems, vulnerable groups need to be empowered to bring their experience to bear on the co-design of technology (Caforio et al., 2021). Eliminating injustice, bias, and discrimination in AI-enabled social service delivery requires the voices of non-recipients, mostly people without a voice, vulnerable populations and minorities, and critics such as NGO that back up for the failures of the public system, not just those of recipients, decision-makers, service providers, or technology producers.

Academia and practitioners justify the inclusion and participation of vulnerable groups in AI design and innovation ethically, politically, and epistemically. Each of these represents a different challenge, although there is a point at which they come into contact. This is expressed in the responsible innovation literature (cf. Ottinger, 2023; Steen, 2021; Valkenburg et al., 2020) as well as literature related to human rights (cf. Krupiy, 2021; Rodrigues, 2020). However, especially compared to AI innovation in the health sector, there is still a tendency to keep these concerns in general terms in the public (administration) sector, for example, by only writing about protection mechanisms for vulnerable groups (e.g., via legal regulations). While this is crucial, it is important to innovate with vulnerable groups (cf. Jarke, 2021) including them as part of co-design processes and not limiting their presence as users, beneficiaries, or customers (cf. Mulvale & Robert, 2021). It has been shown that there is a strong link between the involvement of affected groups and efficiency, which is also important in the context of the allocation of limited resources.

Transdisciplinary research, participatory action research, and innovation co-design already represent vibrant communities of practice in which socio-technological innovation processes include questioning segregation, injustice, and power asymmetries and challenging the status quo (Aldridge, 2015; Krupiy, 2021). Considering this and acknowledging the 'paradigm shift' in public administration increasingly valuing 'equity' (McDonald III et al., 2022), innovating with vulnerable groups in the context of using AI in the public sector is increasingly taking a proactive approach (McDonald III et al., 2022; Bustamante Duarte et al., 2018; Jarke, 2021).

Safe Spaces

Thus, the main risk of innovation in AI-based social assessment for public social service provision concerns a failure to successfully engage stakeholders, especially vulnerable groups, in technology co-design, and of appropriately organising

inter- and transdisciplinary interfaces. Although there are many process factors involved in ensuring eye-level participation of vulnerable groups, which can be methodologically supported by specifically-developed interactive and participative formats (see above). It is also important to provide 'Safe Spaces' to mitigate the risk of crowding out vulnerable groups.

Genealogy of the Concept

The notion of safe spaces has become a popular concept, adapted and applied to very different contexts, from mental health therapeutic practice, to classrooms, and sustainability issues (cf. Boostrom, 1998; Fitzpatrick et al., 2023; Pereira et al., 2015). According to the Oxford English Dictionary (2024), a safe space is 'a place or environment in which people, esp. those belonging to a marginalized group, can feel confident that they will not be exposed to discrimination, criticism, harassment, or any other emotional or physical harm'. The concept of safe space originated in the feminist and LGBT movements of the 1960s (cf. Kenney, 2001). The first safe spaces were locations where people, mainly women, could meet and isolate themselves from the violence of the outside world arising from their gender and race. These places created an atmosphere suitable for resistance, planning defence strategies, and challenging social injustices (cf. Kenney, 2001; The Roestone Collective, 2014). Safe spaces can emerge spontaneously, but they can also be created intentionally. Both forms have been studied and analysed. Studies of safe spaces as purpose-built spaces gather historical experiences not only to understand their evolution and properties, but also to identify best practices in their construction.

The practice of deliberately creating safe spaces has developed significantly in the field of education. These spaces are safe to the extent that they allow for non-violent coexistence in groups that are socially, racially, economically, and politically heterogeneous. They are spaces for students to speak and express themselves freely, protected from any form of violence, be it physical, psychological, or emotional. Care is taken in the construction of the spaces, but above all in the role that educators and teachers play in coordinating the social dynamics of the groups to ensure that safe conditions are in place (cf. Arao & Clemens, 2013; Barrett, 2010; Stengel & Weems, 2010). Experience in building safe spaces has led to their implementation in other domains, learning from the lessons of the past. An emerging area of safe spaces is research aimed at solving complex, uncertain, and socially impactful problems. Problem-based research and participatory action research considers the construction of sufficiently safe spaces relevant (cf. Charli-Joseph et al., 2018, 2023; Pereira et al., 2015, 2020).

Criticisms of the concept of safe spaces come from different fronts, some of them also applying to its expression as safe spaces for the co-creation of knowledge and innovation. The concept of safe spaces is associated with the idea of non-discrimination, protection from criticism, harassment or any other form of physical or emotional harm, yet none of these have been absent in the history of these spaces. Insofar as they are social spaces, the dynamics between their members can generate

violence and power asymmetries, as well as contradictions and paradoxes that are not easy to resolve. For example, in classrooms, students from groups that have historically been socially and economically privileged may feel aggrieved or develop a sense of guilt when issues of social injustice are exposed under a safe space scheme (cf. Arao & Clemens, 2013; Boostrom, 1998). In the case of safe spaces for co-production of knowledge and innovation, the location may inhibit and be openly rejected: the university may seem too strict and elitist and participants may feel better off in another space, for example, their village assembly hall or farming plot (cf. Charli-Joseph et al., 2018). Likewise, opening these spaces to all (stakeholders) can be counterproductive, for example, when members of one group perceive members of another group who have been invited to participate in the safe space as threatening, superior, or corrupt (cf. Charli-Joseph et al., 2018). It has been suggested that the term has begun to lose its meaning and drive, as it has been undertheorised (cf. Barrett, 2010; The Roestone Collective, 2014). However, a great deal of current work on the deliberate implementation of safe spaces and their theorisation focuses on resolving or navigating these internal tensions. Derivative concepts such as brave spaces (cf. Arao & Clemens, 2013; Barrett, 2010), safe enough spaces (Pereira et al., 2020), or paradoxical spaces (cf. Barrett, 2010; The Roestone Collective, 2014) have been proposed. Finally, these spaces have been generated from a Western perspective and conditions. A challenge regarding safe spaces for co-production and innovation is to construct them in a decolonial light, giving rise to a dialogue between diverse ways of knowing (cf. Fazey et al., 2020; Maclean et al., 2022; Mitchell & Matthew, 2018; Turnhout et al., 2020; Zgambo et al., 2018).

The risk is that individual stakeholder groups might not equally speak their minds and contribute their specific perspective and expertise, being cautioned or cowed by their surroundings. This risk is likely to appear when participative formats and venues are not neutral but are representing one involved actor's interest, thus failing to enable horizontal and integrative communication. For example, an academic university environment might be cowing to lower education status groups; and a business environment might be frustrating to people with low income and low socio-economic status. Public agencies would be the worst choice in this regard, because it is there that 'falling through the social net' takes place.

Safe Spaces in AI FORA

Safe Spaces need to work for vulnerable populations and minorities to address value propositions that are not yet given voice in technology development for detecting and avoiding bias and discrimination. The venues for this must be as 'remote from this world' as possible, but also fulfil a set of operational requirements to work across countries and cultural contexts.

A place open to all people, known and trusted by local communities

Safe Spaces need to be welcoming and open to all people in a community regardless of their background. People running the Safe Space need to be recognised as 'one

of them' embedded in the local culture and understanding it from the inside. The Safe Space needs to share the 'cultural footprint' being a part of the societal setting (to know what is going on in terms of problem space, to know the people involved, to speak the language, etc.). High visibility of the Safe Space as part of the local cultural heritage is an asset for fostering further trust and commitment. An institutional background famous for 'building culture' is helpful.

A place dedicated to discussing values and working on ethical questions (of AI)

Often, the situation of vulnerability demarcates the cultural value space based on religious beliefs, political interests, or other normative settings. Safe Spaces are places where people can freely contemplate and critically think about their values, belief sets, and cultural heritage, meeting others in open discourse. It is an additional asset if Safe Spaces have a genuine interest in discussing the ethical aspects of new technologies, especially questions concerning the 'Ethics of AI'.

A place with expertise and own activities in social service provision

Safe Spaces have a background in distribution practices in social service provision. To cater for the 'poor and needy' dedicated to individual wellbeing but also looking for the collective good is an asset of a Safe Space.

A place used to working with science and universities but not academic itself

Working with the project social researchers makes it a requirement that Safe Spaces have expertise in cooperating with academia and are familiar with scientific approaches in general, the universities in their region, and interdisciplinary thinking in research. However, it is important that the Safe Space is not an academic place itself. Safe Spaces need to appeal to people from different levels of education and experience (intellectual, psychological, emotional, social, etc.).

A place with expertise in reconciliation of conflicting value propositions

It is helpful if Safe Spaces have previously acted as 'trustworthy', neutral, and reliable mediators in social conflicts and have a certain tradition, reputation, credibility, and experience for moderating such processes among the local population. It is an additional advantage when Safe Spaces have dedicated expertise in specific participative methods (e.g. non-violent communication, etc.).

A place with appropriate infrastructure in venue, staff and networks

There are organisational and institutional characteristics that are supportive for Safe Spaces. The place must be able to host multi-stakeholder groups and workshops in terms of facilities and staff. Furthermore, it is an advantage if the Safe Space is connected to other similar places internationally, at best as part of a strong global network to enable synergies, cooperation, and comparability of approaches among case studies.

As local intermediaries, i.e. network organisations specialised in interreligious, intercultural, and inter-societal communication, AI FORA has worked with Benedictine monasteries as Safe Spaces. They fulfilled the criteria above and are

'remote out-of-the-world locations'. A first apprehension that monasteries would be perceived as representing 'the church' and equivalents or would be considered as 'too religious' for non-religious people proved to be unfounded and this was confirmed by a feedback questionnaire following inclusive Safe Spaces events to co-design AI technology. The venues worked for case studies as different as the Spanish, German, Chinese, Indian, and the Ukrainian ones—some of the activities will be reported in this volume. An evaluation of the concept will be provided in the conclusions of the book.

Summary and Outlook

Artificial intelligence in public social service provision has major ethical and societal implications for the future of our societies. Inclusive technology co-design uses participatory research methods to compare empirical cases and to create better, i.e. more responsible AI technology adapted to context-specific social value requirements, avoiding bias and discrimination. It relies on a complex networks approach to social innovation using new formats and methodologies aiming at intercultural, interreligious, and international understanding, communication, and reconciliation cooperating with local communities on global issues.

This chapter introduced the concepts that form the basis of the research presented in this volume. Literature on the concept space was reviewed and discussed for allowing an assessment of concepts—their application, differences in implementation and interpretation, and further development in the case studies following.

Concepts operationalise the overall hypothesis that the disconnection of cultural values from AI-based social assessment in national welfare systems increases unfairness such as existing bias and discrimination and at best reflects the fairness concepts of technologically-dominant developers ('colonial AI'). De-contextualised static AI that is unresponsive to value discourse and policy interventions can be a serious impediment to social change. This can lead to the stagnation of social development, to more pressure on vulnerable groups, social conflict and mass migration, increased poverty, nutritional deficiencies, and health problems.

To connect technology to society, i.e. to bring cultural values and dynamic value discourse to AI-based social assessment in national welfare systems, we have used case studies that employ a participatory approach. Inclusive technology co-design engages stakeholders at all levels, among them decision-makers from policy and technology production, public authorities that develop technology policies and employ AI for social assessment in their social service providing agencies, academia and research organisations that invent and research AI technologies, businesses that develop, produce, and sell them, and also communities and citizens including vulnerable groups that are exposed to them.

Actors engage in multi-stakeholder innovation networks bringing their expertise, perspectives, views, and concerns to policy and technology production. Participatory tools are targeted to support and realise participative innovation efforts. Stakeholders

need to be empowered to obtain relevant knowledge as well as to evaluate the decisions made in technological development in terms of societal needs and moral values and—in particular—also the potential alternatives in these developmental processes. For this purpose, they need to be involved at an early stage of technology development. Likewise, stakeholders need to be an integral part of the governance of technology development, being empowered to assess questions of if, when, and how to regulate it.

Stakeholding actors for each case study context have been mapped and structurally investigated by social network analysis. Their different backgrounds, value concepts, economic and social interests, power and agency, cultural views and visions have been the subject of qualitative social research in case studies; their exchanges, interactions, knowledge transfers, value negotiations, and conflicts have been analysed and moderated by participatory action research and low-barrier inclusion methods in sheltered Safe Spaces environments. This approach has also allowed for the interaction and empowerment of vulnerable groups in technological innovation. Data collected and analysed in case study research following this conceptual framework have identified gaps and barriers in existing and requirements for desired systems.

Not all case studies, however, have implemented and used all concepts. And of course, if used, the interpretation of concepts is again context-bound, case-study-specific, and culture-dependent. However, the concept space, its mode and degree of implementation, allows for a common terminology and a measure for case study comparisons between Spain, USA, India, Ukraine, Iran, Nigeria, Germany, Estonia, and China.

By participatory modelling and scenario simulation (to be reported in the next volume of this series), integrated options for 'better AI' are developed as the product of stakeholder interaction for inclusive technology co-design. Concrete changes in policy and technology are expected as the outcome to avoid any type of 'participation washing'. This is not only ensured by the inclusion of public authorities and technology producers such as computer science institutes and software companies in the co-designing multi-stakeholder settings of the case studies; dedicated policy workshops in capitols of case study countries will connect the stakeholder-driven results to the high level of national and international AI governance frameworks. This is to ensure the transfer and application of stakeholder knowledge.

Last but not least, inclusive technology co-design requires inclusive science communication. This volume presenting academic research in case study countries was preceded by 'Angels and other Cows. A Celestial Adventure into AI Worlds, the Social Good, and Unknown Connections', an open-access novel in literary fiction blending genres such as sci-fi, romance, adventure, mystery, and comedy. The task of the novel in this series is inclusive science communication making available research topics, results, and consequences of AI use in the public sector to a broad readership and attracting non-scientists to academic research. This approach of using literary fiction, complemented by a graphic novel, will also be chosen to conclude and evaluate research after the second academic volume on modelling and simulating 'better AI'. Further measures of inclusive science communication

include an open, accessible, and interactive web presence and social media in the digital space.

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz in Germany, are gratefully acknowledged.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahrweiler, P. (2010). Innovation in complex social systems - An introduction. In P. Ahrweiler (Ed.), *Innovation in complex social systems* (pp. 1–25). Routledge. https://doi.org/10.4324/9780203855324

Ahrweiler, P., Gilbert, N., Bicket, M., Sabater Coll, A., Luque Capellas, B., Wurster, D., Siqueiros, J., & Späth, E. (2023). Gamification and simulation for innovation. In C. Elsenbroich (Ed.), *Advances in social simulation*. Springer proceedings in complexity. Springer . https://doi.org/10.1007/978-3-031-34920-1 (in press).

Aldridge, J. (2015). Participatory research: Working with vulnerable groups in research and practice. *Journal of Social Policy, 45*(3), 565–566. https://doi.org/10.1017/S0047279416000106

AlgorithmWatch. (2020). *AI ethics guidelines global inventory*. Retrieved December 19, 2023, from https://inventory.algorithmwatch.org/

Amann, J., & Sleigh, J. (2021). Too vulnerable to involve? Challenges of engaging vulnerable groups in the co-production of public services through research. *International Journal of Public Administration, 44*(9), 715–727. https://doi.org/10.1080/01900692.2021.1912089

Anagnostou, M., Karvounidou, O., Katritzidaki, C., Kechagia, C., & Melidou, K. (2022). Characteristics and challenges in the industries towards responsible AI: A systematic literature review. *Ethics and Information Technology, 24*, 37. https://doi.org/10.1007/s10676-022-09634-1

Arao, B., & Clemens, K. (2013). From safe spaces to brave spaces: A new way to frame dialogue around diversity and social justice. In L. Landreman (Ed.), *The art of effective facilitation: Reflections from social justice educators* (pp. 135–150). Stylus Publishing. https://doi.org/10.4324/9781003447580

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barbrook-Johnson, P., & Penn, A. S. (2022). *Systems mapping: How to build and use causal models of systems*. Springer International Publishing. https://doi.org/10.1007/978-3-031-01919-7

Barreteau, O., et al. (2003). Our companion modelling approach. *Journal of Artificial Societies and Social Simulation, 6*(2) https://jasss.soc.surrey.ac.uk/6/2/1.html

Barrett, B. J. (2010). Is "Safety" dangerous? A critical examination of the classroom as safe space. *The Canadian Journal for the Scholarship of Teaching and Learning, 1*(1), 9. https://doi.org/10.5206/cjsotl-rcacea.2010.1.9

Benjamin, J. J., Kinkeldey, C., & Müller-Birn, C. (2020). Participatory design of a machine learning driven visualization system for non-technical stakeholders. *Lecture notes in informatics (LNI), Proceedings - Series of the Gesellschaft Fur Informatik (GI)*. https://doi.org/10.18420/MUC2020-WS109-287.

Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M.C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM conference on equity and access in algorithms, mechanisms, and optimization (EAAMO '22)* (pp. 1–8). https://doi.org/10.1145/3551624.3555290.

Bischof, A., & Jarke, J. (2021). Configuring the older adult. In A. Peine, B. Marshall, W. Martin, & L. Neven (Eds.), *Socio-gerontechnology: Interdisciplinary critical studies of ageing and technology* (1st ed., pp. 197–212). Routledge. https://doi.org/10.4324/9780429278266-18

Bleher, H., & Braun, M. (2023). Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics, 29*(21). https://doi.org/10.1007/s11948-023-00443-3

Boostrom, R. (1998). 'Safe spaces': Reflections on an educational metaphor. *Journal of Curriculum Studies, 30*, 397–408. https://doi.org/10.1080/002202798183549

Brady, S. R. (2015). Utilizing and adapting the Delphi method for use in qualitative research. *International Journal of Qualitative Methods, 14*(5). https://doi.org/10.1177/1609406915621381

Brown, K. (2019). Vulnerability and child sexual exploitation: Towards an approach grounded in life experiences. *Critical Social Policy, 39*(4), 622–642. https://doi.org/10.1177/0261018318824480

Bustamante Duarte, A. M., Degbelo, A., & Kray, C. (2018). Exploring forced migrants (Re)settlement & the role of digital services. In *Proceedings of 16th European conference on computer-supported cooperative work - exploratory papers.* European Society for Socially Embedded Technologies (EUSSET). https://doi.org/10.18420/ecscw2018_7.

Cacari-Stone, L., Wallerstein, N., Garcia, A. P., & Minkler, M. (2014). The promise of community-based participatory research for health equity: A conceptual model for bridging evidence with policy. *American Journal of Public Health, 104*(9), 1615–1623. https://doi.org/10.2105/AJPH.2014.301961

Caforio, A., Pollini, A., Filograna, A. S., & Passani, A. (2021). Design issues in human-centered AI for marginalized people. *ITAIS 2021 Proceedings,* 5. https://aisel.aisnet.org/itais2021/5

Charli-Joseph, L., Siqueiros-Garcia, J. M., Eakin, H., Manuel-Navarrete, D., & Shelton, R. (2018). Promoting agency for social-ecological transformation: A transformation-lab in the Xochimilco social-ecological system. *Ecology and Society, 23*(2) https://www.jstor.org/stable/26799122

Charli-Joseph, L., Siqueiros-García, J. M., Eakin, H., Manuel-Navarrete, D., Mazari-Hiriart, M., Shelton, R., Pérez-Belmont, P., & Ruizpalacios, B. (2023). Enabling collective agency for sustainability transformations through reframing in the Xochimilco social–ecological system. *Sustainability Science, 18*, 1215–1233. https://doi.org/10.1007/s11625-022-01224-w

Cornish, F., Breton, N., Moreno-Tabarez, U., Delgado, J., Rua, M., de-Graft Aikins, A., & Hodgetts, D. (2023). Participatory action research. *Nature Review Methods Primers, 3*(34). https://doi.org/10.1038/s43586-023-00214-1

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly, 39*(2). https://doi.org/10.1016/j.giq.2021.101666

Dell'Era, C., & Landoni, P. (2014). Living lab: A methodology between user-centred design and participatory design. *Creativity and Innovation Management, 23*, 137–154. https://doi.org/10.1111/caim.12061

Deutscher Ethikrat. (2023). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Stellungnahme*. Retrieved December 19, 2023, from https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion, 99*. https://doi.org/10.1016/j.inffus.2023.101896

Dignum, V. (2021). The role and challenges of education for responsible AI. *London Review of Education, 19*(1). https://doi.org/10.14324/LRE.19.1.01

DiSalvo, C., Clement, A., & Pipek, V. (2012). Communities. Participatory design for, with and by communities. In J. Simonsen & T. Robertson (Eds.), *Routledge international handbook of participatory design*. Routledge. https://doi.org/10.4324/9780203108543

Doll, J. E., Eschbach, C. L., & DeDecker, J. (2018). Using dialogue to engage agricultural audiences in cooperative learning about climate change: A strategy with broad implications. *The Journal of Extension, 56*(2), Article 25. https://doi.org/10.34068/joe.56.02.25

Donia, J., & Shaw, J. A. (2021). Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. *Big Data and Society, 8*(2). https://doi.org/10.1177/20539517211065248

Drury, S., Elstub, S., Escobar, O., & Roberts, J. (2021). Deliberative quality and expertise: Uses of evidence in citizens' juries on wind farms. *Journal of Deliberative Democracy, 17*(2), 10.16997/jdd.986.

Dryzek, J. S., Bowman, Q., Kuyper, J., Pickering, J., Sass, J., & Stevenson, H. (2019). *Deliberative global governance*. Cambridge University Press. https://doi.org/10.1017/9781108762922

Durán, J. M., & Pirtle, Z. (2020). Epistemic standards for participatory technology assessment: Suggestions based upon well-ordered science. *Science and Engineering Ethics, 26*(3), 1709–1741. https://doi.org/10.1007/s11948-020-00211-7

Etienne, M. (2013). *Companion modelling. A participatory approach to support sustainable development*. Springer. https://doi.org/10.1007/978-94-017-8557-0

Etzkowitz, H., & Leydesdorff, L. (1998). A triple helix of university—industry—government relations: Introduction. *Industry and Higher Education, 12*(4), 197–201. https://doi.org/10.1177/095042229801200402

European Commission. (2017). *Horizon 2020*. Retrieved June 22, 2017, from. https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en

European Commission. (2019, April 8). *Ethics guidelines for trustworthy AI*. European Commission. Retrieved December 19, 2023, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission. (2021, April 21). *Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. European Comission. Retrieved December 19, 2023, from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

European Commission. Directorate-General for Research and Innovation. (2007, January). *Taking European Knowledge Society Seriously*. European Commission. Retrieved December 19, 2023, from https://op.europa.eu/en/publication-detail/-/publication/5d0e77c7-2948-4ef5-aec7-bd18efe3c442/language-en

European Council. (2024). *Timeline - Artificial intelligence*. Retrieved June 17, 2024 from https://www.consilium.europa.eu/en/policies/artificial-intelligence/timeline-artificial-intelligence/

Fagerberg, J., Mowery, D., & Nelson, R. (2006). *The Oxford handbook of innovation*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286805.001.0001

Fals-Borda, O. (1987). The application of participatory action-research in Latin America. *International Sociology, 2*(4), 329–347. https://doi.org/10.1177/026858098700200401

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass, 16*(8), e12760. https://doi.org/10.1111/phc3.12760

Fazey, I., et al. (2020). Transforming knowledge systems for life on Earth: Visions of future systems and how to get there. *Energy Research and Social Science, 70*, 101724. https://doi.org/10.1016/j.erss.2020.101724

Fitzpatrick, S. J., Lamb, H., Stewart, E., Gulliver, A., Morse, A. R., Giugni, M., & Banfield, M. (2023). Co-ideation and co-design in co-creation research: Reflections from the 'Co-Creating Safe Spaces' project. *Health Expectations: An International Journal of Public Participation in Health Care and Health Policy, 26*(4), 1738–1745. https://doi.org/10.1111/hex.13785

Funtowicz, S., & Ravetz, J. (1993). Science for the post-normal age. *Futures, 25*(7), 739–755. https://doi.org/10.1016/0016-3287(93)90022-L

Gama, K., Valença, G., Laurendon, C. E. M., Marques, Á. N., Ramos, L. E., Amaral, R., ... & Xavier, G., (2023): Hackathons as Inclusive Spaces for Prototyping Software in Open Social Innovation with NGOs. In 2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS): 58–70.

Geertz, C. (1987). Dichte Beschreibung: Bemerkungen zu einer deutenden Theorie von Kultur. In C. Geertz (Ed.), *Dichte Beschreibung: Beiträge zum Verstehen kultureller Systeme* (pp. 7–43). Suhrkamp.

Geist, M. R. (2010). Using the Delphi method to engage stakeholders: A comparison of two studies. *Evaluation and Program Planning, 33*(2), 147–154. https://doi.org/10.1016/j.evalprogplan.2009.06.006

Gianni, R., & Goujon, P. (2014). *GREAT_Del2.3_final*. Retrieved December 19, 2023, from http://www.great-project.eu/deliverables_files/deliverables02

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (2010). *The new production of knowledge: The dynamics of science and research in contemporary societies*. Sage. https://doi.org/10.4135/9781446221853.

Greenhalgh, T., Jackson, C., Shaw, S., & Janamian, T. (2016). Achieving research impact through co-creation in community- based health services: Literature review and case study. *Milbank Quarterly, 94*(2), 392–429. https://doi.org/10.1111/1468-0009.12197

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37). https://doi.org/10.1126/scirobotics.aay7120

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120. https://doi-org.uchile.idm.oclc.org/10.1007/s11023-020-09517-8

Halskov, K., & Hansen, N. B. (2015). The diversity of participatory design research practice at PDC 2002-2012. *International Journal of Human-Computer Studies, 74*, 81–92. https://doi.org/10.1016/j.ijhcs.2014.09.003

Hansen, H. K., Christensen, L. T., & Flyverbom, M. (2015). Introduction: Logics of transparency in late modernity: Paradoxes, mediation and governance. *European Journal of Social Theory, 18*(2), 117–131. https://doi.org/10.1177/1368431014555254

Hirom, U., Shyama, V. S., Doke, P., Lobo, S., Devkar, S., & Pandey, N. (2017). A critique on participatory design in developmental context: A case study. In P. L. Rau (Ed.), *Cross-cultural design* (CCD 2017. Lecture Notes in Computer Science) (Vol. 10281). Springer. https://doi.org/10.1007/978-3-319-57931-3_52

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence, 4*(4), 799–819. https://doi.org/10.1109/TAI.2022.3194503

Iason, G. (2022): Toward a Theory of Justice for Artificial Intelligence. Daedalus 151: 218–231

Jarke, J. (2021). *Co-creating digital public services for an ageing society: Evidence for user-centric design*. Springer. https://doi.org/10.1007/978-3-030-52873-7

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kenney, M. (2001). *Mapping gay LA: The intersection of place and politics*. Temple University Press.

Kosow, H., & Gassner, R. J. (2008). *Methods of future and scenario analysis: Overview, assessment, and selection criteria*. Deutsches Institut für Entwicklungspolitik gGmbH. Retrieved June 18, 2024, from https://www.idos-research.de/uploads/media/Studies_39.2008.pdf

Krupiy, T. (2021). Understanding digital discrimination: Analysing marshall mcluhan's work through a human rights lens. *New Explorations: Studies in Culture and Communication, 2*(1), 1–22. https://ssrn.com/abstract=4311885

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). University of Chicago Press. https://doi.org/10.7208/chicago/9780226458106.001.0001

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review, 9*(2). https://doi.org/10.14763/2020.2.1469

Lee, M. K., Kahng, A., Kim, J., Yuan, X., Chan, A., Lee, S., Procaccia, A., Kusbit, D., See, D., Nooth-Igattu, R., & Psomas, A. (2019). WeBuildAI: Participatory framework for algorithmic

governance. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–35. https://doi.org/10.1145/3359283

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*(4), 34–46. https://doi.org/10.1111/j.1540-4560.1946.tb02295.x

Lewis, V. A., Larson, B. K., McClurg, A. B., Boswell, R. G., & Fisher, E. S. (2012). The promise and peril of accountable care for vulnerable populations: A framework for overcoming obstacles. *Health affairs (Project Hope), 31*(8), 1777–1785. https://doi.org/10.1377/hlthaff.2012.0490

Liao, Q. V., & Muller, M. (2019). Enabling value sensitive AI systems through participatory design fictions. *ArXiv*. https://doi.org/10.48550/arXiv.1912.07381.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

Löhr, K., Weinhardt, M., & Sieber, S. (2020). The "World Café" as a participatory method for collecting qualitative data. *International Journal of Qualitative Methods, 19*. https://doi.org/10.1177/1609406920916976

Maclean, K., Greenaway, A., & Grünbühel, C. (2022). Developing methods of knowledge co-production across varying contexts to shape sustainability science theory and practice. *Sustainability Science, 17*, 325–332. https://doi.org/10.1007/s11625-022-01103-4

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Putting AI ethics into practice: The hourglass model of organizational AI governance. *ArXiv*. https://doi.org/10.48550/arXiv.2206.00335

McAdam, M., & Debackere, K. (2018). Beyond 'triple helix' toward 'quadruple helix' models in regional innovation systems: Implications for theory and practice. *R&D Management, 48*, 3–6. https://doi.org/10.1111/radm.12309

McDonald, B. D., III, Hall, J. L., O'Flynn, J., & van Thiel, S. (2022). The future of public administration research: An editor's perspective. *Public Administration, 100*(1). https://doi.org/10.1111/padm.12829

Meligrana, J. F., & Andrew, J. S. (2003): Role-playing simulations in urban planning education: A survey of student learning expectations and outcomes. *Planning practice & research, 18*(1): 95–107.

Mikalef, P., Conboy, K., Eriksson Lundström, J., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems, 31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Mitchell, K., & Matthew, S. (2018). Hotspot geopolitics versus geosocial solidarity: Contending constructions of safe space for migrants in Europe. *Environment and Planning D. Society and Space, 38*(6), 1046–1066. https://doi.org/10.1177/0263775818793647

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy and Technology, 33*, 659–684. https://doi.org/10.1007/s13347-020-00405-8

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Mulvale, G., & Robert, G. (2021). Special issue-engaging vulnerable populations in the co-production of public services. *International Journal of Public Administration, 44*, 711–714. https://doi.org/10.1080/01900692.2021.1921941

National Institute of Standards and Technology (NIST). (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. NIST. Retrieved December 19, 2023, from https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Osborne, S., Powell, M., Cui, T., & Strokosch, K. (2022). Value creation in the public service ecosystem: An integrative framework. *Public Administration Review, 82*, 634–645. https://doi.org/10.1111/puar.13474

Ottinger, G. (2023). Responsible epistemic innovation: How combatting epistemic injustice advances responsible innovation (and vice versa). *Journal of Responsible Innovation, 10*(1), 2054306. https://doi.org/10.1080/23299460.2022.2054306

Oxford English Dictionary. (2024). *safe space*. Retrieved June 17, 2024, from https://www.oed.com/dictionary/safe-space_n?tab=meaning_and_use

Pereira, L., Karpouzoglou, T., Doshi, S., & Frantzeskaki, N. (2015). Organising a safe space for navigating social-ecological transformations to sustainability. *International Journal of Environmental Research and Public Health, 12*, 6027–6044. https://doi.org/10.3390/ijerph120606027

Pereira, L., Frantzeskaki, N., Hebinck, A., Charli-Joseph, L., Drimie, S., Dyer, M., Eakin, H., Galafassi, D., Karpouzoglou, T., Marshall, F., Moore, M. L., Olsson, P., Siqueiros-García, J. M., van Zwanenberg, P., & Vervoort, J. M. (2020). Transformative spaces in the making: Key lessons from nine cases in the Global South. *Sustainability Science, 15*, 161–178. https://doi.org/10.1007/s11625-019-00749-x

Pérez-Escolar, M., & Canet, F. (2022). Research on vulnerable people and digital inclusion: Toward a consolidated taxonomical framework. *Universal Access in the Information Society, 22*, 1059–1072. https://doi.org/10.1007/s10209-022-00867-x

Pollini, A., & Caforio, A. (2021). Participation and iterative experiments: Designing alternative futures with migrants and service providers. *Social Sciences, 10*(10), 363. https://doi.org/10.3390/socsci10100363

Ramaswamy, V., & Ozcan, K. (2014). The co-creation paradigm. *Stanford University Press.* https://doi.org/10.1515/9780804790758

Reeves, A., & Loopstra, R. (2017). 'Set up to fail'? How welfare conditionality undermines citizenship for vulnerable groups. *Social Policy and Society, 16*(2), 327–338. https://doi.org/10.1017/S1474746416000646

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology, 4*. https://doi.org/10.1016/j.jrt.2020.100005

Rosenberg, M. (2015). *Nonviolent communication: A language of life* (3rd ed.). PuddleDancer Press.

Schmid, U., & Wrede, B. (2022). What is missing in XAI so far? *KI - Künstliche Intelligenz, 36*, 303–315. https://doi.org/10.1007/s13218-022-00786-2

Schumpeter, J. A. (1912). *The theory of economic development*. Harvard University Press.

Shrivastava, S. R., Shrivastava, P. S. & Ramasamy J., (2017): Force field analysis: An effective tool in qualitative research. J Curr Res SciMed 3: 139–40.

Steen, M. (2013). Virtues in participatory design: Cooperation, curiosity, creativity, empowerment and reflexivity. *Science and Engineering Ethics, 19*, 945–962. https://doi.org/10.1007/s11948-012-9380-9

Steen, M. (2021). Slow Innovation: The need for reflexivity in Responsible Innovation (RI). *Journal of Responsible Innovation, 8*(2), 254–260. https://doi.org/10.1080/23299460.2021.1904346

Stengel, B. S., & Weems, L. (2010). Questioning safe space: An introduction. *Studies in Philosophy and Education, 29*, 505–507. https://doi.org/10.1007/s11217-010-9205-8

Tanguy, A., Carrière, L., & Laforest, V. (2023). Low-tech approaches for sustainability: Key principles from the literature and practice. *Sustainability: Science, Practice and Policy, 19*(1). https://doi.org/10.1080/15487733.2023.2170143

The Roestone Collective. (2014). Safe space: Towards a reconceptualization. *Antipode, 46*, 1346–1365. https://doi.org/10.1111/anti.12089

Turnhout, E., Metze, T., Wyborn, C., Klenk, N., & Louder, E. (2020). The politics of co-production: Participation, power, and transformation. *Current Opinion in Environmental Sustainability, 42*, 15–21. https://doi.org/10.1016/j.cosust.2019.11.009

UNESCO. (2022). *Recommendation on the ethics of artificial intelligence*. UNESCO. Retrieved December 19, 2023, from https://unesdoc.unesco.org/ark:/48223/pf0000381137

UN. Departement of Economic and Social Affairs. (n.d.): Module 7 "Innovating Public Service for Vulnerable Groups": 1–39.

University of Montreal. (2018). *Montreal declaration for a responsible development of AI*. Retrieved December 19, 2023, from https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf

Valkenburg, G., Mamidipudi, A., Pandey, P., & Bijker, W. E. (2020). Responsible innovation as empowering ways of knowing. *Journal of Responsible Innovation, 7*(1), 6–25. https://doi.org/10.1080/23299460.2019.1647087

Van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., ... & Peek, N. (2021): Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. Journal of the American Medical Informatics Association, 28(10): 2128–2138.

Vega-Ramírez, L., Vidaci, A., & Hederich-Martínez, C. (2022). The effect of group work on expressive-artistic activities for the emotional regulation of university students. *Education Sciences, 12*(11), 777. https://doi.org/10.3390/educsci12110777

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications, 11*, 233. https://doi.org/10.1038/s41467-019-14108-y

Voinov, A., & Bousquet, F. (2010). Modelling with stakeholders. *Environmental Modelling and Software, 25*(11), 1268–1281. https://doi.org/10.1016/j.envsoft.2010.03.007

Zamenopoulos, T., & Alexiou, K. (2018, September). *Co-design as collaborative research*. Bristol University/AHRC Connected Communitites Programme. Retrieved June 18, 2024, from https://oro.open.ac.uk/58301/1/Co-Design_CCFoundationSeries_PUBLISHED.pdf

Zgambo, O., Pereira, L., Boatemaa, S., & Drimie, S. (2018). *T-lab for alternative food systems in the western cape*. Centre for Complex Systems in Transition. Retrieved December 19, 2023, from http://www.southernafricafoodlab.org/wp-content/uploads/2017/02/T-Lab-Report-.pdf

Chapter 3
Ethical Aspects of Research on AI-Based Social Assessment

Gerhard Kruip, Elisabeth Späth, and Albert Sabater

Abstract This chapter provides a meta-level *self-reflection* of research on AI-based social assessment with regard to its handling of normativity. It is motivated by the idea that, on the one hand, there cannot be a one-fits-all solution to ethical issues due to the high diversity of cultures and contexts. On the other hand, it is necessary to discuss the possibility of a common position based on shared fundamental norms such as non-discrimination and fairness. These efforts, reflected by the idea of establishing an *Ethical Observatory*, should be relevant to any research on AI-based social assessment. This chapter, therefore, explores various ethical issues related to AI in the context of social service provision, linking them to a reflective approach to social justice and how it could or should be improved by better AI. As different cultural values claim moral validity, it is necessary to scrutinise the relationship between their contextuality and universality. On this basis, discourse ethics is proposed as a model for dealing with cultural differences without abandoning universally valid standards, linking it with the role of safe spaces and stakeholder dialogues. Furthermore, it shows how the justification of moral norms can be reasonably argued. Based on this, important research ethical aspects will be considered.

G. Kruip (✉)
Department of Social Ethics, Faculty of Catholic Theology, Johannes Gutenberg-University Mainz, Mainz, Germany
e-mail: kruip@uni-mainz.de

E. Späth
TISSS Lab, Institute of Sociology, Johannes Gutenberg University Mainz, Mainz, Germany
e-mail: espaeth@uni-mainz.de

A. Sabater
Facultat de Ciencies Economiques i Empresarials, Departament d'Empresa, Edifici Econòmiques, C/ de la Universitat de Girona, Girona, Spain
e-mail: albert.sabater@udg.edu

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_3

Introduction

The aim of AI FORA is to search for *better* AI. But what exactly does *better* mean? From the outset, it was clear that the AI FORA research project would not only focus on criteria such as efficiency, safety, or functionality, but rather on ethical criteria. AI FORA tries to improve AI use in ethically sensitive and highly normative societal domains such as social assessment for public welfare provision. Therefore, already in Chaps. 1 and 2, a lot of ethical aspects are mentioned. This Chap. 3 aims to provide a meta-level *self-reflection* of research on AI-based social assessment with regard to its normative settings and its handling of normativity. Whereas in some areas social sciences legitimately restrict themselves to *descriptive* approaches, analysing social phenomena as they are and as they develop over time, in other projects it cannot be avoided, or it is even the aim of the project to include *normative* dimensions. This does not only require questioning what *is* and *how it is*, but also how the behaviour of people and social institutions or structures *ought to be*. These efforts represent the core idea of an *Ethics Observatory* in this AI FORA research project and should be relevant also for other research on AI-based social assessment. We will not be able to provide a one-fits-all solution to ethical issues, but at least want to reflect on them, show that we are aware of ethical problems related to AI, and discuss the possibility of a commonly shared position. Indeed, all participants in the project share a common vision: they start from fundamental norms such as non-discrimination and fairness.

This chapter begins with a reflection on the goal of *social justice* that should be improved by better AI. We also have to mention some specific *ethical problems of AI* that need to be considered. In both cases, it becomes obvious that there are *cultural differences* between different conceptions of social justice and ethically acceptable AI. Describing these differences is the task of an *empirical* approach to ethics. But all different cultural values claim moral validity and the question arises whether they contradict each other in a conflicting way or whether there can be a common understanding of these values. Therefore, the relationship between *contextuality and universality* of moral norms or values needs to be discussed. *Discourse ethics* offer a model for dealing with cultural differences without a priori abandoning universally valid standards. It also shows how the justification of moral norms can be reasonably argued. Based on this we can also reflect on some aspects of *research ethics* in the context of AI-based social assessment.

Social Justice

The title of this book makes it a little bit more concrete, what is meant by “better AI”. Hereby, *social justice* is a crucial concept. By *social justice* we refer to an ethically sensitive and highly normative societal domain that is often addressed in public and political discourse without a precise definition, while at the same time being

involved in often highly emotional controversies. To approach the concept of social justice, it is useful to start by thinking about justice in general. People often react indignantly to injustice, because it seems to be easier to identify injustices than to explain what justice is (Shklar, 1990). Obviously, everyone has a sense of injustice (Kaplow & Lienkamp, 2005). However, this does not mean that it is always clear what they understand by justice, especially as some things are perceived as unjust by some and as just by others. Empirical research into justice (Liebig & Lengfeld, 2002) has found wide variations across time, context, and culture. Nevertheless, there is a basic rule that is common to all conceptions of justice, namely that partiality must be avoided, i.e. fairness must be achieved (Höffe, 2001). It is therefore not surprising that reference is repeatedly made to fairness in Chaps. 1 and 2. Examples from soccer can be used to illustrate what is meant by this. It would be unfair if one team's goal were bigger than the other's or if the referee penalised one team for a foul but overlooked the other. That is why he is also called *the impartial* referee. An ancient motto for fair trials, which is apparently present in all cultures, is "audiatur et altera pars"—the other party must also be heard. In front of courthouses, justice is often depicted as a blindfolded woman, because the court is concerned with passing judgements without regard to the person. Finally, in regard to impartiality, one can identify the fundamental idea of the equal human dignity of all human beings with its consequence of equal rights for all.

Accordingly, the American philosopher John Rawls (2005) framed his famous theory of justice as a theory of fairness. To find out what justice is, he invites us to try a thought experiment: imagine that the members of a society meet in a primordial state to determine the rules for their future coexistence with no one knowing who they will be in that future society. This forces everyone to put themselves in all possible future positions and so they will therefore not agree to any rules that discriminate against or privilege individuals or groups. They arrive at impartial, fair rules.

Questions of justice arise for very different groups in very different contexts: in families, in schools, in companies, in associations, in political parties, etc. Questions of *social justice* have been particularly important since the emergence of nation states. The following aspects are usually discussed when talking about social justice: How should contributions and income be distributed among the members of a society? How should public goods be financed? How should common risks in life be mitigated by social insurance systems? What are the needs of people in difficult economic situations? Which needs should be met, to what extent and by whom? Egalitarians think that equality is the most important criterion of justice. However, they admit that total equality would not make sense in modern societies using market mechanisms and economic incentives. Nevertheless, John Rawls (2005) argued in favour of the famous *difference principle*, allowing inequalities that are to the greatest benefit of the least advantaged and combined with equality of opportunities. For non-egalitarians (e.g. Nozick, 1974), equality is not important for justice. They insist on the right to liberty, the right to private property and try to establish meritocracy, but also concede that this only can be legitimate when combined with the equality of opportunities. They therefore reduce the level of satisfaction of social

needs to a certain minimum subsistence level, which must nevertheless be compatible to the idea of human dignity (Kruip, 2004; White, 2010). Thanks to the inspiring works of Martha Nussbaum and Amartya Sen, many academics of today insist on the need to enable people and help them to develop capabilities since having capabilities is necessary to use opportunities and realise at least some fairness of opportunities (cf. Sen, 2000). Therefore, equality of opportunities in the education system, for example, has high importance for social justice. However, this must not replace distributive justice. Social services and financial support are very important for equal opportunities, too.

Human Autonomy and AI

As noted in Chaps. 1 and 2, in addition to these general issues related to the understanding of social justice, there are specific ethical questions concerning algorithmic decision-making, which involves delegating decisions based on prevailing judgements to machines, particularly in the context of the distribution of social services. Such use of AI has the potential to either support what people understand by social justice in different contexts—or to hinder social justice. Therefore, general ethical problems of AI have been discussed for some years, and many lists of ethical criteria have been published (cf. Hagendorff, 2020; Grimm et al., 2023). In March 2018, the European Group on Ethics in Science and New Technologies identified human dignity, autonomy, responsibility, justice, equity and solidarity, democracy, rule of law and accountability, security, safety, physical and mental integrity, data protection, privacy, and sustainability as the most important principles (European Group on Ethics in Science and New Technologies, 2018). A year later, the High-Level Expert Group on Artificial Intelligence added the principles of accountability and transparency to its foundations for trustworthy AI, while also insisting on respect for human autonomy, prevention of harm and fairness (Independent High-Level Expert Group on Artificial Intelligence, 2019). In 2021, UNESCO published similar recommendations on the ethics of artificial intelligence (UNESCO, 2021). In the European Union, discussions between member states and the European Parliament lead to a legal regulation on AI, which is the first AI-related law in the world (European Parliament News, 2023). The main idea is that AI systems should be safe, transparent, traceable, non-discriminatory, and environmentally friendly. AI systems should be supervised by humans, not automation, to prevent harmful outcomes. The Catalan Observatory for Ethics in Artificial Intelligence (OEIAC in Catalan) was also developed as part of the Catalonia AI Strategy of the Generalitat de Catalunya, with the aim "to promote the development of an AI ethics which respects current legality, is compatible with our social and cultural norms and human-centred" (Catalan Observatory, 2023). The results of the stakeholder dialogues that took place during this process show, interestingly, that ethical positions range from being a main goal and/or a configurative part of AI. However, there are

also more neutral and partly expectant positions, willing to let the development of AI set its own limits without ethical considerations (ibid.).

More recently, in a similar way, the German Ethics Council published a document in which it insists that the key question for an ethical assessment of AI is whether human authorship and the conditions for responsible action are extended using AI or restricted (Deutscher Ethikrat, 2023). To answer this question, it is necessary to consider in detail the implications associated with the application of AI in different contexts: problems of governance, different economic interests and existing power imbalances in business and politics, the impact on the democratic public sphere, and the urgent need to educate and empower users. AI must not replace human contact and human support, preferably by professional social workers, especially in the context of social assessment for vulnerable groups who need not only financial support but even more social services. Another important aspect is that behind AI (development) there are privileged social groups, such as political and economic elites, who use AI to support and legitimise their own social position. As Rafanelli (2022) puts it, AI is a tool with which "humans exercise power, not a replacement for human power".

As can be seen in the various catalogues of ethical criteria, human dignity is the underlying idea. As Immanuel Kant put it in his famous reflections on autonomy and dignity, human beings, who must always be treated as ends and never merely as means, can only be obliged to obey moral laws given by themselves, but given according to the principle of universalisability. Corresponding to his third version of the categorical imperative, "any maxim is rejected if it is not consistent with the will's role as a giver of universal law. Hence the will is not merely subject to the law, but subject to it in such a way that it must be viewed as prescribing the law to itself, and for just that reason as being subject to the law, the law of which it sees itself as the author" (Kant, 2008, 31). Consequently, the exercise of human power must be legitimate power, and any processes of exercising power through social institutions or technological processes—as in the case of AI—must be controlled and managed by human beings in a democratically justified way.

However, these principles are very general and need to be implemented in concrete situations. Therefore, what can be considered ethically problematic at a concrete level varies from country to country and culture to culture. This is reflected, for example, in different levels of trust in technology and/or politics, different views on whether AI can produce more objective and fairer outcomes, or different ways in which societal value discourses take place or are facilitated, which indirectly affects whether people reject AI, simply accept AI technology, or support the idea that it will produce more fairness. While popular frameworks in AI, including machine learning and natural language processing include notions of group fairness (Dwork et al., 2012) and fair representations (Zemel et al., 2013), without stakeholder participation in an iterative process of design and evaluation, such fairness faces inherent limitations (Anthis et al., 2024). This suggests that, in reality and in everyday life, the design and implementation of (AI) technology and society's response to it—and its normative evaluation—are dynamic, culturally specific processes.

Contextuality and Universality: Empirical and Normative Ethics

Due to the ethical diversity of different cultures, research in AI-based social assessment must be sensitive to different contexts. Only by taking into account concrete situations in different contexts can general norms or values be applied in a responsible way. Therefore, any more concrete ethical reflection must be empirically grounded. Throughout the project, the normative perspective within each case study becomes most concrete when it comes to dialogue with policy-makers, which will also include scientific evidence or even specific recommendations derived from empirical research with vulnerable populations. Therefore, we need to reflect on how to combine descriptive and normative elements.

As mentioned at the beginning of this chapter, scholars in sociology and other empirical disciplines tend to reject normative perspectives, even when they acknowledge that there are "interests that guide knowledge" (Habermas, 1987), such as the motivation to organise a project or its social relevance. Sociology's main concern in the study of ethics and values is the descriptive analysis of existing moral norms and values, which is crucial for a deeper understanding of them: Population surveys or the analysis of people's actions can provide information about the moral norms that people are convinced of and the extent to which (ethical) values can differ within a society or between different societies. By generalising the differences, it is even possible to create groups of cultures, as described in the Inglehart-Welzel cultural map (Chap. 1). However, it is important to note that, paradoxically, refraining from making one's own judgements and perceiving the diversity of values in an unbiased way is often itself motivated by a general moral norm, namely, to take all people seriously with their cultural and individual characteristics and not to patronise anyone. The selection of the different case studies also aimed at including the diversity of cultural backgrounds, with the goal of exchanging views on a possible common understanding of justice and fairness. Therefore, the consideration and recognition of different contexts and the diversity of morals and norms should not be confused with moral relativism, which denies the possibility of any universal norm. Moreover, in the current political situation across the globe, the dangers of identitarian tendencies that understand norms and values only relativistically in relation to one's own position are becoming increasingly clear, so that there are now more and more academic authors who advocate a return to universalist positions (cf. Boehm, 2022).

Logically, it is not possible to infer norms from empirical data alone. It is not permissible to conclude from descriptions to prescriptions. In practical philosophy, as well as in applied ethics in various fields, scholars not only describe existing norms or values, but go a step further and want to argue in favour of the right norms or values in a prescriptive way. There are, of course, different philosophical theories for arguing in favour of moral norms and values. Utilitarian approaches aim to maximise utility, taking into account all human beings, or even all beings capable of suffering. Contractualists try to show that there are rules that correspond to a well-understood egoistic interest of all participants. In the tradition of Kant, many

scholars use the criterion of universalisability to identify moral norms corresponding to the well-known categorical imperative: "Act only according to that maxim which you can at the same time will that it should become a universal law". They claim universal validity for their results, knowing that all moral norms have their own history and cultural background. But the critical evaluation of particular norms can help to recognise whether or not they can legitimately claim universal validity. In this sense, a distinction must be made between the (particular) genesis and (universal) validity of norms.

Descriptive and normative ethics are both important components of any research on AI-based social assessment. At a very fundamental and motivational level, general normative notions of justice are important. But in order to understand different cultural contexts and different application contexts, it is important to describe what kind of moral concepts play an important role. Only by taking into account existing cultural differences and different contexts, it will be possible to find appropriate solutions. Only by referring to the existing ethical beliefs and values in different cultures (reflected in the sub-projects and case studies), these important and relevant particularities will be taken as a point of reference.

Discourse Ethics, *Safe Spaces,* and Stakeholder Dialogue

Diversity is not an end in itself. On the one hand, it is necessary to recognise cultural and other differences between groups and individuals. On the other hand, it allows a fruitful dialogue with all participants, which can lead to a better common understanding of central ethical concepts and goals. If all perspectives are brought together, common normatively relevant insights can emerge, possibly beyond the initial consensus, so that one could speak of a learning process.

It is important to understand that such a common view should not only be conceptualised as a kind of intersection or lowest common denominator of different views. In his debate with the later Rawls, Habermas rejected such an understanding of "overlapping consensus", which only looks at observably similar positions from different cultures, without any real argumentative debate between them that could be understood as a common learning process (Finlayson, 2019). However, in his history of philosophy (Habermas, 2023), he also rejects the view of Jaspers, who tends to understand philosophy as a kind of belief system. For Habermas, what is needed is a post-metaphysical process of deliberation in which different worldviews, including religions, can be integrated, but which seeks a common understanding without relying solely on authorities such as traditions, religions, or powerful individuals or groups.

Habermas' approach is based on an understanding of rationality as communicative rationality beyond instrumental and strategic rationality. In relation to ethics, he proposes an ethical theory of which most important element is a certain understanding of discourse (Habermas, 1990) as a procedure for finding out which moral norms can be regarded as justified. We are not in a position to reconstruct the

various stages of his theory, nor to discuss the arguments against it (but see Finlayson & Rees, 2023). We need to focus on the core of the proposal of discourse ethics. It is important to realise that not every discussion or dialogue is a discourse in his sense. Only a process of communication between different participants counts as a discourse, if it is aimed at reaching a rationally motivated consensus. Such discourses are necessary when there are moral conflicts, when ethical norms are challenged, or when new problems arise without a shared understanding of their ethical implications. Discourse then offers the possibility of renewing or replacing a problematised consensus.

In contrast to Kant, Habermas does not conceive of ethical judgement as a monological process, but consistently as a dialogical or discursive one. His first principle D therefore reads: "Only those norms can claim to be valid that meet (or could meet) with the approval of all in their capacity as participants in a practical discourse" (Habermas, 1990, 66). And in the discourse itself, the rule U for argumentation refers to the idea of universalisability as a kind of bridging principle. Only the rule that meets this criterion is accepted: "All affected can accept the consequences and the side effects its general observance can be anticipated to have for the satisfaction of everyone's interests (and these consequences are preferred to those of known alternative possibilities for regulation)" (Habermas, 1990, 65). D and U imply certain rules of discourse such as no discrimination, no exclusion, no coercion, etc. Habermas justifies these rules (at least in the quoted 1990 article) by his famous transcendental pragmatic argument: It is not possible to argue against these rules because one would be obliged to obey them if one really wanted to argue. Arguing against them would therefore lead to a performative self-contradiction, i.e. a contradiction between the illocutionary act of arguing and its semantic content. In this sense, these rules are inescapable. But for Habermas this is not an ultimate justification in the sense understood by Karl-Otto Apel. Nevertheless, the rules for a discourse that includes all possible participants without discrimination or coercion can be understood as an expression of the idea of the human dignity of all human beings.

Of course, discourse ethics can only justify rather general and abstract rules. The application of these rules in specific situations requires further reflection, which also needs discourses, so-called application discourses (Günther, 1993), with a similar set of discourse rules. It is also important to understand that real discourses will never be perfect, but can only approximate ideal conditions, and that the results of discourses will always remain fallible, so that learning processes are still possible. The consensus reached cannot be imposed on those concerned in an imperialistic way, because it is not a discursively generated result if they have not been involved. The dominance of one culture, e.g. Western culture, is therefore ruled out. What is reasonable cannot be decided in a one-sided process, but is the result of a discourse that is open to all contributions. Discourse provides the opportunity for, as Habermas put it, the "unity of reason in the diversity of its voices" (Habermas, 1996; cf. Gethmann, 2011). Using the concepts of Michael Walzer (2007), we can identify discursive universalism as an alternative to an imperialist or triumphalist kind of universalist thinking, which Walzer calls "covering-law universalism". But the other kind of universalism that Walzer favours, namely "reiterative universalism", does

not directly correspond to discursive universalism either. Walzer assumes that similar learning processes take place in all cultures, but with a particularist focus and based on a process of rethinking one's own tradition. What is neglected here is the fruitful process of intercultural exchange. Cultures are not containers, but always the result of highly hybrid influences. The boundaries between them are permeable. It is precisely this that makes encounters and learning processes possible. Connected to this idea there is the so-called "third space" proposed by Homi Bhabha (2006), where cultural exchanges occur, leading to new forms of cultural expression and identity. The latter can be considered important within the framework of AI for social assessment, as the counternarratives it produces are based on the "negotiation" of space as opposed to hegemonic discourses that homogenise culture and society.

One of the most important general conclusions to be drawn from these considerations is the decision that research on AI for social assessment should take a participatory approach, with a strong emphasis on vulnerable groups, who are the most affected by social injustices and the most important target groups of social services. In terms of such a discursive understanding of justice, the fair participation of all those potentially affected requires the organisation of meetings and workshops in safe spaces (see Chap. 2). In these safe spaces, people will not only present their own perspectives, but they will also argue for or against the views of others, share knowledge, and help each other to improve their way of seeing their own reality and the reality of others. For this reason, the case studies have sought to engage beyond traditional tech circles to include local communities, consumer associations, civil society groups, and cultural groups to ensure a rich tapestry of insights that inform the AI discourse. The active participation of these groups is essential not only for identifying potential risks and opportunities, but also for co-creating solutions that are equitable and culturally sensitive.

The concept of epistemic injustice may be helpful here: as mentioned above, the very issue of the (un)fair distribution of social goods, with or without AI-based technology, touches on central questions of social justice. Knowledge (epistemology) plays a central role in understanding what (in)justice is and how it affects individuals and collectives. Epistemic justice then suggests that everyone should have an equal opportunity to communicate their own views and values; no one should be excluded or inhibited from sharing their knowledge. Miranda Fricker (2007) (cf. Byskov, 2020) provides a compatible framework for addressing different forms of (in)justice mediated by people, organisations, laws and technology, drawing attention to knowledge as the normative bottleneck. She distinguishes between two types of injustice, namely testimonial injustice and hermeneutic injustice, both of which imply a prejudice against a particular speaker or speakers based on their social status, social origin, or appearance. While the former refers to a deflated level of "credibility" (Fricker, 2007) of a person due to prejudice (e.g. a person is not believed because he/she is black), the latter refers to the gap between a person's experience and the available "collective interpretive resources" (ibid.) to understand and evaluate a situation/action as an injustice (e.g. the concept of sexual harassment has only developed over time). This is crucial because the person cannot fully

comprehend his/her own experience, nor can he/she communicate it in an intelligible way to others.

The concept of epistemic injustice can be applied in three different ways. The first aspect—testimonial injustice—relates to the target groups and the research focus of this project: the so-called vulnerable populations and the assessment of people. People are heard by state officials and categorised as needy/not needy, etc.; therefore, the credibility of the person and their personal story is at stake. On the basis of the various case studies, it can be argued that these groups face or have faced forms of testimonial injustice, i.e. they are not believed or taken seriously by the general public or by administrative staff. Furthermore, the second aspect—hermeneutic injustice—which is of course related to the first, refers to the use of AI-based technology, which could increase this form of injustice: there is not only less space (in terms of empathy in human relations, social rights, but also time for communication) for the person to explain himself/herself, but these very conditions also affect the person's ability to make sense of his/her particular position (goals, objectives, ways of articulating oneself). The same applies to those who evaluate the person, because decisions are delegated to technology which, to a considerable extent, anticipates the outcome. The third aspect relates to the international and intercultural dimension, as illustrated in the sections before: criteria for good/bad decisions and good/bad AI are highly contingent across the globe—there is no one-size-fits-all approach to the concrete implementation of social justice that would be perceived as fair everywhere. Fricker's findings can also be applied here, as Segun (2021) points out, calling for a "global engagement with ethics in AI" because values and belief systems differ. At the same time, it is crucial to bear in mind that many contexts are characterised by significant power asymmetries. There needs to be a strong emphasis on co-designing technology with those who are most affected, and often those who do not have power (see Chap. 2). The project's approach, therefore, is fundamentally aimed at avoiding the use of AI for oppression, exploitation, or discrimination.

In addition to the importance of involving vulnerable groups to achieve more discursive rationality, it is also important to involve other stakeholders such as policy-makers and AI experts. On the one hand, this could help to improve research design and the development of AI applications. On the other hand, once strategies for improving AI have been developed, it will be necessary to communicate the results to those who decide on the future application of such technologies. Far from simply imposing ethical norms on policy-makers, one of the main goals of AI research for social assessment must be to focus on the integral role that stakeholders and policy-makers play in navigating the complex landscape of AI. Stakeholders in AI cover a broad spectrum, ranging from AI developers, users, and regulators to the communities directly affected by AI deployments. All of these groups have their own unique insights and concerns, as well as their own capacity for ethical judgement. Their voices are therefore seen as crucial to the development of ethical AI systems.

A comprehensive stakeholder engagement approach is considered to be of paramount importance to local and regional policy-makers for at least three reasons.

Engaging with a wide range of stakeholders provides them with a wealth of perspectives and information, enabling more informed and nuanced decision-making, which is particularly important in the rapidly evolving field of AI, where understanding the impact of the technology on society requires insights from different sectors. It also shows that stakeholder dialogues can help policy-makers identify and anticipate emerging challenges and opportunities in AI, ensuring that policies are not only reactive but also forward-looking. As AI policies need to strike a balance between fostering innovation and ensuring ethical, social, and legal safeguards, this type of engagement with stakeholders from different sectors helps policy-makers to understand the multifaceted impacts of AI and allows them to design policies that balance technological progress with societal welfare and ethical considerations. This is particularly important in the field of AI, which is characterised by rapid change.

Research Ethics

Understanding-oriented discussions, in which the above rules apply, are central to any kind of scientific practice. This is also true even for mathematics and the natural sciences (Kambartel, 1974). Where people are to some extent the objects of research and the knowledge-producing processes are integrated into a participatory approach, the corresponding rules, which are then intended to ensure respect for human dignity, apply in a particular way. This is particularly true in the case of a scientific project that is undertaken explicitly for ethical reasons and on which ethical aspects are to be reflected. Research ethics principles are even more important when vulnerable groups are involved. It is no coincidence that research ethics standards were first developed for medical research.

As research in AI for social assessment has a strong connection to qualitative research, especially with regard to the involvement of vulnerable groups, there is little need to justify why research ethics aspects are important. While research ethics is conceptualised or understood differently depending on the research context (cf. Iphofen, 2020), research in the social use of AI is based on the concept proposed by von Unger et al. (2014): “Questions of research ethics are an inherent part of empirical research practice and arise in all phases of the research process—from the choice of topic and objectives, study design, access to the field, data collection and analysis methods, to questions of publication and use of research results”. (ibid., our translation). From the perspective of project coordination, AI FORA’s research ethics guidelines are mainly based on the document “Ethics in Social Sciences and Humanities”, which was drafted in 2018 at the request of the European Commission (2021). In addition to the well-known principles of good scientific practice, transparency of roles and responsibilities, and informed consent, this document places particular emphasis on the ethical requirements of working with vulnerable groups and respecting different cultures. There are three main areas where research ethics concerns deserve particular attention:

Firstly, as mentioned above, special attention needs to be paid to context sensitivity, which refers not only to culture, but also to the contexts in which AI is used for social assessment. Context sensitivity is mainly realised in the case studies in terms of choice of topic, objectives, access to the field, and data collection. A key aspect that the Ethics Observatory aims to achieve is the transfer of how ethical issues are managed by establishing "adequate ethics monitoring structures" from the outset (von Unger et al., 2014, 22). As already explained, some of these—safe spaces and the inclusion of vulnerable people—are part of these very structures. From a case-study perspective, social science case-study partners and safe space coordinators are responsible for conducting (participatory) research in a responsible manner. This includes risk assessment (research sites, conditions of fieldwork in specific geographical areas) as well as awareness and application of case-study specific relevant criteria. For example, research with refugees requires different and/or additional measures than research with the unemployed. In addition to these context-specific criteria, qualitative/participatory research requires the informed consent of research participants (e.g. workshops, interviews), which should be ensured by the case-study partners.

Secondly, the use of technology/AI-based simulation to create better AI implies research ethics concerns, e.g. analysis methods, issues of publication, exploitation of research in particular. While AI-based simulations may help to uncover or resolve ethical issues, it is important to keep these processes transparent at the same time (cf. Al-Worafi, 2023; Wijse van Heeswijk, 2021) and ideally to have a feedback loop not only with policy-makers but also with vulnerable people themselves.

Thirdly, in relation to the first two aspects, it is desirable to obtain a self-reflection of the case-study partners on research ethics aspects, which may partly overlap with content-related challenges. From the perspective of this very chapter, these insights are crucial for assessing the extent to which an ethical consensus can (really) be reached. On the basis of these analyses—by the case-study partners themselves and by the Ethical Observatory—the validity and added value of the case-study approach will in turn be "checked" and substantiated with arguments. There will be differences as well as similarities between different countries/cultures, but these need to be made explicit and discussed with arguments for and against. This is also crucial for the particularly normative aim of communicating scientific findings or even making recommendations to policy-makers, e.g. regarding the question of whether challenges, pitfalls, solutions, etc. are national or transnational in nature (or: particular vs. universal).

Conclusion and Outlook

In conclusion, these meta-perspective reflections on normativity, research ethics, and dialogue with different stakeholders within the AI FORA project should help to monitor and anticipate ethical challenges, which is important from a project (coordination) perspective as well as from a content perspective. Following Chap. 2,

self-reflection/awareness is important to ensure the goals of inclusive technology co-design, i.e.:

- Value of discourses aimed at bringing cultural values into AI-based social assessment in national welfare systems
- Value of empowering vulnerable groups to achieve greater epistemic justice
- Value of openness regarding the implementation of project results, as the interpretation of concepts is again contextual, case-study specific, and culture-dependent.

More generally, these reflections should also contribute to a bottom-up assessment of ethical issues related to AI (Segun, 2021), together with AI FORA case-study partners. Our experience during the intensive collaboration among AI FORA research partners reveals interesting and important future perspectives on the main ethical and societal challenges of AI in the long term: even among the optimists, there is recognition of the significant journey ahead to better integrate ethical and social considerations into AI development, while emphasising the importance of human dignity and the complex relationship between humans and machines. This perspective is consistent with the cautionary stance against the unethical development of AI and the call for regulation to mitigate the potential risks of uncontrolled AI development. Although a sociological approach to ethics and values is largely descriptive, focusing on the empirical analysis of moral norms and ethical values as they manifest among stakeholders, this approach is important for two reasons. First, it is based on the principle of observing the diversity of moral beliefs and ethical values without imposing normative judgments, thus reflecting a commitment to respect cultural and individual differences. Second, it is motivated by a moral norm of taking all people seriously without bias or patronisation, thereby avoiding moral relativism while recognising the plurality of societal values. Thus, academic research can contribute to the development of a modern or late-modern society based on a deliberative process in a "socio-ecological" perspective, in which competing interests continually negotiate the validity of their claims and seek to reach consensus through mutual understanding and role-playing in practical discourses in an ever-changing complex environment.

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz in Germany, are gratefully acknowledged.

References

Al-Worafi, Y. M. (2023). Ethics in simulation research. In Y. M. Al-Worafi (Ed.), *Comprehensive healthcare simulation: Pharmacy education, practice and research. Comprehensive Healthcare Simulation*. Springer. https://doi.org/10.1007/978-3-031-33761-1_40

Anthis, J. R., Lum, K., Ekstrand, M., Feller, A., D'Amour, A., & Tan, C. (2024). *The impossibility of fair LLMs*. arXiv e-prints, arXiv-2406. https://doi.org/10.48550/arXiv.2406.03198

Bhabha, H. K. (2006). Cultural diversity and cultural differences. In B. Ashcroft, G. Griffiths, & H. Tiffin (Eds.), *The post-colonial studies reader* (pp. 155–157). Routledge.

Boehm, O. (2022). *Radikaler Universalismus. Jenseits von Identität. Unter Mitarbeit von Michael Adrian*. Propyläen.

Byskov, M. F. (2020). What makes epistemic injustice an "Injustice"? *Journal of Social Philosophy, 0*(0), 1–18. https://doi.org/10.1111/josp.12348

Catalan Observatory for Ethics in Artificial Intelligence (OEIAC). (2023). *Who we are*. Retrieved March 15, 2024, from https://oeiac.cat/en/who-we-are

de Wijse van Heeswijk, M. (2021). Ethics and the simulation facilitator: Taking your professional role seriously. *Simulation and Gaming, 52*(3), 312–332. https://doi.org/10.1177/10468781211015707

Deutscher Ethikrat. (2023). Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. *Stellungnahme*. Retrieved June 15, 2024, from https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226). doi:https://doi.org/10.1145/2090236.2090255.

European Commission (DG Research and Innovation). (2021). *Ethics in social science and humanities*. Retrieved March 30, 2024, from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-in-social-science-and-humanities_he_en.pdf

European Group on Ethics in Science and New Technologies. (2018). *Statement on artificial intelligence, robotics and 'autonomous' systems*. European Commission. Retrieved June 15, 2024, from https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1

European Parliament. (2023). *EU AI Act: First regulation on artificial intelligence*. Retrieved 15 June, 2024, from https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Finlayson, J. G. (2019). *The Habermas-Rawls debate*. Columbia University Press.

Finlayson, J. G., & Rees, D. H. (2023). Jürgen Habermas. In E. N. Zalta, & U. Nodelman (Eds.), The *Stanford encyclopedia of philosophy* (Winter 2023 Edition). Retrieved June 15, 2024, from https://plato.stanford.edu/archives/win2023/entries/habermas/

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gethmann, C. F. (2011). Reason and cultures. Life-world as the common ground of ethics. In H. Lange, A. Löhr, & H. Steinmann (Hg.): *Working across cultures. Ethical perspectives for intercultural management* (pp. 213–234). Kluwer.

Grimm, P., Trost, K. E., & Zöllner, O. (Eds.) (2023). *Digitale Ethik. Handbuch für Wissenschaft und Praxis*. Nomos Verlagsgesellschaft. 1. Auflage. Nomos; Academia (Nomos Handbuch). Retrieved June 15, 2024, from https://permalink.obvsg.at/AC16904593

Günther, K. (1993). *the sense of appropriateness: Application discourses in morality and law*. State University of New York Press.

Habermas, J. (1987). *Knowledge and human interests*. Polity.

Habermas, J. (1990). Discourse ethics. Notes on a program of philosophical justification. In J. Habermas (Ed.), *Moral consciousness and communicative action* (pp. 43–115). Massachusetts Institute of Technology.

Habermas, J. (1996). The unity of reason in the diversity of its voices. In J. Schmidt (Eds.), *What is enlightenment? Eighteenth-century answers and twentieth-century questions* (pp. 399–425). University of California Press (Philosophical Traditions, 7).

Habermas, J. (2023). *Also a history of philosophy* (Vol. 1). Polity Press.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Höffe, O. (2001). *Gerechtigkeit*. Beck.

Independent High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. Retrieved June 15, 2024, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Iphofen, R. (Ed.). (2020). *Handbook of research ethics and scientific integrity*. Springer International Publishing.

Kambartel, F. (1974). Ethik und Mathematik. In M. Riedel (Ed.), *Rehabilitierung der praktischen Philosophie* (Vol. 1, pp. 489–503). Rombach.

Kant, I. (2008). *Groundwork for the metaphysic of morals*. Translated and explained by Jonathan Bennet. Retrieved June 15, 2024, from https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf

Kaplow, I., & Lienkamp, C. (Hg.) (2005). *Sinn für Ungerechtigkeit. Ethische Argumentationen im globalen Kontext*. Nomos.

Kruip, G. (2004). Was ist soziale Gerechtigkeit? Grundsätzliche Überlegungen zur aktuellen Sozialstaatsdebatte in Deutschland. In J. Jans (Hg.), Für die Freiheit verantwortlich (221–237). Academic Press; Herder.

Liebig, S., & Lengfeld, H. (Eds.). (2002). *Interdisziplinäre Gerechtigkeitsforschung. Zur Verknüpfung empirischer und normativer Perspektiven*. Campus.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basil Blackwell.

Rafanelli, L. M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data and Society, 9*(1), 1–5. https://doi.org/10.1177/20539517221080676

Rawls, J. (2005). *A theory of justice* (Original ed.). Belknap Press.

Segun, S. T. (2021). Critically engaging the ethics of AI for a global audience. *Ethics and Information Technology, 23*, 99–105. https://doi.org/10.1007/s10676-020-09570-y

Sen, A. (2000). *Inequality reexamined* (6th ed.). Harvard University Press.

Shklar, J. N. (1990). *The faces of injustice*. Yale University Press.

UNESCO. (2021). *Ethics of artificial intelligence. The recommendation*. Retrieved June 15, 2024, from https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

von Unger, H., Narimani, P., & M'Bayo, R. (Eds.). (2014). *Forschungsethik in der qualitativen Forschung*. Springer VS. https://doi.org/10.1007/978-3-658-04289-9_2

Walzer, M. (2007). Nation and universe. Two kinds of universalism. In M. Walzer (Ed.), *Thinking politically. Essays in political theory* (pp. 183–199). Yale University Press.

White, S. (2010). Ethics. In F. G. Castles et al. (Eds.), *The Oxford handbook of the welfare state* (pp. 19–57). Oxford University Press.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International conference on machine learning* (pp. 325–333).

Chapter 4
Participatory Action Research for AI in Social Services: An Example of Local Practices from Catalonia

Albert Sabater, Beatriz López, Roger Campdepadrós, and Cristina Sánchez

Abstract After reviewing some experiences of AI use in social services in Catalonia due to its key role in developing the first AI strategy in Spain, our work adopts an ethics from below perspective through Participatory Action Research (PAR), engaging social workers, policymakers, technologists, developers, and social scientists to assess AI-based social services in Catalonia. Two key meetings under the umbrella of the AI FORA project illuminated the discourse, highlighting the importance of continuous, inclusive evaluation to maintain community relevance and integrity. Methodologically, we employed focus groups for a comprehensive exploration of the advantages and disadvantages of AI in social services, and a World Cafe to gain expertise and perspective around four central themes: (1) data-driven social services for resource allocation, (2) predictive analytics and early intervention, (3) evaluation and continuous improvement, and (4) stakeholder collaboration. The discussion highlights the necessity of continuous ethical reassessment and updates within organizations to maintain integrity, especially in the context of AI in social services, emphasizing the importance of aligning with evolving societal norms and technological advancements. Additionally, it emphasizes the role of organizational culture and digital literacy in adopting AI, advocating for a balanced approach that integrates technological innovation with human empathy and judgement, while addressing the challenges of ensuring ethical AI deployment through proactive, inclusive, and culturally sensitive practices.

A. Sabater (✉) · R. Campdepadrós · C. Sánchez
Facultat de Ciencies Economiques i Empresarials, Departament d'Empresa, Edifici Econòmiques, C/ de la Universitat de Girona, Girona, Spain
e-mail: albert.sabater@udg.edu; roger.campdepadros@udg.edu; cristina.sanchez@udg.edu

B. López
Exit Grup, University of Girona, Carrer Universitat de Girona, Girona, Spain
e-mail: beatriz.lopez@udg.edu

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_4

Introduction

The role of AI in social services significantly influences society's view of AI's integration and advancement. Thus, understanding public trust and societal attitudes towards AI advancements has become a vital responsibility for governments to foster public engagement in AI development, especially as AI begins to play a more prominent role in social services locally. In the Spanish context, although the overarching process of AI incorporation remains slow and thus mirrors that in other administrative sectors (Minguijón & Serrano-Martínez, 2022), it is expected that it will follow three distinct phases, as outlined by Ansón (2017). The first phase marks the initial commitment to a fully digital administration, a stage where many multinational companies based in Spain have progressed but is still developing in certain public administrations, including those within social services. This phase is characterized by a coexistence of traditional, partially computerized, and advanced AI-equipped systems, with the latter being in the implementation and testing stages, making minimal impact. The second phase involves transitioning towards an administration with high information management capacity, where AI-based systems replace traditional activities, requiring fully developed and interconnected information systems. The final phase, known as 'Intelligent Administration', will require comprehensive training and re-skilling of civil servants to handle automated management and social intervention tasks, with AI systems becoming the key players in public management as outlined in the Plan for the Digitalization of Spain's Public Administration 2021–2025.[1] According to the OECD, the implementation is expected to allow for the continual assessment and recalibration of AI systems to align with shifting social values and expectations, ensuring that the technology remains a tool for enhancing social good rather than exacerbating existing social divides (Berryhill et al. 2020).

In the analysis of the Spanish case, Minguijón and Serrano-Martínez (2022) highlight that a gradual and phased approach is critical, and that each step in this process should be firmly rooted in established principles and practices, ensuring that the evolution of AI in social services is both sustainable and responsive to the complex and dynamic nature of societal needs. More importantly for the social services, while AI might enhance the formalization of rules and procedures, potentially leading to more equitable service delivery and reducing unfair practices like systemic corruption or privileged access to social services, another scenario is possible as argued by Newman et al. (2022). Numerous experts warn that heightened automation might lead to a technocracy, a society dominated by rules and machines, where initiating or implementing changes to the system becomes challenging in a manner that aligns only with the organization's objectives, which often include enhancing efficiency, establishing legitimacy, or generating profit (Maris, 2022). Thus, the lack of a consistent and equity-oriented framework concerning AI systems

[1] https://portal.mineco.gob.es/RecursosArticulo/mineco/ministerio/ficheros/210902-digitalisation-of-public-admin-plan.pdf

has become increasingly critical considering the growing adoption of these programs in the digital infrastructures of states and their provision of public services (Walker, 2022), especially when sociotechnical knowledge is lacking (Jang & Landuyt, 2023). The good news is that societal questions have become increasingly acknowledged as crucial among developers of AI technologies, albeit they still show a rudimentary grasp of social dynamics (Joyce et al., 2021). A recurrent example of this are issues of bias in data and AI systems, predominantly perceived as technical problems by technologists when they are actually deeply rooted in preexisting social inequalities (Zajko, 2022).

This critical position does not only come from academics and (some) civil servants. As of 2022, the outlook on innovation in Spain was not so optimistic about technologies such as AI, with a majority (56%) of the population who think that it increases social inequality, compared to 30% who hold the opposite position (Zapata, 2022). In a similar vein, the results of the latest survey of social perception of science and technology from the same year in Spain (EPSCT, 2022) indicated that acceptance of AI is low as people think that the introduction of robots in the workplace contributes significantly to increasing the risk of unemployment (72.7%). While some studies on trust in AI in Spain critically point out that knowledge about big data and AI is moderate (Holgado et al., 2022), the focus on efficiency often overshadows the need for continuous and more detailed research on the perceived and actual benefits of AI, especially in terms of its societal impacts. This complexity necessitates a more informed and balanced discourse about AI, where its potential benefits are weighed against ethical and social considerations.

The current work employs social theory as an analytical framework for examining AI decision-making processes in social services (Joyce et al., 2021). Social theory serves as a prism through which citizens' interpretations of social provision are understood not only the positioning of individuals in relation to others within interpersonal dynamics but also their relative positioning vis-à-vis societal institutions (Bailey et al., 2020; James & Whelan, 2022). The viewpoint anchored in social theory aligns with the notion that an organization's culture achieves ethical robustness when its collective values are regarded as morally upright, not just within the internal confines of the organization but also in the external public sphere or 'downstream' when interacting with algorithmic systems (Christin, 2020; Noble, 2018). Such proactive collective approaches ensure that 'upstream' and 'downstream' processes are entangled, thus sociotechnical systems are more likely to be known 'from genesis to impact and back again' (Roberge & Castelle, 2021: 3).

One way to look at this proactive collective approach is through the collective urgency or Fear of Missing Out (FOMO) to adopt AI in public services, which arguably drives much of the adoption of AI in social services in Spain. Indeed, the AI FOMO in social services is also influenced by various factors across organizational, societal, and national cultural contexts. For instance, one may argue that, due to an organizational culture that value innovation and continuous learning, FOMO is experienced more acutely, although societal attitudes towards technology, underscored by values related to welfare and public service quality, as well as overall digital literacy, appear to shape the collective urgency to adopt AI in public services.

Of course, national and cultural dimensions like uncertainty avoidance, power distance, and the individualism-collectivism spectrum may further modulate this phenomenon as shown by Robinson (2020) in Nordic countries. These cultural layers interact and are expected to influence how organizations within social services navigate the balance between the allure of cutting-edge technology and the imperative to ensure ethical, equitable, and effective service delivery. But since the aim is to create a better, i.e. more accountable AI, one would expect civil society to have much more to offer than simply being the 'moral voice' of society in research and innovation (Ahrweiler et al., 2019).

Thus, a co-construction approach can be used so that there is a real, domain-specific interactive process between the developers and employers of AI systems and the communities that are subject to them. Without the adoption of such approaches, there is a very real possibility of power imbalance between those developing and deploying AI systems and the communities that are subject to them, and this situation can be further amplified when historically marginalized and underrepresented groups are not involved. Therefore, an approach that involves critical thinking and empirical research (a posteriori methods) based on social theory, with multidisciplinary teams of researchers, practitioners, policy makers, and citizens is needed to maximize equity and transparency (Lepri et al., 2018).

Experiences of AI-Based Social Services in Catalonia

In this section, we review some experiences of AI use in social services in Catalonia due to its key role in developing the first AI strategy in Spain, known as Catalonia.AI,[2] and for a multi-sectoral, cross-cutting people-centred plan that prioritizes the use of AI in sectors such as health and public services. Further, Catalonia showcases a significant level of digitization and AI implementation as shown by the fourth edition of the DESI Catalunya 2022 digital economy and society index, which applies the European Commission's methodology specifically to Catalonia. According to the report compiled by the Cercle Tecnològic de Catalunya and backed by the Department of Business and Labor of the Generalitat of Catalonia, Catalonia ranks fifth in Europe for digital advancement for the fourth year in a row, with a digitization rate of 63.7% (nearly 12 points above the EU average), thus at the forefront of digital development in Europe alongside countries such as Denmark, Finland, Sweden, and the Netherlands.

Although the use of AI in the realm of social services is not as common as in business (one out of ten companies already use AI systems in Catalonia in 2022[3]), there are nonetheless various examples of AI-based social services that have been

[2] https://politiquesdigitals.gencat.cat/ca/economia/catalonia-ai

[3] https://elmon.cat/moneconomia/es/innovacion/84-empresas-catalanas-incorporar-inteligencia-artificial-ia-2022-24207/

gradually integrated. The introduction of AI-powered tools for citizens as part of primary social service includes Barcelona's social aid simulator, which uses algorithms to assess individuals' socio-economic status and recommend appropriate social assistance. These notable AI-based social services are called the DPR system and are used by the City Council of Barcelona.[4] The AI system is designed to process the notes of social workers to categorize the issue and the individual's request, and to offer recommendations for responses, tailored to the available resources and services from the municipality. The DPR system uses machine learning and has been trained with data from 300,000 interviews conducted by Barcelona City Council's social services.

Gavà City Council's project Gavius[5] represents a cutting-edge initiative aimed at developing a smart, virtual assistant for local administrations, designed to inform citizens about relevant social aids using data in innovative ways. Prioritizing privacy and data minimization, Gavius employs AI techniques that avoid the use of individual data, while guiding users through necessary procedures to apply for social benefits. Thus, the system facilitates the processing and receipt of the small grants through a user-friendly mobile application, ensuring a quick, simple, and privacy-respectful experience. According to Gavà City Council, beneficiaries are identified via a highly secure and private mechanism, thus enabling them to apply for social benefits and manage applications, decisions, and potential aid directly through the system. The system is also used by the City Council of Mataró to provide social workers tools for personalized communication about aid options tailored to individual cases.

Sumar[6] is another project developed by social service organizations in Girona for information retrieval, data interpretation, and data visualization for social workers. Although the system is not an AI per se, it is clear that it aims to become one with the initial design and a common controlled vocabulary for information interoperability in social services. An important aspect of this project has been the harmonization of variables that have previously been worked on with social professionals, including people served, active records, beneficiaries of services, etc. as a way to facilitate the detection of patterns as well as possible errors and anomalies in the data. In Girona, there is also the PIGALL project,[7] which aims to improve the quality of social services by the City Council of Girona with criteria of effectiveness, efficiency, justice, and inclusion. An interesting element of the PIGALL project is that it will be based on a mixed-approach that combines quantitative and qualitative analysis of a wide range of available data in order to develop a better understanding, extract knowledge, and carry out general and specific actions to improve social services.

[4] https://isocial.cat/en/dpr-a-smart-tool-to-facilitate-the-work-of-primary-care-social-service-professionals/

[5] https://gavius.eu/the-project/?lang=en

[6] https://www.sumaracciosocial.cat/

[7] https://exteriors.gencat.cat/ca/ambits-dactuacio/afers_exteriors/ue/fons_europeus/detalls/noticia/20230313_ia-girona

Additionally, there are some AI-based projects in social services that have focused on very specific domains such as the Welcome[8] project or the Nidus[9] service. The former is based on developing a multilingual personal assistant tailored to support the reception and integration of migrants and refugees from Mediterranean and Middle Eastern countries. The AI-powered assistants interact with users either in real-world or augmented settings to facilitate information exchange and personalized assistance on various aspects of the reception and integration process, including support in administrative procedures as well as language and professional training. The latter project, NIDUS, is a digital platform for securely storing digital copies of crucial documents in the cloud for the homeless in order to address the critical issue of document loss often faced by those without stable housing. The AI tool also facilitates the completion of various administrative processes.

Although the great majority of examples demonstrate the local implementation of AI-based social services, some applications are also found at regional level. For instance, the Employment Service of Catalonia (SOC) also employs AI to review job listings on platforms like Infojobs and FeinaActiva and to predict a job seeker's chances of employment based on their profile, education, and experience, while also suggesting pathways to enhance their employability. Comprising several Basic Areas of Social Services (BASS) in Catalonia, there is the Digitalis[10] project, an AI-based social services designed to enhance communication and relationships among professionals and between professionals and users in the BASS. It includes a self-management service for users and group activity management and a virtual assistant for information requests that also enables users to access their social history and to manage interactions with social service professionals, including scheduling and communication. This AI-based tool also focuses on decision-making, planning, and evaluation for social workers by providing them dashboards for the 18 participating Basic Areas. Other Spanish regions that are developing AI tools for the provision services include Andalusia with the Cohessiona[11] system, Castilla-La Mancha with the HSU system,[12] and in Galicia with the HSUE.[13]

[8] https://isocial.cat/en/welcome-personal-assistant-for-newcomers/

[9] https://isocial.cat/en/nidus/

[10] https://isocial.cat/en/digitalissb-impuls-de-la-digitalitzacio-dels-serveis-socials-basics-de-catalunya/

[11] https://www.juntadeandalucia.es/presidencia/portavoz/social/148059/consejodegobierno/IgualdadPoliticasSocialesyConciliacion/Cohessiona/historiasocialunicaelectronica/convenios-municipales/diputaciones/carpetaciudadana/ventanillaelectronica

[12] https://historiasocial.castillalamancha.es/historiasocial/

[13] https://politicasocial.xunta.gal/es/areas/inclusion-social/servicios-sociales-comunitarios/historia-social-unica-electronica-hsue

Assessing AI-Based Social Services in Catalonia

As we have seen in the previous section, Catalonia has been at the forefront of integrating AI into its social service sector in Spain. While this integration marks a significant shift in how AI-based social assessment can potentially streamline processes, personalize services, and enhance resource allocation, there is also a risk of perpetuating existing biases, overlooking marginalized groups, and creating new forms of inequality. Thus, we consider it is important to provide a comprehensive understanding of how AI is reshaping the landscape of social services, highlighting both the opportunities and challenges that come with such a transformative approach, particularly for the most vulnerable who are, in turn, overrepresented in social services (Ahrweiler et al., 2024; Birhane, 2021).

For this purpose, we adopt an ethics from below perspective (Posada, 2021) through Participatory Action Research (PAR) to assess AI-based social services in Catalonia. This approach is particularly relevant when considering the roles of social workers in the middle of an AI transformation that impacts on vulnerable populations. Thus, we engage with several stakeholders that are related to various AI-based social services to understand their ethical viewpoints and assess whether the development and application of AI in social services is guided by the needs and concerns of those it serves. As part of the AI FORA[14] project, the PAR methodology was presented in two separate meetings on May 6th 2022 and May 4th 2023 to various stakeholders. They were initially recruited from various social service organizations involved in the AI FORA project. Each organization provided the contact details of at least one frontline and one administrative staff member to be involved in the AI FORA project. However, they were also encouraged to disseminate study information and invite interested individuals to contact the research team at the University of Girona for participation. Thus, purposive sampling with snowballing was employed, where participants were asked to refer colleagues who might be interested in participating in a two-step recruitment strategy (participants were first contacted either via email and/or telephone). After collecting various expressions of interest, we invited all those interested to attend two AI FORA meetings in 2022 and 2023.

The stakeholders that finally came to the meetings represented various segments of the population, especially marginalized or vulnerable groups (community organizations of migrants and advocacy groups for workers); professionals who administer social services and have first-hand experience with the practicalities and challenges of integrating AI into their work (social workers and service providers); government officials responsible for the implementation of AI in social services (policy makers); engineers, data scientists, and AI specialists who design and build AI systems (AI technologists and developers, mostly academics); and scholars who study the social, ethical, and legal implications of AI (social scientists). Two main groups of stakeholders were missing albeit they were also invited in both occasions,

[14] https://www.ai-fora.de/

namely companies that supply AI technologies and related services to city councils for AI-based social assessment and are responsible for the technical quality and performance of AI solutions used in social services (technology vendors); and public and private entities that provide financial support for AI projects in social services (funding agencies and donors). In order to encourage reflective and constructive discussions in private, the media or general public were not invited despite they play a role in shaping public perception and discourse around the use of AI in social services.

In the first meeting that took place at Montserrat Abbey, 15 local stakeholders that included social workers and service providers (7), policy makers (3), AI technologists (2) and social scientists (3) discussed the advantages and disadvantages (i.e. the status quo of AI) in social services through a focus group methodology. The main aim of the discussion of the focus group was to address the current state of AI use for social services and possible alternatives to the status quo based on participatory approach as a way to provide better use of digital technology in general and of AI in particular. Building upon the foundations laid at Montserrat Abbey, the second meeting was convened in a mid-size city recognized for its engagement in various AI projects for social services. This gathering, larger in scope, included 24 academics with computational and social science backgrounds from around the globe, alongside 10 local stakeholders comprising social workers, service providers, and policymakers. The participants were organized into four groups, each tasked with exploring different facets of AI-based approaches in social services: (1) data-driven social services for resource allocation, (2) predictive analytics and early intervention, (3) evaluation and continuous improvement, and (4) stakeholder collaboration.

The second meeting was particularly significant because it took place in real social services setting that has already experience in managing and delivering AI-based social services, thus providing a practical backdrop to the theoretical discussions from the initial Montserrat meeting.

It is important to highlight for the purposes of context that these two meetings were preceded by two serious games that resembled a typical AI-based social assessment situation in local social service provision. The first game was on social services related to employment management and unemployment benefits (gamification 1), whereas the second game focused on the management and allocation of benefits based on multiple complex needs (game 2). These serious games allowed stakeholders to gain further experiential perspectives on the integration of AI in decision-making in local social services both for the general population and for vulnerable populations. Most importantly, the two serious games were aligned with the ethos of PAR, which promotes collaborative inquiry and action for social change. By engaging a diverse group of stakeholders, the two meetings aimed to co-create participatory frameworks within the AI FORA project for building and critically discussing AI social assessment technologies in a user-friendly environment (the Better-AI Lab). The methodology and procedures of this comprehensive study are elaborated and described elsewhere (WP1 of the AI FORA project).

A Primer on Advantages and Disadvantages of AI-Based Social Services (First Meeting)

In the Montserrat Abbey focus group, the participants embarked on a comprehensive exploration of the advantages and disadvantages of AI in social services, ensuring that diverse perspectives, especially from the local context in Catalonia, were adequately represented and considered.

One of the key advantages of AI identified in this session was its unparalleled capacity for data analysis, with stakeholders acknowledging that AI's ability to process large volumes of complex data could lead to more informed decision-making in social services. This was seen as particularly beneficial in addressing the multifaceted issues faced by diverse populations, where traditional methods might fall short, especially for local services that increasingly deal with diverse and ageing populations. However, the discussion also brought to light several disadvantages and concerns associated with the implementation of AI, including the critical issue of AI perpetuating existing biases. Additionally, there were apprehensions about the potential dehumanization of social services, with an over-reliance on AI leading to a reduction in personal interaction and understanding. Throughout the session, it was evident that while AI was seen to present significant opportunities to enhance social services, there are also substantial challenges that need to be addressed. The following are the main key issues discussed and some quotes from stakeholders.

The discussions initially centred on the inherent advantages of AI, particularly its efficiency in handling both structured and unstructured data. This aspect appeared to be particularly important as stakeholders focused on the idea that AI might allow for the processing of a vast array of data types—from neatly organized databases to more complex, unstructured data like text, images, and even spoken language. From this perspective, a stakeholder (policy maker) manifested that '*if AI can process both structured and unstructured data, it is a game-changer for us in social services. We're dealing with a myriad of data types daily, ranging from structured demographic information to unstructured inputs like user interviews and case notes so AI's versatility in handling these diverse data formats is not just a matter of efficiency; it's about unlocking deeper insights into the needs and challenges of those we serve. This capability allows us to tailor our services more effectively to each individual's unique situation'*.

The use of AI to process and analyse large datasets rapidly was repeatedly mentioned as a potential significant advantage in identifying at-risk populations and emerging social issues in community health, employment, and housing and to proactively identify areas or groups that require intervention. Within this context, another recurrent issue in the discussion was that AI can handle repetitive and time-consuming administrative tasks such as data entry, appointment scheduling, and case file management more efficiently and effectively, and that the automation of administrative tasks could reduce the administrative burden on social workers, thus allowing them to focus more on direct client interaction and care. For instance, a stakeholder (social worker) said that '*if AI can help us reduce paperwork, we will be*

able to spend more time in the field, engaging with people face-to-face, understanding their needs better, and providing more compassionate and effective service'. This segment of the conversation recognized that in social service settings, each individual and community presents a unique tapestry of needs and circumstances, with stakeholders mostly hopeful that AI might stand as a transformative tool to customize services that align more closely with the specific requirements and contexts of each beneficiary.

As the discussion entered AI-driven personalized services, the majority of stakeholders highlighted both potential and challenges in how they should approach social care with AI. While it was repeated that the ability to harness data and shed light on the unique needs of each individual is promising, a degree of scepticism became particularly pronounced when the stakeholders delved into the issues of data and AI bias. The group collectively recognized that while AI has the potential to help through more personalized solutions, there is an inherent risk of these systems perpetuating existing societal biases if data biases are not dealt with. This was clearly a common concern among all stakeholders, who emphasized that AI data is still primarily dependent on historical statistical data. As one stakeholder (data scientist) in the group pointed out, '*the data we feed into AI systems is not free from historical prejudices. If unchecked, these biases could manifest in the AI's decision-making processes, potentially leading to unfair or discriminatory outcomes'*. The conversation also highlighted the challenge of ensuring diversity and inclusivity in the data used to train AI systems, particularly in relation to vulnerable populations that tend to be overrepresented in the provision of social services. At this point, a stakeholder (social scientist) remarked, '*Data is often not representative of the entire population, especially marginalized groups, and this lack of representation can lead to AI systems that are biased against these groups, further marginalizing them in the process of service provision'*. This comment sparked a discussion on the need for conscientious data collection beyond using historical data, with the group concurring that addressing these issues would require a multifaceted approach that is not only technical such as improving algorithmic transparency and fairness but also a broader engagement with diverse communities to understand and integrate their experiences into AI development.

Another issue that was clearly part of the conversation was the risk of becoming overly reliant on technological solutions within sectors that traditionally prioritize human interaction and understanding. In this sense, as AI technology becomes more integrated into social services, there is a growing view that it might overshadow the intrinsic human qualities that are crucial in social service provision, particularly interpersonal relations, emotional intelligence, and ethical considerations. As one stakeholder (social worker) put it '*it's crucial we don't lose sight of the human element—the very core of social services. It's about finding the right balance, ensuring we cater to the distinct needs of everyone in our community while remembering that at the end of the day, these are real people with stories that go beyond what data can tell us'*. Building on this critical viewpoint, the stakeholders delved deeper into the autonomy and privacy concerns associated with the use of AI in social services, particularly the ability of individuals to make informed decisions about their own

lives, emerged as a central theme. Hence, there was a palpable concern about the potential erosion of autonomy among social workers. At this point, a stakeholder (AI specialist) in the group raised a poignant question, '*While AI can guide us in the social service provision and provide support in the decision-making process, it is always important to have in mind the following: are we risking social workers' ability to make choices for themselves?'* This reflected an apprehension that an over-dependence on AI-driven recommendations might lead to a scenario where human agency is overshadowed by algorithmic decision-making.

Privacy concerns were another critical issue that came to the fore. The extensive data required to power AI systems, including sensitive personal information, raised red flags among the group of stakeholders. A stakeholder (policy maker) emphasized, '*In our quest to take advantage AI, we mustn't compromise the privacy of those we're trying to help'*. This statement underlined the ethical quandary of balancing the benefits of AI in providing tailored services with the imperative of safeguarding personal data. At this point, all stakeholders acknowledged that while anonymized data could mitigate some risks, the potential for breaches and misuse remained a significant worry. A stakeholder (social scientist) added, '*Every piece of data we feed into these systems is a fragment of someone's personal story. How do we ensure this information is used respectfully and responsibly?'* The concern raised by the social scientist stakeholder also focused on the handling of sensitive personal information within the context of AI-driven social services as data like social records and financial information are crucial for determining the kind of support an individual might need.

The stakeholder's assertion pointed to two key points: the need for robust data security measures, especially when AI is dealing with sensitive information; and the importance of ethical handling practices, which include obtaining consent from individuals before collecting and using their data and being transparent about how the data will be used as well as ensuring that the data is used solely for its intended purpose. Another topic that became recurrent in the conversations about possible disadvantages in AI-based assessments was the potential lack of cultural sensitivity and context understanding. For instance, as one stakeholder (community organization) put it '*an AI system might not fully comprehend cultural nuances in family dynamics or communication styles, leading to recommendations that are not culturally appropriate or effective. This lack of cultural sensitivity can result in services that are not only ineffective but also potentially harmful, alienating the very communities they are meant to serve'*.

Further, it was mentioned that with the growing use of AI in social services there was also the risk of widening the digital divide and this posed a substantial challenge because it risks of exacerbating existing inequalities rather than mitigating them. For instance, one of the most cited problems of the digital divide was that there still exists a disparity in access to computers, smartphones, and reliable internet connections and AI could potentially widen existing gaps in service provision if AI-driven services, which often rely on internet connectivity and digital devices, are less reachable for vulnerable populations. Within this context, one stakeholder (policy maker) stressed that '*the elderly, people with certain disabilities, and*

individuals from lower socio-economic backgrounds are likely to face higher barriers in this respect, which can prevent them from benefiting from AI-enhanced services'. Thus, it was highlighted that technology and AI services are often designed with a particular user in mind, which is particularly problematic for the provision of social services as this means that AI-driven innovations might only benefit a subset of the population, potentially neglecting those who might need these services the most. Finally, it also became evident that some stakeholders (social workers in particular) worried about the reduction in human interaction as a result of an over-dependence on AI. In this sense, since human interaction is a critical component of social work, it was noted that the quality of social services could suffer if AI replaces too much of the human element, as building trust and understanding with people often requires a personal touch that technology cannot offer. In this sense, a stakeholder said emphatically *'the empathetic, compassionate, and nuanced understanding that human social workers provide cannot be fully replicated by AI'*.

A Primer on AI-Based Social Services with Local Participatory Efforts (Second Meeting)

The World Cafe event served as a dynamic forum for stakeholders from various sectors to converge and deliberate on AI-based social services. These discussions, rich in expertise and perspective, revolved around four central themes: (1) data-driven social services for resource allocation, (2) predictive analytics and early intervention, (3) evaluation and continuous improvement, and (4) stakeholder collaboration. Each theme represented a crucial aspect of the collective effort to address the multifaceted challenges in the local provision of social services.

The first theme revolved around the pivotal role of data in enhancing the efficiency and effectiveness of social services by taking into account that (big) data plays an increasingly pivotal role, especially in the ability to utilize data to improve the targeting and allocation of social services. Stakeholders discussed leveraging demographic, socio-economic, and community-specific data to identify areas in dire need of social services. Whilst the discussion was mostly centred on how data could be harnessed to identify specific geographic areas or population groups most in need, various policy makers highlighted the need for collaboration with academia and civil society to ensure the integrity and applicability of data. They also stressed the importance of this data in shaping policies that are not only effective but also culturally sensitive and locally relevant. Academics brought to the table discussions on creating ethically-minded algorithms that could process complex data sets, while civil society emphasized their role in grounding this data with real-world insights and local nuances. Within this context, a stakeholder (policy maker) commented that '*by collaborating with civil society and academia, we ensure our data-driven strategies are grounded in reality, leading to equitable and transparent resource allocation'*. Although this statement highlighted the sector's commitment to

leveraging quantitative and qualitative data for more targeted interventions, from the perspective of vulnerability theory, both academics and social workers also brought to the table discussions on ethical robustness in data-driven resource allocation, which involves ensuring that the data used does not inadvertently perpetuate existing inequalities or vulnerabilities.

Following the second theme, which emphasized the proactive use of predictive analytics in social service provision, the discussions highlighted the importance of ensuring that the use of these tools for early intervention is coupled with safeguards against overreach and misinterpretation, particularly in relation to their potential use for social determinism. This perspective was seen as crucial when considering the deployment of predictive analytics in AI-based social services, as there is a risk that the models could lean towards a deterministic viewpoint, suggesting that individuals' futures are predestined by their past and current social behaviour. At this point, stakeholders (particularly social workers) called for an ethical approach to predictive analytics that involves not only technical accuracy but also a broader understanding of social determinants and their impact on individuals. Further, a stakeholder (social scientist), while acknowledging the importance of research in predictive analytics, brought to light a fundamental principle that often gets overlooked: the difference between association and causation. This distinction is critical in the context of predictive analytics as the academic elaborated, '*while our models can detect patterns and correlations in data, we must remember that correlation does not imply causation. Just because two variables are associated doesn't mean one causes the other'*. This insight sparked a broader conversation about the limitations and ethical considerations in the use of predictive analytics in the local provision of social services. The participants discussed instances where predictive models, based solely on associative data, might lead to interventions that are misaligned with the actual causative factors of a person's situation. Another stakeholder (also social scientist) in the group emphasized this point further, '*if we're not careful, we risk acting on false correlations, leading to interventions that could do more harm than good'*. This point served to stress the importance of complementing predictive analytics with qualitative research and human judgement to understand the underlying causes behind the identified patterns.

The third theme of the focus group revolved around the pivotal role of evaluation and continuous improvement in the realm of social services, especially in an era increasingly dominated by AI-based assessments. This theme was grounded in the understanding that for social services to remain effective and relevant, they must be subject to on-going scrutiny and refinement. As AI becomes more integrated into these services, the group recognized the heightened need for robust frameworks to assess their impact accurately and make necessary adaptations based on empirical evidence. In this sense, a stakeholder (AI specialist) in the group highlighted that '*in an age where data-driven solutions are becoming the norm, the significance of independent, objective evaluations cannot be overstressed. It's imperative that we continually scrutinize the efficacy of these services to ensure they are achieving their intended purposes and adhering to ethical standards'*. Hence an approach that involves continuous reflection on the impact of services, with a commitment to

making improvements that are responsive to the needs and rights of service recipients was called upon. Various stakeholders (particularly academics and civil society representatives) emphasized that this also means being open to public scrutiny and feedback, recognizing that ethical robustness is achieved not just through internal assessments, but through engagement with and accountability to the broader community.

Furthermore, the discussions emphasized that evaluation methodologies should not be confined to quantitative metrics alone. A stakeholder (social worker) added, '*while data and numbers are important, we must also consider the qualitative aspects of service impact. Stories, experiences, and feedback from the community provide insights that numbers alone cannot'*. This perspective was echoed by everyone but particularly from academics and representatives from civil society organizations who stressed their crucial role in this evaluative process. In line with the AI FORA approach, they advocated for a participatory approach to evaluation, one where community feedback is actively sought and incorporated. Hence, the group collectively called for evaluation frameworks that are holistic, encompassing both quantitative and qualitative data, and transparent in their methodology and findings. They argued that such comprehensive evaluations are essential for not just assessing the current effectiveness of services, but also for identifying areas of improvement and adapting services to better meet the evolving needs of the community. The consensus was that continuous improvement in social services, particularly in the context of AI and data-driven approaches, requires a multifaceted, inclusive, and transparent evaluation strategy, ensuring that the impact on communities is genuinely understood, valued, and addressed.

The final theme underscored the essence of the World Cafe event itself—the power and necessity of collaborative efforts among different stakeholders in addressing complex social challenges. The event presented a rich tapestry of ideas and strategies, emphasizing the power of data-driven approaches in social services and the vital role of collaboration across sectors. The discussions highlighted that effective solutions to social challenges require a multifaceted approach, combining the analytical capabilities of academia, the policy-making power of public administrations, and the ground-level insights of civil society organizations. A stakeholder (a civil society representative) summed it up by saying, '*our collaborative efforts are the backbone of developing comprehensive, data-driven solutions for social services that truly address the needs of vulnerable populations in housing, mental health, and migration'*. This sentiment was echoed throughout the discussions, with each group acknowledging the unique contributions and perspectives that others brought to the table.

Discussion

The viewpoint of stakeholders is clearly anchored in vulnerability theory and is aligned with the notion that an organization's culture achieves ethical robustness when its collective values are regarded as morally upright (Kruip et al., 2024). However, to maintain this ethical integrity, it is not sufficient for an organization to establish a one-time ethical framework. Instead, there is a need for periodic reassessment and updates of these ethical standards (Rothschild, 2016). This is particularly important in response to emerging ethical challenges and responsibilities in AI-based social services, which are often driven by evolving societal norms, technological advancements, and changing regulatory landscapes (Bradley et al., 2020; Jobin et al., 2019). However, as the prevalence of AI systems with human-like capabilities such as intelligent agents increases, the complexity and opacity of how these artificial agents learn and evolve introduce heightened levels of uncertainty and unpredictability. This inherent complexity is exacerbated by the fuzzy and dynamic boundaries of AI systems, which complicate efforts to ensure that the values and intentions of the designers are effectively embedded within these systems. This phenomenon of a 'value shift', where the operational values of AI systems may diverge from their initially intended values over time, poses a unique and significant challenge and highlights the importance of rigorous oversight to mitigate unforeseen impacts and ensure alignment with societal norms and ethical standards through responsive governance structures.

In the two meetings held as part of the AI FORA project case study, stakeholders clearly emphasized the necessity of balancing AI's analytical capabilities with human empathy and judgement. Both the focus group and World Cage signalled that the organizational culture in Catalonia plays a pivotal role in shaping the adoption of AI in social services, especially as cultures that champion innovation and calculated risk-taking are particularly prone to FOMO, partly driven by a desire to harness new technologies but also for a collaborative nature of social services, often characterized by networks and associations that may amplify FOMO through peer influence. In other words, witnessing the successful implementation of AI by counterpart organizations may intensify the urge to adopt similar AI technologies in social services. Nonetheless, since adopting AI within social services is deeply intertwined with the cultural fabric, particularly regarding attitudes towards technology, social values and, above all, digital literacy, only a higher level of digital competence may foster an increasing engagement with AI technology. Therefore, although our case study might exemplify a society with a warm embrace of technological innovation that often experiences heightened FOMO, partly driven by positive media narratives, the role of digital literacy along with the collective social values emphasizing welfare, equity, and the quality of public services cannot be understated enough. Of course, at the national level, cultural dimensions may further influence both AI adoption and the manifestation of FOMO in social services. For instance, on the one hand, the pronounced power distance is expected to make the AI adoption in social services predominantly flow from top-down directives,

with FOMO potentially being fuelled by leadership rather than grassroots demand. On the other hand, the collectivist cultures, which emphasize communal harmony and consensus, might experience a moderated form of FOMO, where the collective decision-making process naturally tempers the rush towards new technologies, thus fostering a more deliberate and inclusive approach to embracing AI in social services.

Such proactive steps could ensure that the organization remains responsible for the public good, aligned with contemporary ethical expectations as well as resilient to vulnerabilities arising from dynamic external and internal environments (Stivers et al., 2021). While AI specific regulations lack 'teeth' or meaningful enforcement (Zajko, 2022), this grassroots approach ensures that ethical considerations are not just top-down mandates but are woven into the very fabric of the organizational culture. Of course, it also necessitates some conditions: accountability to society, fostering moral autonomy and a mutual trust climate, and engaging in ethical deliberation (Martínez et al., 2021). But without active societal involvement in decision-making models, institutional measures can only partially address ethical principles in AI. Indeed, this scenario presents a fertile ground for further exploration and inquiry, particularly within the context of initiatives like the Spanish AI FORA case, where it is shown that there is an opportunity to delve deeper into the sociotechnical intricacies of AI systems in social services.

Addressing these challenges clearly requires a multifaceted approach that combines technological innovation with qualitative research as an important avenue to explore deep insights into assessing AI-based social services to offer a much more nuanced understanding of what is needed and that goes beyond traditional quantitative data. Other qualitative approaches such as ethnographic research has also provided useful and how AI systems might inadvertently perpetuate social inequalities or are misused for social control, especially affecting the less privileged as Eubanks (2018) cautioned. Thus, the application of AI ethics through qualitative methods to assess AI systems in social services should also serve to prevent AI technology from being used in what the same Eubanks (2018) calls 'low-rights environments', referring to the testing grounds that different organizations may use first for the poor but eventually for all.

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz in Germany, are gratefully acknowledged.

References

Ahrweiler, P., Gilbert, N., Schrempf, B., Grimpe, B., & Jirotka, M. (2019). The role of civil society organisations in European responsible research and innovation. *Journal of Responsible Innovation, 6*(1), 25–49.

Ahrweiler, P., Gilbert, N., Kampis, G., Jurány, Z., Sabater, A., Bicket, M., Luque, B., & Wurster, D. (2024). Using ABM and serious games to create "Better AI". In P. J. Giabbanelli, D. Ruiz-Martín, B. Oakes, & R. Cardenas (Eds.), *Proceedings of the 2024 annual simulation conference (ANNSIM'24), May 20–23, 2024*. American University.

Anson, A. (2017). *Las tres fases de la automatización de la administración pública*. Available at https://trabajandomasporunpocomenos.wordpress.com/2017/04/26/las-tres-fases-de-la-automatizacion-del-sector-publico/ [Consulted on 15/1/2024].

Bailey, S., Pierides, D., Brisley, A., Weisshaar, C., & Blakeman, T. (2020). Dismembering organisation: The coordination of algorithmic work in healthcare. *Current Sociology, 68*(4), 546–571.

Berryhill, J., Heang, K., Clogher, R. & McBride, K. (2020). *Hola, Mundo: la inteligencia artificial y su uso en el sector público*. Documentos de trabajo de la OCDE sobre gobernanza pública, 36. Available at https://oecd-opsi.org/wp-content/uploads/2020/11/OPSI-AI-Primer-Spanish.pdf [Consulted on 15/1/2024].

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns, 2*(2).

Bradley, C., Wingfield, R., & Metzger, M. (2020). *National artificial intelligence strategies and human rights: A review*. London & Stanford. Available at https://www.gp-digital.org/wp-content/uploads/2020/04/National-Artifical-Intelligence-Strategies-and-Human-Rights—A--Review_April2020.pdf [Consulted on 15/1/2024].

Christin, A. (2020). The ethnographer and the algorithm: Beyond the black box. *Theory and Society, 49*(5–6), 897–918.

EPSCT (2022). Encuesta de percepción social de la ciencia y la tecnología en España. : Fundación Española para la Ciencia y la Tecnología.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Holgado, P. S., Calderón, C. A., & Herrero, D. B. (2022). Conocimiento y percepción de la ciudadanía española sobre el big data y la inteligencia artificial. *Icono 14, 20*(1), 15.

James, A., & Whelan, A. (2022). Ethical'artificial intelligence in the welfare state: Discourse and discrepancy in Australian social services. *Critical Social Policy, 42*(1), 22–42.

Jang, K., & Landuyt, N. G. (2023). Limited benefits of technological advances in human service organizations: Going beyond the hype using sociotechnical knowledge management system. *Journal of Social Service Research, 49*(4), 426–446.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., & Shestakofsky, B. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius, 7*, 2378023121999581.

Kruip, G., Späth, E., & Sabater, A. (2024) "Ethical Aspects of Research on AI-based Social Assessment". In Ahrweiler, P. (Ed.), Participatory Artificial Intelligence in Public Social Services: From Bias to Fairness in Assessing Beneficiaries, Artificial Intelligence, Simulation and Society Series, Cham: Springer.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy and Technology, 31*, 611–627.

Maris, E. (2022). The humanities can't save big tech from itself. *Wired*. Available at https://www.wired.com/story/ethicis-big-tech-humanities/ [Consulted on 15/1/2024].

Martínez, C., Skeet, A. G., & Sasia, P. M. (2021). Managing organizational ethics: How ethics becomes pervasive within organizations. *Business Horizons, 64*(1), 83–92.

Minguijón, J., & Serrano-Martínez, C. (2022). La inteligencia artificial en los servicios sociales. *Cuadernos de Trabajo Social, 35*(2), 319–329.

Newman, J., Mintrom, M., & O'Neill, D. (2022). Digital technologies, artificial intelligence, and bureaucratic transformation. *Futures, 136*, 102886.

Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.

Posada, J. (2021). *Why AI needs ethics from below*. AI Now Institute. Available at https://posada.website/publication/why-ai-needs-ethics-from-below/Posada2021AINow.pdf [Consulted on 15/1/2024].

Roberge, J., & Castelle, M. (Eds.). (2021). *The cultural life of machine learning: An incursion into critical AI studies*. Palgrave Macmillan.

Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society, 63*, 101421.

Rothschild, J. (2016). The logic of a co-operative economy and democracy 2.0: Recovering the possibilities for autonomy, creativity, solidarity, and common purpose. *The Sociological Quarterly, 57*(1), 7–35.

Stivers, C., Pandey, S. K., DeHart-Davis, L., Hall, J. L., Newcomer, K., Portillo, S., et al. (2023). Beyond social equity: Talking social justice in public administration. *Public Administration Review, 83*(2), 229–240.

Walker, J. (2022). *You are part of the machine: Understanding algorithmic discrimination in artificial intelligence* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass, 16*(3), e12962.

Zapata, A. (Ed.). (2022). *Informe COTEC: Anuario 2022*. COTEC.

Chapter 5
Specialists and Algorithms: Implementation of AI in the Delivery of Unemployment Services in Estonia

Triin Vihalemm, Maris Männiste, Avo Trumm, and Mihkel Solvak

Abstract The case study examines the utilization of an AI-based tool to evaluate unemployed individuals who receive welfare services from specialists at the Estonian Unemployment Insurance Fund (EUIF). In this case, the machine collaborates with human decision-makers to enhance advising unemployed clients. Specifically, the automated decision-support tool provides background information to EUIF consultants by assessing the likely time when clients will find employment. This assessment is based on data related to the current labour market situation within the relevant segment for unemployed individuals, considering factors such as training, residence, and education. By analysing documents and conducting interviews with EUIF consultants, the authors explore various models for sharing decision-making responsibility between humans and machines based on the core values of AI implementation in Estonian society: effectiveness of information processing and the fairness of decisions made by machines compared to humans.

Introduction

This study contributes to the core question of the AI Fora project: How do specific values and cultural context impact AI-based social assessment for public service provision? Authors examine this issue from the perspective of professional culture and work ethos among public welfare systems' specialists. Despite lacking decision-making authority, these specialists significantly influence the local ecosystem of AI-based welfare services. Their acceptance or rejection of AI tools shapes public

T. Vihalemm (✉) · M. Männiste · A. Trumm
Faculty of Social Sciences, Institute of Social Studies, University of Tartu, Tartu, Estonia
e-mail: Triin.vihalemm@ut.ee; maris.manniste@ut.ee; avo.trumm@ut.ee

M. Solvak
Faculty of Social Sciences, Johan Skytte Institute of Political Studies, University of Tartu, Tartu, Estonia
e-mail: mihkel.solvak@ut.ee

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_5

perceptions of AI-related values, risks, and expectations. Moreover, it informs managerial deliberations on implementation policies and contributes to the formulation of local AI-related policy.

By analysing the empirical case of AI implementation in the Estonian public sector, this chapter explores various models for sharing decision-making responsibility between humans and machines. We focus on welfare specialists who were provided with AI assistants to support their professional activities. The discussion centres on the effectiveness of information processing and the fairness of decisions made by machines compared to humans.

Before delving into the detailed analysis, let us provide a brief overview of the local context in Estonia and the specific case under examination.

According to the Inglehart-Welzel World Value Survey, individuals in Estonian society evaluate personal ambitions, believe in technological progress, and respect individual autonomy (World Value Survey, n.d.). Notably, public perceptions of AI in Estonia, as portrayed in the media, are predominantly positive. Researchers have observed that the country's pro-technology stance played a crucial role in shaping a positive national identity after the Soviet occupation (Annus, 2022). This identity combines ecological values—highlighting Estonians' strong connection to nature—with new digital values, emphasizing the widespread use of digital services (ibid). Estonian author Valdur Mikita succinctly captured this fusion of tradition and modernity with his expression: "The true Estonian holds Skype in one hand and a small mushroom knife in the other" (Mikita & Kerge, 2017). This suggests that despite embracing urban digitization, Estonians continue their time-honoured tradition of foraging for forest berries and mushrooms (ibid). Regarding AI implementation, Estonian society exhibits most positive evaluations and homogeneity in associated values, compared to countries like Germany (Masso et al., 2023). Key values associated with AI adoption include security, accountability, privacy, justice, equality, and transparency (ibid). The general socio-cultural context is conducive to implementing AI when it is secure and just. The analysis considers both social justice and economic efficiency in the utilization of public funds.

The Estonian social welfare system operates on universalistic, equal solidarity funding principles, and its design and distribution of welfare services align with the Bismarckian low-spending welfare model. Established in 2001, the *Estonian Unemployment Insurance Fund* (EUIF) administers unemployment insurance benefits. Since May 2009, it has also organized labour market services to assist unemployed individuals in finding new employment. The EUIF operates as a quasi-governmental organization and functions as a legal entity under public law. Most of its services are subject to regulation by the Labour Market Services and Benefits Act (Riigi Teataja, 2023). It conducts its activities independently of the government, guided by a mission and operational rules defined by the Unemployment Insurance Act (EUIF, n.d.). Collaborating closely with various partners, the EUIF offers services to jobseekers, employees, employers, and young people. The organization operates public labour market services across 32 regional departments (EUIF, n.d.).

Annually, approximately 70,000 people register as unemployed in Estonia, each with diverse backgrounds, strengths, and obstacles. They can access EUIF services either by physically visiting the nearest EUIF office where consultants register them

as unemployed, or virtually through the EUIF website. Applications undergo scrutiny by an automated system, which determines, based on data provided by the unemployed individual, whether they are entitled to compensation (Raudla, 2020). Usually, the unemployed will first meet with the *consultant*, who will check if any data are missing or need to be corrected in the EUIF database. If the unemployed person has any *physical, psychological, or other problems which prevent finding a new job quickly*, she is redirected to a *case manager*. The case managers fall into two categories. The first-type case managers deal with clients requiring more intensive support in their job search endeavours. This involves collaborative analysis of the clients' situation, frequent client meetings, and leveraging multiple services to assist them. The clients under the purview of second-type case managers exhibit reduced working capacity. These individuals need assistance in identifying their strengths and overcoming barriers to securing suitable employment. EUIF statistics reveal that an employment consultant spends an average of 25 min per client and manages a portfolio of 220–250 individuals. Conversely, a case manager allocates approximately 45 min per client and oversees a portfolio of 100−120 individual clients (Kraavi, 2023). *Both consultants and case managers are referred to as welfare specialists in further analysis.*

The EUIF's 350 specialists meticulously analyse a substantial amount of information to create tailored plans for assisting individuals in their job search. While this process requires time, it is crucial to swiftly grasp the person's situation and initiate targeted efforts. The primary *information system* employed by EUIF staff in their daily work is the register of persons listed as unemployed and job-searchers and of the provision of labour market services (EMPIS-2), which can retrieve data from other Estonian registers. This integration streamlines data entry for specialists, eliminating the need for manual input. All systems utilized by EUIF are meticulously registered and further detailed within the state information system management system (RIHA).

The study focuses on *decision-support tool known as OTT* (an Estonian acronym for "decision support"). OTT is a *machine learning application that computes the probability of an unemployed person finding employment within 180 days and subsequently becoming unemployed again* within the same timeframe. Additionally, it identifies the factors influencing these two outcomes. The primary purpose of OTT is to assist consultants in providing tailored assistance to unemployed clients based on their individual needs. By leveraging a *scoring system*, OTT helps allocate time resources of specialists effectively. It reduces excessive support for clients who do not require it while increasing support for those who do.

OTT analyses individual jobseeker employment and unemployment histories spanning the past 5 years. It also considers socio-demographic and education data, ICT literacy, past and current benefit information, and membership in certain high unemployment risk groups. Additionally, OTT incorporates labour market information, including overall inflow into employment, outflow into unemployment, and the number of suitable vacancies in the region where the unemployed person resides. In total, 45 indicators contribute to its assessment. The analytical side of OTT operates on R software. Initially, it employed a random forest model, which was later replaced by a gradient-boosted decision tree model in 2022. OTT integrates data collection

from the data warehouse, pre-processing of data, model evaluation, and calculation of risk scores using machine learning techniques. The updated results are imported back to the data warehouse, where they become accessible to consultants. Interactive dashboards provide consultants with data on each individual client, including the client's overall portfolio risk distribution and various functional views of the model output. OTT's model undergoes retraining quarterly, while new scores are issued nightly to reflect the latest data changes for each unemployed person.

OTT emerged from a collaborative effort involving the EUIF and its development partners, including the University of Tartu's Centre of IT Impact Studies (CITIS). The project, titled "Profiling and Policy Evaluation Tools for EUIF", spanned from 2018 to 2020. Notably, the software development was undertaken by Nortal, while the data warehouse development was handled by Resta. EUIF developed and integrated the OTT system into its IT infrastructure in 2019. Subsequently, testing and piloting commenced in five branch offices during mid-2020. By October 2020, the system was fully deployed across all offices. The primary purpose of OTT is to provide welfare specialists with a quick summary of each client's situation. Furthermore, it offers an overview of all clients based on OTT's assessment, enabling the setting of priorities according to the extent of each client's need for assistance. Welfare specialists are also required to provide feedback on the scoring by assessing its accuracy for each client. They briefly explain why they believe the scores are either too optimistic or pessimistic. While the human decision-maker retains the authority to override the AI system scores, feedback statistics reveal that 93% of welfare specialists consider the OTT scoring to be accurate. In 4% of cases, they perceive the system as overly optimistic in evaluating client risks, while in 3% of cases, it is deemed to be too pessimistic (Kraavi, 2023).

Primarily designed for operational use within the EUIF, OTT's output also informs management decisions. A dedicated dashboard enables managers to examine client risk distributions and portfolio structures of employment consultants and case managers. If workload issues arise or certain case managers demonstrate greater expertise in assisting high-risk clients, reassignment becomes feasible. Furthermore, the system's output serves labour supply and demand analysis at regional and national levels by aggregating individual risk scores and distributions based on various filters.

From a management perspective, the OTT system has garnered resounding success. Notably, it won the Best Data-Driven Service Award in 2021. Furthermore, it serves as a prominent example of AI and machine learning application within the public sector, as recognized by the EC's Joint Research Council AI Watch (Misuraca & Noordt, 2020). This achievement has spurred the generation of ideas for future utilization and further experimental developments of tools. In new development, OTT scores will directly influence the choice of the first consultation channel for clients. Whether it is face-to-face, phone-based, online video counselling, or an online self-service environment, this decision will be partially automated, driven by risk assessment.

The case study was conducted about a year after the OTT system was fully deployed across all offices of EUIF. Then the welfare specialists were granted

autonomy to decide whether to incorporate the information provided by OTT into their real-world client advising. Given the tool's relative novelty, specialists' knowledge about it varied, influencing their internalization and interpretation of OTT. Consequently, the interpretation of research evidence must be approached with caution. No definitive conclusions can be drawn regarding whether AI assistance consistently shaped specialists' perspectives on clients and their advisory needs. Similarly, we cannot definitively conclude that AI assistance was never utilized. Our analysis is not cause-and-effect type experiment; it rather draws inspiration from the case study, highlighting issues related to the professional culture and work ethos of welfare specialists as they navigate the implementation of AI-based solutions in public welfare services.

The study aims to answer to the following research questions:

1. How do automated solutions and datafication transform the welfare sector and influence the decision-making processes surrounding social service provision? What options exist for reorganizing decision-making and integrating technology?
2. What were the experiences of specialists and middle managers at the Estonian Unemployment Insurance Fund during the implementation of the automated decision-support tool for welfare services targeting unemployed citizens?
3. What opportunities and risks arise from incorporating artificial intelligence into the decision-making process for welfare service provision in the future?

Case Study

Conceptual Framework

AI in the Public Sector and in Welfare Provision

In preceding decades, governments harnessed digitalization to enhance services by transitioning them online, thereby streamlining administrative processes and reducing paperwork. However, contemporary administrations now embrace a broader spectrum of digital technologies, integrating various AI tools into the creation and delivery of public services. These technologies span from facial recognition systems employed in policing (Zilka et al., 2022) to predictive models addressing future social service needs (Bright et al., 2019).

In a comprehensive context, AI encompasses "[a] collection of interrelated technologies used to solve problems autonomously and perform tasks to achieve defined objectives without explicit guidance from a human being" (Hajkowicz et al., 2019).

Scholars have coined terms such as the "digital welfare state" (Alston, 2019) and the "data welfare state" (Kaun et al., 2023) to describe the evolving landscape of social protection. Increasingly, digital data and technologies drive welfare policies, serving functions that automate, predict, identify, survey, detect, target, and even penalize (Alston, 2019: 3). Specifically, these technologies facilitate identity

verification, assess eligibility for welfare benefits, calculate entitlements, prevent, and detect fraud, assign risk scores, classify cases, and enable communication between welfare authorities and beneficiaries.

As new tools infiltrate civil servants' work, a streamlined bureaucracy emerges, where digital systems assume a more prominent role than their human counterparts (Zouridis et al., 2020). These civil servants, known as street-level bureaucrats, directly engage with citizens and implement governmental policies (Lipsky, 1980). However, the traditional face-to-face street-level bureaucracy has undergone a metamorphosis. It now exists as screen-level or system-level bureaucracy, where IT systems partially or fully take over decision-making tasks from human street-level bureaucrats.

Busch (2019: 115–116) categorizes automation into distinct levels: (1) limited automation, i.e. computers propose decision alternatives, but humans make the final choices; (2) considerable automation (infocracy), where computers analyse and execute decision alternatives, subject to potential human overruling; (3) canocracy or full automation, where computers analyse and generate decision alternatives; and (4) robocracy or complete automation, with computers making decisions autonomously.

Considine et al. (2022) propose five client–advisor–technology interaction models in public service provision: (1) technology-free model, where full face-to-face contact between street-level bureaucrats and clients; (2) technology-assisted model, where advisors use technology as decision-support tools; (3) technology-facilitated model, where clients receive some services without human interaction; (4) technology-mediated model, where interaction involves clients and technology, with advisors providing assistance during technical issues; and (5) technology-generated model or fully automated digital service provision.

The implementation of AI-based technologies in the welfare sector fundamentally alters the character of welfare provision and redefines the role of human service providers.

AI and the Discretionary Power of Service Providers

Social welfare specialists are highly educated professionals, who operate autonomously and expect professional trust (Giest & Raaphorst, 2018). However, the increasing delegation of decision-making to AI systems diminishes their autonomy and the value of their expertise. In public policy research, discretion grants specialists the freedom to make contextually fitting decisions within established constraints (Lipsky, 1980/2010). Yet, AI systems transform public sector officials into mediators rather than decision-makers (Wihlborg et al., 2016).

The implementation of AI systems can both support and disrupt human decision-making. It leads to the emergence of diverse administrative and fairness models (Carney, 2020). Employment administrations, influenced by managerialism, prioritize cost reduction through efficiency and efficacy (Penz et al., 2017). Street-level bureaucrats' discretion is shaped by input requirements imposed by automated

systems. Pre-programmed rules dictate case processing and decision translation (Zouridis et al., 2020). Consequently, programmers' discretion becomes central to public organizations, eroding specialists' autonomy. This may discourage civil servants from utilizing digital tools (Gofen, 2014).

For civil servants, the primary benefit of AI implementation lies in freeing them from mundane tasks, such as responding to frequently asked questions (Mehr et al., 2017). Ideally, this liberation allows them to focus on personalized care and other valuable activities (Sun & Medaglia, 2019).

Automated systems offer the potential to enhance decision-making by eliminating human arbitrariness, thereby improving accuracy, consistency, objectivity, and efficiency (Binns, 2022). However, due to the inherent limitations of mathematical models, AI systems may struggle to make nuanced judgements in individual cases or provide context-specific decisions (ibid). As demonstrated by Veale and Brass (2019), AI systems often function as "decision-support systems" rather than autonomous agents. While they inform and assist decision-making, the final verdict remains with specialists (Bullock, 2019). However, this arrangement does not guarantee superior outcomes. Specialists may become overly reliant on AI rationality (Young et al., 2019), failing to correct system errors and inadvertently amplifying automation bias (Peeters, 2020).

Nevertheless, most specialists critically evaluate AI-generated information alongside their professional expertise (Selten et al., 2023). Alon-Barkat and Busuioc (2022) contend that human decision-makers tend to trust AI recommendations that align with existing stereotypes and biases, often neglecting to rectify machine errors (Selten et al., 2023). Consequently, decision-support systems may inadvertently perpetuate discrimination against vulnerable groups, such as immigrants (Desiere & Struyven, 2021).

AI and the Citizen-Client Perspective

Citizens' direct interactions with various public sector organizations, ranging from law enforcement to social welfare, are increasingly digitized. These encounters necessitate proficiency in navigating automated systems. Access to digitized (un) employment services is contingent upon digital skills. When implementing AI systems in social welfare provision, adherence to the OECD principle of "nobody is left behind" becomes crucial. This principle ensures that viable alternatives exist for individuals lacking digital competencies or means to access online services (OECD, 2022:9).

However, evidence suggests that mere digital skills are insufficient for effective utilization of e-government services online. Citizens must also comprehend system functioning. As highlighted by scholars (Grönlund et al., 2007; Döring, 2021), administrative literacy—a seamless ability to navigate bureaucracy—is essential. Notably, encounters differ between traditional face-to-face settings and digital contexts. In traditional interactions, welfare specialists assume responsibility for proper

processes. Conversely, in digital encounters, citizens bear the burden of seeking information and understanding service processes.

The SyRI case (as summarized by Algorithm Watch 2020) exemplifies the challenge: even skilled citizens find it nearly impossible to question automated decisions when system transparency is lacking. Unaware of how their data influences decision-making, citizens face significant risks. In welfare service provision, AI systems pose complexities due to the unpredictable nature of human behaviour and the potential for long-term negative consequences arising from automated decisions.

The adoption of AI systems in public service delivery hinges on the trust placed by citizens and specialists in these automated tools, which are designed to enhance decision-making. Research indicates that trust varies based on system characteristics. For instance, citizens tend to trust AI for assessing job suitability due to its impartiality towards personal attributes like race and gender. However, when it comes to replacing human advisers with chatbots, trust diminishes (Starke & Lünich, 2020; Araujo et al., 2020).

Consequently, the utilization of AI systems necessitates a careful balance between potential benefits and risks. In the subsequent section, we delve into an empirical case study involving the implementation of AI within Estonian unemployment services. Within this context, we explore the opportunities and risks associated with different models of decision-making responsibility shared between human and machine actors.

Methodology of Case Study

The empirical case study combined individual in-depth interviews with (a) development and regional managers of EUIF and (b) welfare specialists, i.e. employment consultants and case managers working with clients. Additionally, desk research was conducted on thematic legislation, research reports, and other relevant documents.

Within the case study, four expert interviews were conducted with EUIF managers in October and November 2021. The expert sample included (1) the head of the EUIF regional branch; (2) the head of the jobseeker and employer services department at EUIF; (3) an internal training specialist within the jobseeker and employer services department; and (4) two OTT developers from the partner organization.

Nine interviews were carried out with welfare specialists of EUIF who use OTT in their daily work. These interviews spanned three regions: Tartu, Tallinn, and two counties in southern Estonia. The sample structure was determined based on insights from expert interviews that clarified EUIF specialists' working arrangements. Consequently, the final study sample included representatives from three specialist categories dealing with different client segments: At first, employment consultants, serving as entry points for clients, most frequently utilize OTT. They redirect newly registered unemployed individuals with a lower probability of promptly re-entering the labour market and requiring additional assistance and time in job searching to

case managers. For second, case managers who support clients with health-related work restrictions use OTT less frequently. Thirdly, two interviews were conducted with office managers responsible for organizing work, primarily using OTT as a management tool.

Throughout the study, research ethics principles were rigorously adhered to.

The interview guide encompassed inquiries about work experience at EUIF, changes in work arrangements, efficiency criteria, requisite (background) knowledge for effective jobseeker advising, experiences with OTT, the timing of OTT forecasts (prior to or after the initial client meeting), evaluations of OTT usefulness, mandatory assessment of OTT predictions, observed mistakes and biases in OTT forecasts, and necessary future developments to enhance OTT's utility for consultants.

Research team members conducted the interviews, which typically lasted 45–60 min. The interviews were facilitated via the remote communication platform teams. Notably, no recordings were made (and thus none were stored); interviewers instead took notes and completed anonymous memos following each interview.

Results of the Case Study and Implications for the Theoretical Model of AI Involvement in Welfare Service Provision

Contrasting Views of Stakeholders Regarding OTT

The EUIF decision-support tool OTT was launched in October 2020 and has since undergone regular updates by the development team. The development process and daily utilization of OTT involve various stakeholders. Key stakeholders associated with OTT include the central management team, IT system development team, data model/algorithm development team, local office managers responsible for organizing work, and specialists as end-users.

At the close of the first year of implementation, divergent perspectives and attitudes emerged among these stakeholders regarding the aims, functions, effects, and impacts of the OTT decision-support tool. The central management team perceives OTT as a decision facilitator, enhancing decision quality and serving as a tool to boost work efficiency. Additionally, it aids in predicting future labour market conditions, contributing to the development of improved labour policies. As succinctly expressed by one management team member:

> Thanks to the decision-support tool, our consultants can instantly decide which actions to take next, e.g. whether to help the person improve their computer skills or re-evaluate their work capacity/…/OTT has been in use for less than a year, but we can already say that we are very satisfied with the tool. Now we can make further developments based on practical user experience.

The system development team strives to ensure high-quality system usability, optimal user experience, and seamless compatibility with other EUIF systems. Meanwhile, the data model development team focuses on creating a machine

learning-powered tool capable of predicting individual-level unemployment risks. This tool also recommends effective interventions and training strategies for clients, with the overarching goal of achieving the highest possible prediction accuracy.

Local office managers, in turn, utilize OTT to monitor employment consultants' and case managers' client portfolios. For them, OTT serves as an organizational tool, evenly distributing workloads between consultants and providing crucial support to case managers handling complex cases.

Welfare specialists as end-users should derive the most benefit from the tool. However, interviews revealed that after using OTT for approximately 1 year, their engagement was lacklustre. The tool was perceived as an "object to be tested", and providing feedback on OTT accuracy was considered an additional duty:

> Basically, OTT for me is a duty to give feedback. I look at everything that OTT thinks will increase or decrease employment. I agree with some things.

Despite the widespread suspicious attitude, some interviews revealed more positive experiences and opinions. The decision-support tool in some cases could be useful and the predictions provided by the system might give a new viewpoint about the client for further decisions. It could also be useful for employment consultants to get an overview of client characteristics:

> Well, he gives such a general picture that you can at the same moment see who has been employed for a long time, which reduces the probability of finding a job.

Nevertheless, in their daily work with their clients, specialists seldom used OTT. As one of the interviewed specialists said:

> Honestly, I have to say that such use, such analysis is limited. The use of the system is not as active as it could be. There are several reasons for that, but the fact that consultants cannot see concrete benefits seems to be the main reason.

The transparency of decisions made by the AI system posed challenges for end-users. As highlighted in an interview with another specialist:

> It is possible that certain features of OTT remain unclear. A more comprehensive explanation of how OTT provides support, generates data, and facilitates interpretation is necessary. Although I have reviewed the 22-page manual, it has undergone updates over time. Additionally, I attended training sessions on OTT, but further training and detailed explanations may enhance overall support.

While usage practices were still evolving, insights from interviews with end-users informed management decisions. Organizational steps were taken to further emphasize the benefits of the AI system. However, the interviews also underscored the fundamental challenges associated with using automated decision-making in unemployment services, particularly in terms of sharing responsibility between humans and machines.

The Impact of Human Interaction on Service Efficiency

Despite diverse educational and professional backgrounds, the interviewed specialists shared a common mission: helping people. They identified themselves as social counsellors or social workers. For them, establishing trust-inspiring connections, empathetic listening, and understanding clients were of paramount importance. Their primary goal was to support clients in achieving their objectives and navigating various challenges. While finding employment was often part of this process, it was not always the sole focus. Specialists handling complex cases emphasized that the speed of reintegration into the labour market might not be their primary objective. As expressed by one specialist:

> My main goal is to help the clients go their own way. Sometimes it is not getting a job at all. There are also goals where the client may not be able to go to work and may need to perform a new work ability assessment instead.

The professional responsibilities of specialists encompassed various administrative tasks: data collection, verification, updates, and information mediation. They also crafted action plans for clients, documented interactions, and engaged in counselling. Administrative duties consumed significant time, leaving limited room for actual counselling. On average, sessions lasted approximately 30 min, occurring once every 30 days. Regrettably, according to interviewed specialists, this time constraint adversely impacted service quality:

> The clients will know if you do not have time for them. They are less confident and cooperative.

OTT failed to assist specialists in managing administrative duties to allocate more time for meaningful client communication. Consequently, they perceived limited benefits from its use. The inability to incorporate health-related information or enhance the AI system's assessment by considering client motivation for re-entering the job market fostered scepticism among end-users regarding adherence to OTT prognoses:

> It's important to understand the client, to put yourself in his/her situation. The system doesn't see what happens inside the person.

In general, specialists strongly believe that human interaction, empathy, and a personalized approach to clients are pivotal for success. While OTT scores were perceived as static, specialists placed greater importance on the subjective information they personally collected. The inability of OTT to incorporate specialists' additional input—such as motivation assessments or case complexity—along with the absence of dynamic evaluations or simulations hindered meaningful collaboration with the system.

Need to Integrate Different Information Systems and Mandatory Feedback on the System

In their professional roles, welfare specialists draw upon data from diverse sources. The central information system, EMPIS-2, serves multiple functions. It acts as an environment for data entry and provides individual client profiles, encompassing socio-demographic information, employment records, and details about interactions and activities related to EUIF. OTT, the decision-support tool, is integrated within EMPIS-2. However, specialists encountered challenges in distinguishing between the functions of OTT and EMPIS-2, often conflating them.

For specialists accustomed to working directly with people, the underlying connections to machine learning, quantitative modelling, and statistical profiling—the true functions and mechanisms of OTT—remained unclear:

> Especially in the beginning, a lot of questions arose. I don't like that kind of prediction and assumption at all.

Specialists had also problems understanding the cause–effect relationships used in the model and they were quite suspicious of the decision-support tool. For them OTT was like a new unknown acquaintance that they needed to get to know:

> OTT somewhat perplexes me. I do not understand how it works.

For office managers responsible for organizing specialists' client portfolios, the function of OTT—arranging clients within consultants' portfolios based on the probability of finding jobs—was deemed to be advantageous. This capability allowed them to categorize clients as either "simple" or "complicated", facilitating strategic planning and the proposal of appropriate measures:

> I really like this list (the one where they are ranked from most likely to least likely to be employed); it gives a complete picture, complete context.

For head specialists and managers, OTT served as input for developmental interviews with personnel and facilitated the compilation of client portfolios for consultants. Additionally, specialists were tasked with evaluating the proposed probabilities of job placement—assessing whether they were accurate, overly optimistic, or unduly pessimistic. This assessment occurred between the 35th and 65th day of a client's unemployment period. Notably, the interviews revealed that most specialists hesitated to provide feedback on the system's predictions. However, recognizing the need to enhance prediction quality and accuracy, feedback was mandated for specialists.

According to the interviewed specialists, OTT's prognoses were predominantly accurate. Despite this overall assessment, some predictions drew criticism. The critique centred on two main points: (1) incomplete data model—the existing data model lacked essential aspects that influenced job placement probabilities, such as motivation and health-related information; (2) static prognoses—OTT predictions remained fixed at specific time points and did not adapt during the client engagement process. The current system lacks consideration for factors beyond its scope.

A person is a complex entity, encompassing attitudes and concerns that elude quantification by the system.

To advance the application of AI systems, two avenues emerge: (1) enhanced integration, i.e. achieving greater synergy between diverse information systems and AI-driven decision-support tools, and (2) educational and skills enhancement: empowering welfare specialists with deeper comprehension of automated decisions.

These measures are essential for bridging the gap between system limitations and the holistic reality of human experience.

Technology Acceptance by Specialists

User acceptance and confidence play pivotal roles in the further development of any emerging technology. Numerous frameworks and models have been devised to elucidate user adoption of novel technologies. These frameworks predominantly draw upon the theory of reasoned action (Fishbein & Ajzen, 1975), which delineates the interplay between attitudes and behaviours in human actions.

To contextualize interview results within the realm of technology acceptance, we applied the unified theory of acceptance and use of technology (Venkatesh et al., 2003; see Fig. 5.1). This theory posits four key constructs: (1) performance expectancy, i.e. anticipated effectiveness of the system; (2) effort expectancy, i.e. perceived ease of system use; (3) social influence based on communication with peers; and (4) enabling conditions, i.e. availability of necessary resources. Gender, age, experience, and willingness to engage with the system act as moderating factors influencing behavioural intention. Notably, the sample of interviewees in the current case study exhibited homogeneity, resulting in an insignificant impact on behavioural intention.

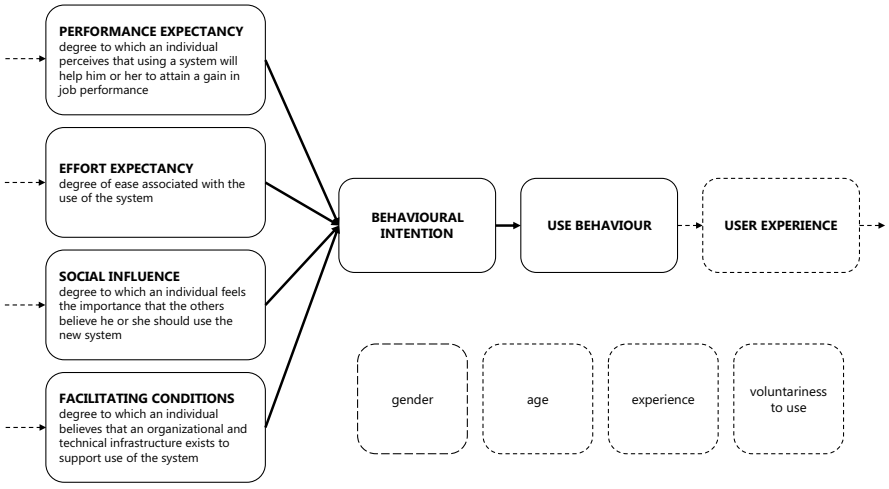

Fig. 5.1 Unified theory acceptance model (Adapted from Venkatesh et al. 2003)

Busch (2017, 2019) proposed the Digital Discretion Acceptance Model (DDAM), building upon the Unified Theory of Acceptance and Use of Technology (UTAUT). In this model, the central concept is the Perceived Importance of Discretion (PID), which signifies the degree to which specialists perceive professional judgement as crucial for decision-making. As our previous analysis indicated, while specialists acknowledged the efficiency of decisions that consider all potential reasons for unemployment, such comprehensive assessments were time-consuming. This suggests that system use can be more efficient when clients are segmented, and automated decision support is not universally applied. Performance expectancy emerged as the most relevant factor explaining the modest adoption of OTT. As previously mentioned, specialists assume the role of "social counsellors", tasked with thorough client analysis—a task not yet achievable by a mere "machine". Despite some positive experiences and beliefs, prevailing suspicion and confusion outweighed these factors. To enhance performance expectancy, it is imperative to bolster knowledge management strategies within the EUIF. Regular training, opportunities for experience sharing, and effective motivation strategies for staff are essential to maximize the effective use of the decision-support tool.

Effort expectancy directly correlates with system usability and user experience. The interview findings do not reveal any significant issues in this regard; minor inconveniences encountered could likely be resolved by the system development team.

Regarding the widespread modest use of the system, the argument that "others do not use it either" (negative social influence) strongly reinforces negative intentions towards system adoption. The impact of unwillingness is contentious: while it imposes a formal obligation to use the system, it often leads to reluctance and perfunctory (non-useful) task performance. Several strategies can mitigate negative social influence and promote positive adoption: (1) recognizing and disseminating success stories from effective system users; (2) highlighting those who adeptly utilize the system as role models for others; and (3) linking system use to career advancement, i.e. considering skilful system use as a criterion for professional growth.

The conditions facilitating behavioural intention partially overlap with the aforementioned factors. Providing comprehensive training, accessible user manuals, and straightforward instructions enhances willingness to engage with the system. However, the existing instructions were complex, training occurred long before implementation, and crucial aspects were inadequately covered, leaving insufficient time for meaningful feedback. To address this, consider revising the requirement for compulsory feedback, integrating it seamlessly into specialists' workflow. Delegating feedback responsibilities selectively or during specific periods could offer a pragmatic solution. Additionally, reorganizing consultants' tasks to allocate sufficient time for system-related activities would be beneficial.

Conclusion

This case study sheds light on the implementation of artificial intelligence (AI) within the Estonian public sector, specifically focusing on welfare specialists. The AI tool's information was strategically employed for managerial decision-making, resulting in the equitable distribution of consultants' workloads and increased consultation time with clients facing extended periods before re-entering employment. The overarching objectives of AI implementation were twofold: operational efficiency and the reduction of long-term unemployment.

However, this study primarily addressed one dimension: service time. While the AI-supported assessment positively impacted efficiency, its effectiveness in content-wise aspects of consultation services remains uncertain. Specialists exhibited resistance to using the information provided by the OTT tool for two reasons: organizational challenges, such as insufficient training and additional tasks, and the tool's static design, lacking opportunities for specialists to contribute additional insights.

Notably, the study did not explore other critical aspects, including service quality and fairness. Furthermore, the desired objective—specialists seamlessly integrating AI-derived employment probability data with subjective information about jobseekers' motivation, health, and situational obstacles—remained unfulfilled. The potential of a more dynamic and interactive AI tool for client servicing remains an open question. The active participation of specialists in the design of AI systems holds promise as a strategy to enhance technology acceptance. When AI systems are exclusively top-down creations, there exists a risk that they may not align with the actual needs of specialists, existing work processes, or citizen requirements. Notably, specialists reported encountering new tasks resulting from AI implementation, which were not seamlessly integrated into their daily workflows.

This empirical case study serves as a catalyst for theoretical discussions regarding the opportunities and risks associated with integrating artificial intelligence into decision-making processes for social service provision. Drawing inspiration from prior research (Busch & Eikebrokk, 2019; Busch, 2019; Considine et al., 2022), we propose a conceptual framework for navigating choices in social service delivery (see Fig. 5.2). This framework considers several critical factors, including case complexity, expected decisions, digital accessibility, skills, and the mental readiness of end-users—both unemployment service clients and welfare specialists. Figure 5.2 illustrates that a one-size-fits-all approach to technological solutions may not yield success. Instead, tailoring technical interventions to direct different cases and clients towards appropriate channels proves more effective. By acknowledging the multifaceted landscape of social services, we can foster AI systems that truly align with the diverse needs of stakeholders.

The complexity of cases, as depicted on the vertical axis of Fig. 5.2, depends on several interrelated factors hindering job finding. These factors include legislative vagueness, case uniqueness, and sensitivity. Alongside multifaceted and complex cases, there are simpler situations that necessitate formal administrative decisions.

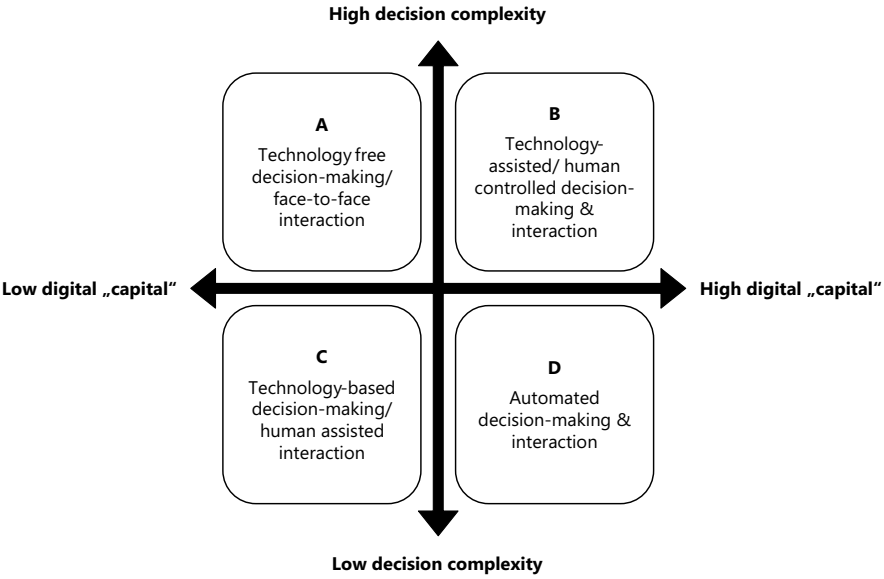

Fig. 5.2 Conceptual field of choices for social service provision. Authors' own work

The horizontal axis represents "digital capital", encompassing access to technical resources, knowledge, skills, and willingness to use them. This dimension distinguishes end-users with limited digital assets and inadequate skills from those proficient in interacting with technological systems. In more intricate scenarios, both clients and specialists deem human contact and decision-making essential, particularly when digital capital is constrained on both sides (quadrant A in Fig. 5.2). Solving complex cases involving digitally empowered end-users may benefit from technology and data-driven decision support, but ultimate decision control should rest with human experts, whether street-level or screen-level bureaucrats (quadrant B). Conversely, less complex cases involve formal administrative procedures like unemployment registration, service applications, and information checks. These cases align well with standardized AI-based solutions, which do not require specialist discretion. Depending on user digital capital, fully automated systems (quadrant D) or human-assisted systems (quadrant C) can be employed.

The model for utilizing AI systems to support human advisors, who act as intermediaries for all services, ensures high-quality handling of complex unemployment cases. However, it necessitates service providers with greater digital capital. Achieving this requires educating and professionally training welfare specialists to comprehend the operational principles and evaluations of AI systems. The adoption of data-driven decision support has the potential to enhance decision quality and fairness. To mitigate decision-making bias, screen-level bureaucrats should be trained to critically assess AI decisions. When AI systems are designed to be more interactive, they can provide additional assistance in human decision-making, including the incorporation of individual-specific information. While implementing

this approach would enhance professional efficiency, it poses challenges in terms of managerial efficiency.

The utilization of AI systems in welfare service provision offers the potential to enhance managerial efficiency and grant digitally capitalized clients greater spatial-temporal flexibility. However, it is crucial to recognize that digitally non-capitalized clients remain highly vulnerable. To address this, a critical feedback system on AI decisions must be developed and integrated into the service system for continuous improvement. The proposed segmented approach to welfare service delivery involves AI supporting digitally capitalized clients in less complex cases, where some services can be mediated without human intervention. Simultaneously, human welfare specialists play a pivotal role in assisting digitally vulnerable clients and handling intricate cases that necessitate individualized solutions. By adopting this model, specialists can allocate more time to complex cases. Notably, the segmented approach holds promise, allowing for the achievement of both managerial and professional efficiency, particularly when implementing team-based case management.

Conclusion and Outlook

This case study investigates the integration of AI within the Estonian public sector, with a specific focus on welfare specialists. These professionals were equipped with an AI assistant named OTT to enhance their advisory activities for unemployed clients. The primary objectives of AI implementation were twofold: efficiency and effectiveness. The case study yielded positive results in terms of operational efficiency because office managers utilized AI to plan specialists' schedules and workloads, thereby allowing for increased consultation time with clients facing extended periods before re-entering employment. However, the effectiveness dimension—specifically service quality and fairness—fell beyond the scope of this specific case study due to encountered limitations. Firstly, definitive conclusions regarding the consistent impact of AI assistance on specialists' perspectives towards clients and their advisory needs remain elusive, primarily because the consideration of OTT scores was voluntary. Secondly, the study relied on interviews rather than on-site observations, limiting the analysis of real-life behaviours. Nonetheless, the authors theoretically explored various models for distributing decision-making responsibilities between humans and machines.

This chapter highlights the initial findings from the Estonian case study, which focused on welfare specialists' experiences. The subsequent phase of the AI FORA project in Estonia will shift its focus more explicitly to citizens and their requirements concerning AI systems in the context of (un)employment services. Particular attention will be directed towards potentially vulnerable groups to gain a deeper understanding of their interactions with and needs related to AI systems.

Despite Estonia's ongoing development of various AI solutions over the years, empirical studies that comprehensively explore both the utilization and implementation of these systems within organizations, as well as their perception by citizens,

remain scarce. Consequently, the results of this study offer valuable insights not only for the specific public sector institutions planning AI system implementation but also for a broader audience.

The study's outcomes have already been shared with the management of the Estonian Unemployment Insurance Fund (EUIF), enabling them to enhance their implementation processes. Additionally, we intend to present these findings at various conferences and on our university's webpage. Furthermore, given that many future public sector specialists and experts are currently enrolled in courses or pursuing degrees at Tartu University, this case study will serve as a valuable resource for educational purposes.

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

Appendix: Sample of the EUIF Consultants' Interviews

Professional position	Working time in EIUF	Education	How many clients one has to handle currently	Age group (until 29, 30–50, 51–70)
Case manager I	1.5 years	Social work	50–100	30–50 (30)
Employment consultant	2 years	Cultural management and social work	Ca 100	30–50
Employment consultant	16 years	Previously been medic, worked as local government social advisor	100–150	51–70
Employment consultant	2 years	Not known	Ca 200	30–50
Employment consultant (works also as a trainer/mentor for new employees)	3.5 years	HR specialist and tailor-stylist	100–150 (because teaches others, otherwise same position has 250–300)	30–50
Case manager II	4 years	Social work	50–100	30–50
Case manager II	15 years	Biology	50–100	30–50
Head consultant (formerly worked as employment consultant)	6 years	Business management	11 people in her team (consultants she manages)	30–50
Head consultant	Not known	Not known	14 people in her team (consultants she manages)	30–50 (30)

References

Algorithm Watch. (2020). How Dutch activists got an invasive fraud detection algorithm banned. *AlgorithmWatch*. Retrieved 15.11.2023, from https://algorithmwatch.org/en/syri-netherlands-algorithm/

Alon-Barkat, S., & Busuioc, M. (2022). Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory, 33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Alston, P. (2019). *Report of the special rapporteur on extreme poverty and human rights*. A/74/48037. Retrieved 11.09.2023. https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25156

Annus, E. (2022). A post-soviet eco-digital nation? Metonymic processes of nation-building and Estonia's high-tech dreams in the 2010s. *East European Politics and Societies and Cultures, 36*(2), 399–422.

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and Society, 35*(3), 611–623.

Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation and Governance, 16*(1), 197–211.

Bright, J., Ganesh, B., Seidelin, C., & Vogl, T. M. (2019). Data science for local government. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3370217

Bullock, J. B. (2019). Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration, 49*(7), 751–761.

Busch, P. A. (2017). The role of contextual factors in the influence of ICT on street-level discretion. *Proceedings of the 50th Hawaii international conference on system sciences*. Retrieved 15.11.2023.https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/63aa4375-3018-4c7c-8403-384bf3d2a186/content

Busch, P. A. (2019). *Digital discretion acceptance and impact in street-level bureaucracy*. Doctoral Thesis. https://doi.org/10.13140/RG.2.2.28111.53924.

Busch, P. A. & Eikebrokk, T. (2019). Digitizing discretionary practices in public service provision: An empirical study of public service workers' attitudes. *Proceedings of the 52nd Hawaii international conference on system sciences*. Retrieved 15.11.2023. https://scholarspace.manoa.hawaii.edu/items/5a522798-c3c2-4c49-9d56-c1f2fdc4ac62

Carney, T. (2020). Artificial intelligence in welfare: Striking the vulnerability balance? *Monash University Law Review, 46*(2), 23–51.

Considine, M., McGann, M., Ball, S., & Nguyen, P. (2022). Can robots understand welfare? Exploring machine bureaucracies in welfare-to-work. *Journal of Social Policy, 51*(3), 519–534.

Desiere, S., & Struyven, L. (2021). Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy, 50*(2), 367–385.

Döring, M. (2021). How to bureaucracy: A concept of citizens' administrative literacy. *Administration and Society, 53*(8), 1155–1177.

EUIF. (n.d.). *Webpage of Estonian unemployment insurance fund*. Retrieved 15.11.2023. https://www.tootukassa.ee/en/footer/about-tootukassa

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Addison-Wesley.

Giest, S., & Raaphorst, N. (2018). Unraveling the hindering factors of digital public service delivery at street-level: The case of electronic health records. *Policy Design and Practice, 1*(2), 141–154.

Gofen, A. (2014). Mind the gap: Dimensions and influence of street-level divergence. *Journal of Public Administration Research and Theory, 24*(2), 473–493.

Grönlund, Å., Hatakka, M., & Ask, A. (2007). Inclusion in the e-service society-investigating administrative literacy requirements for using e-services. In M. A. Wimmer, J. Scholland, & Å. Grönlund (Eds.), *Electronic Government. EGOV 2007. Lecture Notes in Computer Science, 4656* (pp. 216–227). Springer.

Hajkowicz, S. A., Karimi, S., Wark, T., Chen, C., Evans, M., Rens, N., Dawson, D., Charlton, A., Brennan, T., Moffatt, C., Srikumar, S., & Tong, K. J. (2019). *Artificial intelligence: Solving problems, growing the economy and improving our quality of life*. CSIRO Data61, Australia.

Kaun, A., Lomborg, S., Pentzold, C., Allhutter, D., & Sztandar-Sztanderska, K. (2023). Cross-currents: Welfare. *Media, Culture and Society, 45*(4), 877–883. https://doi.org/10.1177/01634437231154777

Kraavi, T. (2023). Estonian unemployment insurance fund and data science projects. *Conference presentation "artificial Intelligence - Shift from Strategy to execution"*. TAIEX Expert Mission on Artificial Intelligence Institutional Capacity Building, 20–22 August 2023.

Lipsky, M. (1980/2010). *Street-level bureaucracy: Dilemmas of the individual in public services*. Sage.

Masso, A., Kaun, A., & Van Noordt, C. (2023). Basic values in artificial intelligence: Comparative factor analysis in Estonia, Germany, and Sweden. *AI & Society*, 1–16.

Mehr, H., Ash, H., & Fellow, D. (2017). *Artificial intelligence for citizen services and government*. In Ash Center for Democratic Governance and Innovation. Harvard Kennedy School, no. August. Ash Center, Harvard Kennedy School.

Mikita, V., & Kerge, R. (2017). Eesti märk peaks olema 'Igav liiv ja tühi väli'. *Eesti Loodus, 3*, 46.

Misuraca, G., & Noordt, C. (2020). *AI watch, artificial intelligence in public services: Overview of the use and impact of AI in public services in the EU*. European Commission, Joint Research Council, Publications Office. Retrieved 15.11.2023. https://data.europa.eu/doi/10.2760/039619

OECD. (2022). Harnessing digitalisation in public employment services to connect people with jobs. *Policy brief*. https://www.oecd.org/els/emp/Harnessing_digitalisation_in_Public_Employment_Services_to_connect_people_with_jobs.pdf

Peeters, R. (2020). The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making. *Information Polity, 25*(4), 507–522.

Penz, O., Sauer, B., Gaitsch, M., Hofbauer, J., & Glinsner, B. (2017). Post-bureaucratic encounters: Affective labour in public employment services. *Critical Social Policy, 37*, 540–561. https://doi.org/10.1177/0261018316681286

Raudla, M. (2020 February 20). *Töötukassa automatiseeritud infosüsteem otsustab raha jagamise sekunditega*. Virumaa Teataja/Postimees https://virumaateataja.postimees.ee/6904529/tootukassa-automatiseeritud-infosusteem-otsustab-raha-jagamise-sekunditega.

Riigi Teataja. (2023). *Labour market services and benefits act*. Retrieved 15.11.2023. https://www.riigiteataja.ee/en/eli/ee/506062014001/consolide/current

Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). "Just like I thought": Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review, 83*(2), 263–278.

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data and Policy, 2*, E16–E278. https://doi.org/10.1017/dap.2020.19

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly, 36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Veale, M., & Brass, I. (2019). Administration by algorithm?: Public management meets public sector machine learning. In M. Veale & I. Brass (Eds.), *Algorithmic regulation* (pp. 121–149). Oxford University Press. https://doi.org/10.1093/oso/9780198838494.003.0006

Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

Wihlborg, E., Larsson, H., & Hedström, K. (2016). "The computer says no!" - A case study on automated decision-making in public authorities. *49th Hawaii international conference on system sciences (HICSS)* (pp. 2903-2912). IEEE.

World Value Survey. (n.d.). https://www.worldvaluessurvey.org/photos/EV000190.JPG

Young, M., Bullock, J., & Lecy, J. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance, 2*(4), 301–313.

Zilka, M., Sargeant, H., & Weller, A. (2022). *Transparency, governance and regulation of algorithmic tools deployed in the criminal justice system: A UK case study*. doi:https://doi.org/10.1145/3514094.3534200.

Zouridis, S., Van Eck, M., & Bovens, M. (2020). Automated discretion. *Discretion and the quest for controlled freedom* (pp. 313–329).

Chapter 6
AI Use in the Asylum Procedure in Germany: Exploring Perspectives with Refugees and Supporters on Assessment Criteria and Beyond

Elisabeth Späth

Abstract This chapter takes the AI-based Dialect/Language recognition software ("Language Biometrics Assistance System," acronym: DIAS), used by the Federal Office for Migration and Refugees (German acronym: BAMF), as an example for assessing asylum seekers' "eligibility." This software should support decision-makers in identifying the country and/or region the refugee is coming from, based on their language features, as more than half of those who apply for asylum do not have their passport (anymore) or other supporting documents. Based on document analysis, the chapter presents the most important stages of the asylum procedure, its AI component, and the political, legal, and technical context. Empirical research conducted based on interactive sessions, such as a world café, was enriched by former exploratory interviews with important stakeholders supporting refugees as well as desktop research to present current discourses. The experienced assessment criteria in the asylum procedure and beyond, highlighting the experiences of those affected by these assessments, namely refugees, and of those "guiding" refugees through the different procedures, are illustrated and analyzed in this chapter. Furthermore, these insights are discussed in the light of the (current) use of AI for assessment, exploring its implications for fairness through the lens of legitimacy (of asylum bureaucracy) and agency of refugees.

Introduction

Globally, and particularly in Europe, there is increasing interest in using artificial intelligence (AI) in the context of "migration management," which is critically evaluated by a growing number of researchers (Amelung & Galis, 2023; Beduschi,

E. Späth (✉)
TISSS Lab, Institute of Sociology, Johannes Gutenberg University Mainz, Mainz, Germany
e-mail: espaeth@uni-mainz.de

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*,
Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_6

2021; Molnar, 2020), highlighting the potential risks and human rights violation toward refugees (Ahmad, 2021). While this literature mostly focuses on AI systems used at national borders, there is comparatively less research on using AI *within* national borders (e.g., Bither & Ziebarth, 2022; Ozkul, 2023). Related to this, there is a lack of (general) questions relevant for the welfare state, i.e., who is eligible to which social services, and the use of technology assisting public administration to implement these decisions. Refugees represent a specific group since they do not have citizenship in the respective state in the first place and have different legal rights. Furthermore, importantly, there are different types of "protection categories," consisting of a certain of criteria, an asylum seeker is entitled to (or not). The assessment, therefore, is crucial as it is very "decisive" for their individual future, whether asylum is granted in the first place and which "protection category" is given, i.e., possibilities for integration and also access to social service provision.

In this chapter, the use of the AI-based Dialect/Language recognition software ("Language Biometrics Assistance System," DIAS), being part of the "Integrated Identity Management Tools" (IDM-S-Tools) by the Federal Office for Migration and Refugees (in German: Bundesamt für Migration und Flüchtlinge; acronym: BAMF), is taken as an example for assessing asylum seekers' "eligibility" during the asylum procedure. In the context of asylum procedure, this tool represents an important part of identity verification, as more than half of those who apply for asylum do not have their passport (anymore) or other supporting documents; this software should support decision-makers identifying the country and/or region the person is coming from based on their language features. In the following, the most important stages of the asylum procedure, its AI-component, and the different protection status will be briefly explained. Furthermore, the political, legal, and technical aspects of this technology are presented, mainly on the basis of relevant documents made accessible by the BAMF. Additionally, exploratory interviews with important stakeholders working and supporting refugees in the asylum procedure and beyond as well as literature research were conducted to present state-of-the art discourses. These insights provided the basis for identifying important research gaps, leading to an empirical research design focusing on exploring perspectives of refugees and supporters on assessment criteria within the asylum procedure and beyond (integration processes). Thereby, the chapter aims at illustrating, on the one hand, the perceived assessment criteria in the asylum procedure and beyond, highlighting the experiences of those affected by these assessments and of those "guiding" refugees through the different processes. These important insights shedding light on current assessment criteria, and practices, were collected via interactive sessions. On the other hand, an evaluation, and some reflections, based on document analysis, will be provided to validate the insights gained through data collection, to understand some relevant discrepancies and the normative dimension of "assessment." This analysis of empirical data will be discussed in the light of the (current) use of AI for assessment, exploring its implications for fairness through the lens of legitimacy (of asylum bureaucracy) and agency of refugees.

The Assessment and Its AI Component in the Asylum Procedure

First, the asylum procedure in Germany will be briefly explained. The main source for the description is the Handbook Germany.[1] However, all information was also cross-checked with the description of the "asylum procedure" provided by the BAMF, which is—under the leadership of the Federal Ministry of the Interior—the responsible agency for asylum applications. Arriving in Germany, asylum seekers need to inform an authorities' office (e.g., the police, foreigners' office) or an initial reception center. The registration takes place in an initial reception center, where personal information (name, date of birth, country of origin) is collected and a photo is recorded as well as fingerprints. The asylum seeker receives a proof of arrival ("*Ankunftsnachweis*"). Whether the initial reception center is "competent" ("zuständig") until the official start of the "asylum procedure," i.e., asylum application at the BAMF, depends partly on the country of origin of the respective asylum seeker; this is because not all the different nationalities are represented in the reception centers and not all branch offices process all countries of origin (Grote, 2021). In case the initial reception center is not responsible, the so-called "First Distribution of Asylum Seekers" is performed, based on the computer program EASY ("Erstverteilung der Asylbegehrenden"), which is defined by the *Königsteiner Schlüssel*.[2] This determines whether the asylum seeker can stay in this initial reception center or will be sent to another city. The respective competent reception center is thereby responsible for the accommodation, access to food, and a more comprehensive registration. At the competent reception center, the BAMF "already provides support in this procedural step," i.e., the employees of the registration office for asylum procedures ("Asylverfahrenssekretariat," acronym: AVS), by the means of the IDM-S-Tools in case the identity or country of origin of an asylum seeker cannot be determined or an obviously faked document has been given to the authorities (Grote, 2021). These tools consist of (a) automatic face recognition, (b) name transliteration, (c) automatic dialect and language recognition, and (d) the analysis of data devices (Bundesamt für Migration und Flüchtlinge, 2018).[3] The software called "Language Biometrics Assistance System" (or "DIAS") is used since 2017 for Arabic-speaking asylum applicants, and since 2022, also Farsi, Dari-Persian, and Pashtu (Deutscher Bundestag, 2022; Lulamae, 2022). In this step, asylum applicants are asked to describe a picture in their language by phone. In an official document (cf. Deutscher

[1] Handbook Germany is an information platform for refugees and immigrants in Germany. The project has been initiated by the Neue deutsche Medienmacher*innen association at the end of 2016. The website is supported/funded by the Federal Government Commissioner for Migration, Refugees and Integration as well as Ministry of the Interior (HandbookGermany Website, 2024).

[2] The so-called "Königstein Key" determines how many asylum seekers a Federal State (Land) must take in. This is based on tax revenue (2/3 share in the assessment) and population size (1/3 share in the assessment); the quota is recalculated annually (Grote, 2021).

[3] The "practice of regularly analysing" mobile devices, however, has been just ruled by the highest Court in Germany the German Federal Administrative Court to be illegal (Palmiotto & Ozkul, 2023).

Bundestag, 2022), it is stated that applicants are informed about the procedure and the purpose of the language sample before the language test is taken. This includes the information that the result of the language test (report) will be kept on file (ibid.).

The next step is the asylum application ("*Asylantrag*") at the BAMF branch, which represents the official start of the asylum procedure. The asylum seeker is asked for personal data again (country of origin, city, family, school or job, religion, and escape route); a photo as well as fingerprints are taken. The asylum seeker is normally supported by an interpreter. Afterwards, the BAMF checks (cf. "*Dublin Procedure*") whether the asylum procedure can take place in Germany or whether the person should apply for asylum in a different country (e.g., in case the person already has given their fingerprints in a different country, or they already applied for asylum in a different country). Under certain conditions, Germany is able to send asylum seekers back (cf. "*Dublin-Verordnung III*"). After verification whether Germany is the responsible country for the asylum procedure, an invitation to the hearing ("*Anhörung*") is sent to the asylum applicant. The hearing represents the most important step in the asylum procedure and takes place in a BAMF office, where an interpreter is present, too; the asylum applicant is asked by a BAMF staff member about their life in their country of origin, the reasons for their escape, and the escape route; there are varying numbers in regard of the amount of questions (according to the Handbook, there are around 40 questions). The BAMF employee is taking notes on everything reported and is filing a protocol, which will serve as the basis for decision-making (whether asylum is granted and which protection category is given). In the hearing, it is the second time that DIAS might be used (cf. Grote, 2021). According to the official document of the Bundestag, in case information given by the applicant is contradictory or if there is a need for clarification, the hearing is the stage where any doubts should be dispelled. In case the identity or country of origin of the applicant is still uncertain, it is possible to organize a separate language expertise ("Sprachgutachten").

After the hearing, it can take months until the asylum applicants receive the decision ("*Entscheidung*") of the BAMF. There are different types of "protection categories" for refugees; however, only three of four are evaluated as "positive" decisions. These are the (a) acknowledgment of entitlement to asylum, (b) award of refugee protection, and (c) award of subsidiary protection. Concerning the first category (a), asylum is granted to individuals who are persecuted for political reasons by the state or a state-related organization in their home country and who arrive directly in Germany (Art. 16a Basic Law). In the case of (b), individuals are recognized as refugees in case they are persecuted by non-state-related actors (§ 3 AsylG, this article is based on the Geneva Convention); they have a residence permit for 3 years. Individuals who are not entitled to asylum or a refugee status are granted subsidiary protection (§ 4 par. 1 Asylgesetz (Asylum Act)). Subsidiary protection (c) applies if (a) or (b) is not adequate, i.e., if "there is a threat of serious harm in the country of origin." The residence permit issued under these circumstances is only valid for 1 year and can be extended. The most important differences here compared to asylum and refugee status are that persons under subsidiary protection do not receive a permit for traveling purposes and that reunion with family members

is more restricted and entails different procedures (Levy, 2020). Another "protection category" is the imposition of a ban on deportation (§ 60a AufenthG, "Duldung"), implying that deportation to their home country might lead to human rights abuses, danger to life, health, or freedom. This decision is "negative" because the tolerated status is not equated with a residence permit ("*Aufenthaltstitel*") and is only valid for a certain period, for 1 year or less. There are two ways in which the asylum application can be rejected: (1) the BAMF does not see any reason why protection is necessary or (2) the BAMF supposes that the applicant did not tell the whole truth during the hearing (and that the person migrated to Germany mainly for economic reasons ("*offensichtlich unbegründet*"). The decision of the BAMF cannot be challenged ("angefochten") in the first application. It can be only challenged by filing a new asylum application (second and follow-up applications), stating that there has been a change to the factual or legal situation (Federal Office for Migration and Refugees, 2024). Importantly, under certain conditions (integration measures), the legal "status" can change (see next section).

The Assessment Beyond the Assessment in the Asylum Procedure

The previous section highlighted the most relevant stages—and "categorizations" in the asylum procedure, as most important assessment. However, to understand the implications of this assessment, it is important to consider the stages *beyond* the assessment on their status. As pinpointed by Raschke (2023), national migration policy and integration policy are completely interrelated, since the better the "perspective to stay," the "easier" it will be to integrate and become less dependent on service provision. In this way, assessment practices having an impact on the residence permit do not end at the BAMF's doorstep but are continued at those places responsible for integration "on the ground" (cf. Fig. 6.1). In the case of having a "tolerated status"; for example, doing an apprenticeship can legally increase the chance "to switch from a tolerated status to a residence permit" (Bauer & Schreyer, 2019). However, this involves many steps in between and is only possible if the person is living in Germany for several years, has a job, and makes full use of "integration measures" (e.g., reaching a certain level of German).

The most relevant actors involved in the "integration process"[4] regarding refugees on the federal government level are the Ministry of the Interior and Community, closely cooperating with the BAMF, as well as the Ministry of Labour and Social Affairs, closely cooperating with the Federal Employment Agency. At the municipality (local) level, the most important actors are the following: Federal Office of Migration & Refugees Branch, foreigners' offices, social welfare offices, housing

[4] A detailed description of these, such as access to language courses, housing, work, etc., is outside the scope of this chapter; it should be emphasized that many different actors are involved in them.

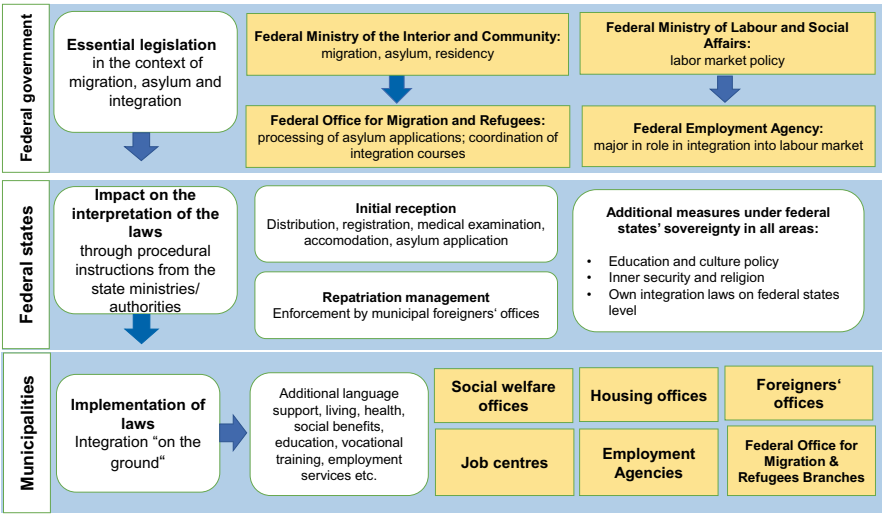

Fig. 6.1 "The most relevant state actors and tasks for the integration of refugees," own translation; Source: See Original: Sixtus, F., Kiziak, T., & Klingholz, R. (2019). *Von individuellen und institutionellen Hürden. Der lange Weg zur Arbeitsmarktintegration Geflüchteter*. Diskussionspapier 23, p. 11. Berlin-Institut für Bevölkerung und Entwicklung

offices, and employment agencies, which again work closely together with job centers.[5] Although "integration" already takes place *during* the "asylum procedure," it is getting more complex after the "decision" of the BAMF: this is mainly due to the fact that there exists a "*path dependency*" related to the legal status of the person (e.g., subsidiary protection or deportation ban), and also because it happens on the *local level* (see Fig. 6.1, integration "on the ground").

These policy processes do not happen in a vacuum, but in the dynamic field of actors interacting with each other: considering the different actors and resorts, Schammann (2015) and others (Dahmen et al., 2017; Muy, 2020) indicate that there is a tension between "a *regulatory approach* (migration control) and a *welfare state approach* (labor market integration and integration and securing livelihoods)" (Schammann, 2015). This tension is transferred or maintained on the federal level (Federal Ministry of the Interior vs. Federal Ministry of Labor and Social Affairs) and municipality level (foreigners' office vs. welfare office). These local, contextual factors are important for understanding assessment practices and its outcomes: since the past 5 years, the federal structure and the margin of discretion are depicted in (empirical) studies, too, where differences in decision-making are highlighted. This does not only refer to the high variance among the German Länder concerning the recognition rates (e.g., Riedel & Schneider, 2017), but also to the different

[5] Foreigners' offices play an outstanding role (cf. Eule, 2017) because they are responsible for executing the issuance and/or the refusal of a residence permit, and also, are responsible for clarifying the identity of refuges (including obtaining identity papers), if the identity is still not clear.

implementation/application of social service provision on the municipality level (Meyer et al., 2021; Rambøll Management Consulting GmbH and Nationaler Normenkontrollrat, 2014; Schammann, 2015). Some authors even label it as "asylum lottery" (Schneider & Riedel, 2017).[6] These contextual factors will be reconsidered in section "Discussion".

Political and Legal Context of Introduction of IDM-S Tools

To understand the *rationale* of the BAMF to use AI in the asylum procedure in the first place, its context, policies, and institutional infrastructures supporting these practices are considered, mainly by the means of document analysis (Bowen, 2009): internal documents as well as three relevant Working Papers published by the BAMF (Grote, 2018; Grote, 2021; Tangermann, 2017) as well as the BAMF's Digitalization Agenda 2022. Apart from that, important changes of legal regulations will be considered, too.

Political and Legal Basis: From Efficiency, (Data) Quality, (Inner) Security to Fairness

The plan to use IDM-S tools can be traced back to 2015. The high number of asylum applications, predominantly due to war in several countries, e.g. in Syria, led to the fact that the BAMF was overburdened with these, to a considerable extent due to a lack of an adequate number of (skilled) personnel. As mentioned before, the "establishment of identity" (Tangermann, 2017) is a crucial step. While the asylum seeker needs to "credibly substantiate their history of persecution", the respective BAMF employee needs to "balance" the current information on the country of origin and the legal basis of the Asylum Act (ibid.). According to Tangermann's explanation, this requires "specific knowledge, experience, and intuition" (ibid.). Back in 2015/2016, the main challenge was to implement training as well as instruction for many new employees, especially those who conduct the "hearing" (ibid.). However, the difficulty of "establishing the identity" of a person stays *omni-present* in the different stages of the asylum procedures (see section "The Assessment and Its AI Component in the Asylum Procedure"; registration; personal interview; issuance of residence permits), if asylum seekers cannot show any supporting document. Next to the issue of the lack of personnel, on the other hand, there were also organizational and technical issues. Until 2015, the registration of asylum seekers was

[6] Schammann and Gluns (2021), for example, have identified some of the causes for these kinds of variations at the local level, related to civil engagement, financial situation of the community or district, and also the political willingness of decision-makers at the local level. Generally, there seem to be differing views, however, on the extent to which these differences are linked to the governing political parties.

handled differently in the Länder; this implied that a "central registration" (e.g., identity check, including biometric data such as fingerprints) did not take place until the application was submitted to the BAMF (IT-Planungsrat, 2018). There was no "data pool" for basic data or standardized classification features, which led to the fact that "personal data was often collected redundantly" and was erroneous (partly due to the difficulty of spelling names) (ibid.). The delay of processing asylum applications, as pointed out by Grote, caused confusion and uncertainties, among authorities as well as asylum seekers: "[…] *lengthy and at times confusing (data processing) procedures at the individual authorities involved in the asylum procedure were identified as neuralgic points, which led […] to multiple registrations as well as status, responsibility and benefit uncertainties for the asylum seekers and authorities involved*" (Grote, 2021). To tackle these issues—mainly related to efficiency and data quality, the BMI initiated a "project group on the digitalization of asylum procedures" (PG DAS, "Projektgruppe Digitalisierung des Asylverfahrens"), while the IT Planning Council, who coordinates the collaboration between the federal state and the Länder in respect of information technology, was asked to adopt the role of the coordinator (IT Planning Council and BMI 2018). While these efforts became effective in May 2016, introducing the IDM-Tools in 2016, further political and technical developments need to be considered in this context; there were a few cases where individuals applied for asylum with different types of identities (Hahn, 2020); next to the case of Amnis Amri (having ten different identities), who killed several people in Berlin, there was the quite bizarre case of Franco, a far-right German soldier, who could successfully register as an asylum seeker from Syria causing considerable criticism leveled at the BAMF (ibid.). In a summarizing statement in 2018, the IT Planning Council and BMI concluded: "*This situation led to inefficient processes, a lack of data quality, a lack of political control, duplicate identities with the possibility of misuse of services and identity concealment, and ultimately to considerable security risks*" (ibid.).[7] In this way, efficiency, the quality of data, and security risks were strongly interrelated and provided an argumentative basis:

- **Data quality**: to improve data quality, i.e., data on the person applying for asylum, stock data ("Bestandsdaten"), and new data to be recorded
- **Plausibility**: identification of connecting points of the information given by the applicant and/or the received information (perspective of those who process these data)
- **Safety-related aspects**: limitation of possibilities concerning the international concealment of one's identity by the applicant. Provision of a better and more comprehensive information status ("Informationsbild") for the (BAMF) employee (own translation; BAMF 2018)

[7] This is why many involved actors as well as researchers, such as Hahlen and Kühn (2016), pointed to the vulnerability of the system and suggested to speak of an "administration crisis" rather than a "refugee crisis."

Between 2014 and 2017 only, there were 19 different types of laws and regulations related to asylum politics implemented (cf. Grote, 2018, cf. Fig. 3). One of these refers particularly to the use of a language analysis, *§ 16 Sicherung, Feststellung und Überprüfung der Identität (§ 16 Abs. 1 Satz 3 und 4 Asylum Act): "An analysis of the language of the applicant to determine his or her probable country or region of origin is a measure foreseen by the Asylum Act, provided that the applicant is informed of this beforehand (Section 16 subs. 1 third and fourth sentences of the Asylum Act."* In effect, a language analysis is not mandatory but may be used if there are "reasonable doubts about the identity of the applicant" (Tangermann, 2017). While there generally are not many publications of the BAMF reporting on or explaining their AI-based tools, the document "Digitalization Agenda 2022" (Federal Office for Migration and Refugees, 2021a, b), being part of its public relations work (p. 30), is more explicit on technology that is already in use and/or being developed. In contrast to previous documents, fairness is used as an argument for using AI: "*In the asylum area, assistance systems based on artificial intelligence and digital identity management, which implements the best possible and fair processing of asylum applications for all parties involved, will be of great importance in the future"* (p. 24). While the values of efficiency, (data) quality, and (inner) security are more or less explained in the different documents, the value of fairness is not further explained or conceptualized in more detail in these.

Technical Information on DIAS, Including Statistical Data on Its Use

The following technical information on DIAS as well as some related statistical data on its use is based on an internal document ("Sprachbiometrisches Assistenzsystem—Unterstützung der Feststellung von Herkunft und Identität im Asylverfahren") as well as a request by the left-wing party (Die Linke) to the Federal Government (Bundesregierung). According to the answer of the Federal Government (2022), DIAS is based on the software Nuance Speech Suite; the training data is retrieved from the Linguistic Data Consortium (LDC) at the University of Pennsylvania as well as a small amount was procured via Clickworker GmbH. The Arabic models were trained on anonymized language samples of the BAMF. According to the internal document, the process of introducing the DIAS was structured along three "stages": a proof-of-concept phase (2–4 weeks), a pilot phase (3–6 months), and a transitional phase (permanently) in which the product is transferred into "sustainable structures and processes."

One question in the "request document" addresses the cases in which the data ("Angaben") of the asylum seeker could be affirmed or refuted. It is stated that between the years 2020 and 2022, the verification of an asylum applicant's identity without any documents could be "supported" ("gestützt") in 76–79% of cases, and in 21–24% of cases "not supported" ("nicht gestützt") with the DIAS. Another important question raised concerns the "recognition rate" ("Erkennungsquote") of the DIAS. The answer: the "recognition rate" of the Arabic dialects was about 80% in the time period 2017 until 2020. After first-time implementation of language

model training in 2021, the recognition rate is up to 85%. The recently introduced Persian language models (Dari and Persian/Farsi) reach a recognition rate of 73.07% and Pashto a recognition rate of 77.7%. It is assumed that the recognition rate will be increased due to trained language models. Between 2020 and 2022, Syria, Algeria, Iraq, and Morocco were the countries of origin mostly detected. The document, furthermore, contains information on the number of language samples, its data protection and recording mechanism, and statistical information on how many times DIAS was used in different BAMF branches.

Exploratory Interviews Regarding Changed Practices of Assessment and the Role of AI

To explore, and understand, relevant aspects of assessment practices as well as AI usage in this context, five exploratory interviews with welfare organizations, representatives, and refugee commissioners from Christian churches were held in the summer of 2021.[8] The topics of these interviews centered around their perspectives on changed practices of assessment and distribution of social service provision in the past 5–10 years in reference to the dynamics of political, societal, and technological development and, secondly, their perspectives on the status quo of digitalization and the (possible) use of artificial intelligence. In respect to both questions, there was a considerable overlap of experiences, perspectives, and mentioned problems and challenges, while the latter were generally very much emphasized. In respect of the first question, one mentioned positive aspect was the fact that since 2015/2016 structures for counseling and social support have very much improved. However, there was a general line of agreement that "not much has happened or changed" in the past 10 years, particularly referring to the "intercultural opening" in the public administration ("Interkulturelle Öffnung in der Verwaltung"), and training of intercultural competence of administration staff. Related to this, a paradox was emphasized by two stakeholders: while Germany is striving to become a more "diverse society," creating initiatives to attract skilled workers from abroad and implementing integration measures for refugees, people "from abroad," refugees in particular, are confronted with big bureaucratic hurdles (e.g., getting a Visa and/or work permit), and also face racist behavior in institutions. Furthermore, large "fluctuations" referring to changing emotions and opinion of people living in Germany toward migration politics and refugees can be observed: while there was a "welcome culture" in 2015/2016 and general appreciation and acceptance of migration politics, resentment against so-called "foreigners" has grown in the last couple of years, resulting in considerable intake for the right-wing party "Alternative

[8] While the exploratory interviews took place in August 2021, it is important to "contextualize" the empirical data taking place in July 2022; due to the war in the Ukraine starting in 2022, discourses changed, having an impact also on research with refugees and supporters.

for Germany" ("Alternative für Deutschland"; acronym: AFD). Apart from that, the margin of discretion in public administration generally poses a challenge for fair decision-making. The federal structure was addressed, too, by all stakeholders: the heterogeneity of the Länder concerning their financial situation and resources represents an important factor for successfully organizing and implementing supporting structures for social service provision and integration measures.

Against this background, according to all interviewed stakeholders, artificial intelligence does not play a role at all in their daily work; to the contrary, the lack of digitized processes within the administration and services provided to recipients was highlighted by most interviewees; some stakeholders, like the job center and the foreigners' office, still require using fax because it is supposed to be "safest" means in relation to data protection. They stressed, too, the high relevance—and the difficulty—of exchanging relevant information among authorities because of, on the one hand, the incompatibility of databanks, and on the other hand, the sensitive data of refugees. In respect of AI use in the asylum procedure, two main concerns were raised: One touches upon the "complexity of knowledge" needed to perform the assessment (e.g., intercultural competence) and to assess the credibility of a refugee's story (background, escape route, etc.) in a setting of time constraints. The other aspect mentioned refers to "quality assurance" of the data and the related processes (e.g., who determines what data are used as input) and the question of "monitoring structures" (e.g., more-eye principle to ensure equity right from the start). The lack of knowledge, or awareness, that AI is increasingly playing a role in asylum procedures is reflected in the literature, too. Regarding the German context, there are only a few organizations or NGOs like AlgorithmWatch and blogs like Netzpolitik.de (via Frag-den-Staat.de), who regularly report on the BAMF's usage of DIAS (Biselli, 2017; Lulamae, 2022).

The insights gained based on document analysis, literature research, and exploratory interviews helped to identify some important research gaps and define a research focus: first, as for now, the question whether asylum applications are processed "in the best possible and fairest way for all parties" with the help of AI—or to put it in other words—whether a fair assessment is really taking place—is so far only answered by two stakeholder groups: BAMF/BMI and technology providers. The example of the DIAS indicates a *lack of (public) discourse*; apart from that, the case of the mobile data processing illustrates that law is lagging behind technological development. Secondly, there seems to be a *discrepancy in respect to the use of technology in the field*: on the one hand, there are several initiatives taken by the BAMF to implement digital workflows and high technology, including AI; on the other hand, the lack of digitalized services, for example, in foreigners' offices, does not support making the asylum procedure and integration (e.g., access to social service provision, work permission) more efficient and "good." Furthermore, this is exacerbated by the *margin of discretion* and the *federal structure*, marked by the "tension" of different actors involved (cf. Schammann, 2015). All these layers matter in evaluating whether assessment(s) contribute to (more) fairness.

Empirical Research Focus and Methods

In the following, the empirical research therefore centers around those who experienced "being assessed" by the BAMF and also around those who support refugees and asylum seekers in the different processes of the asylum procedure—and beyond. As previously indicated, asylum procedure and integration are interrelated as the specific legal category has implications for access to social service provision, education, work, etc. The empirical research design therefore deals with the *experienced assessment criteria, relevant in asylum procedure and integration procedure*, and furthermore, partly due to recent political developments in Germany, and the world, *other aspects that are relevant to them in the context of "being assessed."*

The empirical material collected during a 1.5-day workshop[9] with refugees and supporters in July 2022 will be presented in the following section. Before, the stakeholders participating in the workshop should be shortly presented. There were ten refugees (R) from three different countries: two refugees from Iran, four refugees from Afghanistan, and four refugees from Syria; six of them were male, four of them were female, and the age range was ca. 25 years and 70 years. There were 8 supporters (S); two of them were female, six were male, and their age ranged from ca. 45 years to 80 years. The majority of supporters were voluntary helpers. Some of them are active in civil society organizations, while others work in welfare organizations. The recruitment was based on a "snowball sampling" via supporters, mainly for two reasons. First, due to the potentially sensitive information regarding refugees' experiences, it was crucial to ensure a certain level of trust.[10] Secondly, related to the former aspect, there were two main "criteria" for the selection of refugee participants, considering their "status" and the number of years they already live in Germany. In respect of the research questions posed in the previous paragraph, it was necessary that (at least) the vast majority has "gone through" the asylum procedure and integration measures, such as language courses. This was crucial, too, to have German as main language during the workshop. Most refugees taking part in the workshop spoke German almost fluently; three of them preferred to speak in English; two of them needed additional help for translation. A last remark refers to the division into "refugees" and "supporters"; importantly, three refugees could be also considered as "supporters" because they are helping other refugees and migrants in many ways; however, due to their experiences in the asylum procedure, they will be referred here as refugees.

In the following, the methods used will be shortly introduced, namely *triangulation, research world café,* and *grounded theory*, which provided the basis for coding the world café. To address the first part of the research question, an interactive

[9] The author wants to thank David Wurster, Blanca Luque Capellas, and Andrew Chan for their support in data collection.

[10] Beforehand, there have been clarifying conversations on the content and methods employed in the workshop as well as anonymization of data, between the researcher and the supporters, partly with refugees themselves too.

session was created, in which refugees and supporters were asked to explain the criteria they felt assessed by and which criteria should play a more dominant role in the process and/or which should not play a role at all. This interactive session focused on "pinpointing" criteria on a wall, where everyone was asked to agree or disagree. While the perceived, or experienced, assessment criteria are presented, there will be also a reference to the "actual" criteria, following the method of *data triangulation* to account for their reliability and validity (Denzin, 2006; Flick, 2007). The source of reference here is Grote's Table on "What data is collected at which time by whom, how is it collected and where is it stored" (Grote, 2021).[11] This document is taken as a point of reference, on the one hand, since it gives a comprehensive picture of which kind of information is relevant to which kind of state actors. The downside, on the other hand, is that there is no prioritization of "criteria" shown, which might be compared with the perceived assessment criteria. While these data may not represent criteria per se, they indicate (the amount of) relevant information related to identity verification, reflected in the list of categories and sub-categories. These will be shortly listed here (Grote, 2021, see Table 6.1):

A second source for exploring relevant assessment criteria is based on the 40 investigative questions during the hearing, which were compiled by the Handbook Germany. The aim of this overview is to provide potential asylum applicants/refugees with information, stating that it is very important "to be well-prepared" for the hearing (Handbook Germany 2024). The advantage of this source is that it is very detailed—in some regards more detailed than the previously mentioned table—so there are criteria which are not mentioned in Grote's table, such as the indication to military service or membership of a specific tribe or ethnicity (which is relevant, too, in the case of "perceived" assessment criteria, cf. 4.1). At the same time, these questions aim at preparing the asylum seeker/refugee for the potential private and sensitive questions that might be asked by the BAMF, which is most obvious in the case of questions related to family members/conditions (about 10 questions). Due to this "bias," it is hard to draw conclusion on what matters to the BAMF most. Therefore, this source is rather used to support the conclusion drawn on the comparison "perceived vs. actual" criteria.

To address the second part of the research question, the *research world café* was chosen as a method. The world café is defined by Schiele et al. (2022) as a "method of explorative data collection as part of a qualitative research approach, gathering experts in a workshop, which share their knowledge by rotating between several discussion tables, each are focusing on a particular aspect of the overall topic" (p. 281). Importantly, the method is suitable for *exploration* as well as for *verification of themes* (Löhr et al., 2020). Furthermore, the method is suited to create a

[11] The Working Paper by Grote (2021) is part of Germany's contribution to the European Migration Network Study "Accurate, timely, interoperable? Data management in the asylum procedure": "The study is conducted in all participating EU Member States and Norway according to common specifications. The results of the national study are subsequently incorporated into a comparative synthesis report, which provides a pan-European overview of the measures and challenges with regard to data management in the asylum procedure in the Member States" (ibid.).

Table 6.1 Relevant data (categories) and sub-categories recorded by the BAMF; own illustration; Source: Grote (2021), cf. Table 6 "What data is collected at which time by whom, how is it collected and where is it stored"

Data/Categories	Sub-categories
Name	1. Current name 2. Birth name 3. Previous name 4. Pen name (Alias) 5. Religious name 6. Artist name 7. Monastic name 8. Other names 9. Gender
Biometric data	1. Photo 2. Fingerprints 3. Eye color 3. Height 4. Date of birth 5. Citizenship(s) 6. Country of origin
Place of birth	1. Hometown 2. Region 3. Country 4. Date of arrival in Germany 5. Last place of residence in the country of origin
Contact details	1. Phone number 2. Email address 3. Current address 4. Responsible reception center 5. Competent immigration authority 6. Responsible land 7. Civil status
Education	1. School attendance 2. Academic studies 3. Apprenticeships 4. Occupation 5. Language skills 6. Profession
Supporting documents	1. Passport 2. Travel document 3. Passport substitute 4. Reasons for fleeing 5. Reasons for not wanting to be returned to the competent Member State (− > Dublin procedure) 6. Previous applications 7. Information on the route taken 8. Religious affiliation (voluntary indication)
Vulnerability/ Health	1. Unaccompanied minor 2. Pregnant 3. Disabilities 4. Elderly 5. Single parent with minor child(ren) 6. Victims of human trafficking 7. Mental disorders 8. Victims of torture, physical or sexual violence (*health-related criteria are considered by the medical service)
Accompanied by	1. Spouse or civil partner 2. Children 3. Parents 4. Other relatives
Family members	1. Name 2. Residency 3. Citizenship

space for facilitating "dialogue and mutual learning" (ibid.). This approach combines several advantages (cf. Schiele et al., 2022, Table 3. Method comparison). Next to generating and refining ideas and topics, one overall objective of the research world café is "testing of knowledge," which is, here, relevant to the affected people and the researchers. In practice, this overall objective was realized by having only one topic fixed and two topics the participants could choose themselves. The "fixed" topic was about ideas dealing with potential(s) of using "algorithms" in the context of asylum procedure and beyond (Table "Algorithm"). During focus groups on the day before, it became clear that especially refugees wanted to discuss more on the issue of apparently different assessment criteria used for Ukrainian refugees (Table "Ukraine").[12]Furthermore, a topic brought up by refugees as well as supporters was possibilities to participate or change processes in society, but also in politics (Table

[12] Since the invasion of Ukraine by Russia in February 2022, it was the first time that the European Union directive Temporary Protection Directive (TPD; Council Directive 2001/55/EC) ("Massenzustromslinie"), introduced in 2001, was applied. The goal of TPD is to provide an immediate, temporary protection for displaced people from outside the external border of the Union. In practice, this directive implies a considerable facilitated access to social service provision, labor market, education, housing, and health services. This is relevant to acknowledge because, since then, (public) discourse changed, impacting empirical research as well.

"Participation Possibilities"). At each table, there was a moderating person of the research team, introducing the topic and supporting the participants to write down their thoughts and ideas on flipcharts (the rotation time was about 15–20 min). After the sessions, the findings were presented to the whole group in a plenary session. The analysis of the transcripts followed a *grounded theory* approach (Strauss & Corbin, 1990): on the one hand, to distinguish more clearly the perspectives of refugees and supporters and also, on the other hand, to identify main themes and related aspects. Especially in the context of qualitative research with participants of a different culture, it is beneficial to emphasize the voices of these participants and possibility that they themselves "give meaning to the data" (Manning, 2017). Therefore, "in vivo coding," using the words used by the participants themselves, was applied whenever suitable and possible (ibid.).

Results I: "Being Assessed"—Experienced Assessment Criteria

The interactive session with refugees and supporters provides some interesting insights (see Table 6.2): regarding (1) the weighing mechanism (e.g., the first six categories are considered to be very important), (2) which criteria should not play a role, and (3) which criteria should play a more prominent role. In the following description, the criteria are not presented in chronological order, but in the context of the desiderata of refugees and supporters and in comparison to the criteria illustrated by Grote (cf. Table 6.1).

The criterion of "**Country of Origin**" was of great importance; there is a considerable overlap with the criteria illustrated by Grote, especially the main categories "Name," "Place of birth," and, "Supporting documents" and high numbers of subcategories; however, according to the participants, aspects like "individual case decision," "war," and different forms of "persecution" should have more relevance. This is mirrored, too, in the fact that "**Political opinion"** and "**Ethnicity**" are not listed in the table by Grote; here, they are mentioned as having "relative" significance in the assessment; however, these two criteria can be sub-ordinated to the level of "individual case decision."

Another interesting criterion is "**Religion**." This category was, more surprising, mentioned as having high relevance in the assessment procedure. Refugees and supporters stated that religion "should not play a role"; in this context, refugees and supporters referred to discriminatory practices; e.g., someone with a Muslim background and/or belief should not be treated differently than someone with a Christian background/belief. Interestingly, there is a considerable discrepancy in comparison to the overview offered by Grote: here, religion ("religious affiliation") is a small sub-category of "Supporting Documents," and, furthermore, there is a "voluntary indication." This suggests, too, a discrepancy between "actual vs. perceived" criteria.

The criterion "**Support**" represents rather an outstanding category because it is the only one which does not reflect an attribute or intrinsic characteristic of the

Table 6.2 Perceived assessment criteria and wishes regarding criteria

Assessment criteria	In sum	Refugees	Supporters	Should play a (more prominent) role	Should not play a role at all
Country of origin	13	6	7		
Religion	12	4	8		Religion
Profession/Job	12	5	7	Profession/Job (Qualification/Training)	
Support[a]	11	4	7		Support
Documents (Pass/ID)	10	3	7	Documents	
Education/ Apprenticeship	10	3	7	Education/ Apprenticeship	
Political opinion	9	3	6		
Health	9	3	6		
Ethnicity	6	2	4		
Family conditions	5	0	5		Family conditions
Personal charisma[a]	5	1	4		Personal charisma
Experience of violence	4	1	3		
Language skills[a]	4	2	2		Language skills
Army affiliation[a]	3	1	2		
Age[a]	3	1	2		Age
Gender	2	1	1		Gender
				Individual case deliberation	
				War	
				Political persecution	
				Religious persecution	
				Family persecution	

[a]Was mentioned first by supporters

respective refugee, but an external condition. Of course, this category is not captured in the table offered by Grote. The participants attached great relevance to this criterion because it "makes a difference" whether a refugee has support before and during the asylum procedure, normally having a positive impact on the outcome. However, importantly, like religion, the criterion should not play a role because it increases the *dependency* of refugees on supporting infrastructures.

"**Profession/Job**" and "**Education/Apprenticeship**" represent two other important criteria, too, which seem to play an important role. This is reflected in the table by Grote, too. While from a legal perspective, this criterion does and should not play a role, like in the case of religion, refugees and supporters refer to its importance; the reason *why* it plays a role (cf. "actual" criteria), e.g., in the hearing, is, however, purely speculative at this stage. Based on the experiences of refugees and supporters, there is a reason why this should play a more important role: in the context of

the asylum procedure, the level of profession and education indicates that the person/asylum applicant had reasons to leave the country of origin, indirectly showing that they did not come to Germany for economic reasons in the first place; however, this is, of course, only valid for those who have a high level of education and professional background.

"**Health**" is relatively relevant in the assessment, according to the participants. This more or less corresponds to the table by Grote; it is a highly individual—and sensitive—criterion and of course might intersect with other criteria (e.g., gender, disability). The same applies to "**Experience of violence**" (which might be associated with "vulnerability" and "health" in Grote) as well as to "**Family members/ conditions,**" which is reflected, too, in Grote's table ("Accompanied by" and "Family members"). The criterion "**Army affiliation**" was mentioned by a smaller number of participants; this criterion is, however, not mentioned at all in Grote's table. The criteria "**Language skills,**" "**Age,**" and "**Gender**" were mentioned as well as, but implied significance to a lesser extent. According to the participants, these criteria should not play a role at all. In respect to the "actual" criteria, there does not seem to be a big discrepancy.

Results II: World Café, Table "Ukraine"—Perceived Assessment Criteria Regarding Ukrainian Refugees

Mainly due to space constraints, there will be only one table/topic of the world café presented here. To contribute to demonstrating what matters to the research participants, refugees, and supporters, and to shed more light on the role, and meaning, of "assessment" in this particular context, some results regarding perceived assessment criteria considering Ukrainian refugees are illustrated.

Throughout the analysis and the coding process of relevant aspects mentioned, a "red threat" has emerged: the overlapping but also differing perspectives of refugees and supporters. Generally, there was considerable consensus on the fact that refugees from Ukraine have ***easier access to the welfare system***. This refers to multiple layers, such as general social service provision, access to visa, and acknowledgment of education/diploma. These aspects were "narrowed down" by the moderator by the term ***prioritization***, especially in bureaucratic processes. This was very much linked to *"**being Ukrainian is enough**" (S3; R1):*

> The government is rich for Ukrainian people than other refugees […] they did for the Ukrainian like Ausweis, Erlaubnis for work, Schule … (R7)

> You get help if you are Ukrainian. You are specially assessed (=bewertet) [….] (S7)

Furthermore, the different assessment and thereby prioritization were also linked to "***European identity***":

> […] they are quasi-Europeans (S4: Yes, they are), although they do not belong to the EU, but they are then identified as Europeans (R5: considered (=betrachtet) […] (S2)

Related to this, refugees and particularly supporters emphasized the ***political will*** as being decisive:

> Yes, and the will was there from the outset (R5: "These are prejudices") to Europeanize the Ukrainians and pull them to the West (S4)

This last aspect mentioned already indicates the slightly different perspectives of supporters. They demonstrated knowledge concerning, on the one hand, special financial benefits for those who support refugees, for example, offering them a space in their flat or house. On the other hand, they pointed to ***difficulties related to easier access to social service provision***:

> It was also a struggle at first. Should they get money from the job centre or should they get money according to the Asylum Seekers Benefits Act [….] (S2)

Apart from that, they linked the arguments mentioned before—easier access to welfare system and prioritization—to political and partly economic reasons: "and she [referring to current Ministry of Exterior] said you should all come, in reality it's quite … Yes, let's say economic reasons" (S4).

In contrast, refugees predominantly reported in which ways Ukrainian refugees are "***more proximal***" to Germans; based on this, a set of criteria emerged, which are illustrated in Fig. 6.2:

Furthermore, there were two quotes referring to racism in the context of "being Ukrainian is enough": "*That is obvious racism in the system*" (R10) and "*latent racism*" (R1). Importantly, refugees differentiated, too, how these "differences" are implemented:

- via law: "*Afghan people, paragraph 25, section 2, Ukrainian paragraph 24*" (R6)
- via politics: "*Yes, this [other judgment] comes from a high level, from politicians or something […] This is worse than when normal people talk about this*" (R4); "*The politicians who say "Ukrainians are smarter or very highly qualified or something" Maybe yes, but such statements are not right[….] in every society there are the good and the bad, that's how it is*" (R4)
- via society: "*public perception plays a role in assessment*" (M1)
- via media: "*Yes, and it's also faster in the media, in social media or something*" (R1); "*Media pressure*" (R1)

Analysis of Empirical Data: Exploring Assessment Criteria

The collected empirical material (sections "Results I: "Being assessed"—Experienced Assessment Criteria" and "Results II: Worldcafé, Table "Ukraine"—Perceived Assessment Criteria Regarding Ukrainian Refugees") will be analyzed in more detail in the following. Regarding this analysis, it is important point out what criteria and processes are applied or implemented regarding refugees (depending on the country of origin). This is important to reflect upon the normative dimension of

Culture:

- *"Ukrainian culture is a, is a criterion for reception [of refugees]" (R1)*
- *"There is a big difference between Ukrainian culture and Iraqi or Syrian culture, but Ukrainian culture and German culture are similar." (R1)*

Easier integration due to culture:

- *"I might come from Iraq or something, and I want to stay or live in a German family, then it's difficult for me. But it's easier for the Ukrainians, right?" (R4)*
- *"Perhaps the Ukrainians integrate more easily than the others." (R1)*
- *"Difference between autocratic countries and democratic countries" (R4, referring to refugees from other countries)*
- *"Culture shock, that sometimes takes up to two years" (R4, referring to refugees from other countries)*

Geography:

- "*But it's also about geopolitics. Because if Ukraine is close to me, then I have to be very careful. And because of this reason, they have such priority." (R4)*

Religion: "*religious proximity*" (R4)

Appearance: "*maybe the look as well*" (R1; R5)

Fig. 6.2 Perceived assessment criteria regarding Ukrainian refugees; being "more proximal"

assessment, gained by the empirical information, i.e., what should (not) play a role in the process and/or where do the participants see discriminatory practices.

The perceived, and actual, differences in assessment criteria—regarding refugees from Syria, Afghanistan, Iran, and Ukraine—have been clearly articulated by refugees and supporters: in the first dataset, this has been illustrated by the discrepancy ("is vs. ought"), referring to "country of origin" and "individual case decision" evaluated as most important criterion. This is interesting since these aspects reflect the legal basis for a "protection status." The participants' perception and evaluation might be an indication that, on a macro level, political category of (un)safe country practically "outweighs" the legal category. The more "individual" motivation/reason for escaping the country of origin, legally crucial, seems to play a secondary role. This is significant for understanding how the criteria are applied, indicating a tension between political objectives or policies and legal obligations.[13] The observation that there is another discrepancy ("is vs. ought") regarding "religion"—as it supposedly plays a role, but it should not according to the participants, stands along with this argument.[14] This stands in contrast to criteria reported to play a role for Ukrainian refugees; to be acknowledged as a refugee *and* having resources/access to various types of social welfare provision is associated with "being Ukrainian,"

[13] This discrepancy can be further underlined by the questions made public by the Handbook; in fact, only 2–3 questions deal with the reasons for fleeing.

[14] However, this evaluation, too, depends on the individual context; for example, in the case of religious persecution, religion as an attribute needs to be taken into account.

"political will," "European identity," but also "being more proximal" (e.g., religion, geography, appearance) to Germans and to German culture.

Another aspect which appears in both empirical datasets is the professional and/or educational status. Regarding the first dataset, profession/education is relevant in the assessment process; regarding the second dataset, political and economic reasons are mentioned as well by participants to explain why Ukrainian refugees are assessed differently. Nonetheless, from a political and legal point of view, this criterion is not and ought not to be relevant regarding whether someone gets a "protection status." In an interview, the Representative of the Federal Government for Migration, Integration and Refugees (at the same time being the Representative for Anti-Racism) clarified that "*the right to asylum is a human right. We must separate labour migration and escape. However, it is important that refugees who have been here for a certain period of time can also find work*" (cf. interview conducted by L. Kottmann in 2023). The last part of this quote addresses an important aspect of integration, namely (having access to) work. Considering this, Ukrainian refugees had, in fact, a "more favourable legal conditions, being exempt from the asylum procedure, employment bans, and were granted immediate residence permits for at least the initial two years" (Brücker et al., 2023), in particular, having more direct access to the labor market (Mediendienst Integration, 2024).

The fact that "support" is a high-ranking "criterion," but should not play a role, reflects this (dependency-)dilemma: while "support" does not represent a criterion for the assessment practice itself, it has a supposedly considerable impact on the processes related to the assessment (within the asylum procedure) and the processes beyond (→ integration). In sum, the assessment criteria and practices applied in the case of refugees from countries like Syria and Afghanistan, and those from Ukraine, in fact do make differences among people, which very much corresponds to the idea of "othering": "*the act of treating someone as though they are not part of a group and are different in some way*" (Cambridge Dictionary, 2024). Importantly, "othering" seems to stand in direct opposition to a "sense of belonging" and the desire to create "equality among refugees," which was explicitly mentioned by one refugee at the table "Ukraine":

> It's important for us and we hold a lot of events or workshops about these things, so there has to be equal treatment for everyone. That is important. If I, for example, [organize] an event for the other, who are not Ukrainians, and they say: "We are second class here" and that's what destroys the integration process. And it is important to be all refugees equal, so that the integration process goes faster and better for all. That's ehh… Some people maybe have this feeling of "I'm not in", "I don't have this belonging", "I don't belong to this society" and then, "I can't move on", "I can't do anything positive" or something. (R4)

> Or the sense of belonging. That's an important dimension. (…). This belonging is important. […] when you see that I'm in, I can do something. Ehm, and I don't feel excluded, and I can, I want to do that. (R4)

These quotes do not only show the normative role of assessment practices, but also what they mean in detail for those affected: for the individual person not only the feeling of being excluded but also practical barriers such as access to the labor market or language courses (e.g., Kosyakova & Brenzel, 2020). From a temporal

perspective, this does not just mean that they are excluded, but remain excluded and are hindered to develop a "sense of belonging." The process of categorization, i.e., determining who is "worthy/unworthy" of getting asylum and (a certain type of) social service provision, does not stop at borders but is continued within the national borders. Hinger (2020) frames this by the idea of "integration through disintegration" (Hinger, 2020). The insights gained clearly point to issues, which are related to "latent racism"—or, as coined by one participant, "special assessment"—and the concern/request by affected people—refugees themselves, and also supporters, who are aware of the challenges of the system—that assessment criteria, and practices, should not make a difference among refugees.

Discussion

In the following, the results of the empirical material will be discussed in the light of AI use and some implications for fairness, highlighting legitimacy (of asylum bureaucracy) and agency (of refugees) as relevant concepts.

The first aspect refers particularly refers to the asylum procedure and the BAMF's "legitimization" of using AI in the asylum process. As for now, the question whether asylum applications are processed "in the best possible and fairest way for all parties" with the help of AI—or to put it in other words—whether a fair assessment is really taking place—is so far only answered by two stakeholder groups: BAMF/BMI and technology developers. At the same time, how the application of AI contributes to more fairness is not made transparent. The use of AI as a tool to verify the identity is in line with the politics of the BMI and the BAMF, trying to use generalizable data (here: language patterns): the accumulation of "fact-speaking" data, supporting documents, and AI-based technology are justified with making "better decision," i.e., via efficiency, (data) quality, and (inner) security. Creating efficiency and objectivity via "greater standardization," "large amounts of data," and "increased consistency" (Kinchin & Mougouei, 2022) is often used as an argument by governmental institutions in asylum bureaucracy, which mostly happens at the cost of "individuals' human rights, including privacy and security, and raises concerns about vulnerability and transparency" (Nalbandian, 2022). Regarding the concrete case of the DIAS, the AI-based system relies on a probability assessment, which is closely related to the geographical location, based on "generalizable" information. Insights of empirical research have shown, or confirmed, that the criteria for assessment are very much focused on the "country of origin." While this information is supposedly important due to probability, and therefore credibility, reasons, the current focus on geographical location as a crucial criterion for the protection status—being increased/co-determined by AI technology— can run the risk of neglecting more complex reasons or motivations for escape. The empirical data have shown that a political direction, including a "political" assessment who is entitled to (which kind of) protection status, *already* tends to neglect complex, individual refugee backgrounds. This tendency may be increased by AI, as it is currently applied.

Following this line of argument, AI used for identity verification increases the (artificial) construction of "social stratification of refugee groups and defining refugee identities" (Ahmad, 2021). In terms of legitimacy, the current AI use may contribute to strengthening asylum bureaucracy, but does not immediately or automatically (i.e., also considering the existing error rate of AI, i.e., ca. 15–20% of DIAS) lead to more fairness toward individual cases, and also certain groups.

The second fairness-related aspect highlights some problems in assessment practices taking place, in particular, beyond the asylum procedure: on the one hand, there are several initiatives taken by the BAMF to implement digital workflows and high technology, including AI, making processes more efficient; on the other hand, the lack of digitalized services, for example, in foreigners' offices, does not support making the asylum procedure and integration (e.g., access to social service provision, work permission) more efficient and "good"; to the contrary, refugees encounter many difficulties which can be (partly) traced back to an insufficient digital workflow within and among institutions. In 2021, the National Regulatory Control Council in Germany, established in 2006 as an independent body of experts to advise the Federal Government, the German Bundestag, and the Bundesrat on reducing bureaucracy and improving legislation, remarked the following: the COVID-19-pandemic and the "refugee crisis" have fundamentally illustrated that "a modern state must offer digital administrative services, make its internal processes digitally compatible and improve the exchange of data across disciplinary boundaries and administrative levels" (Nationaler Normenkontrollrat, 2021). Kühn and Heimann (2021) have provided a comprehensive analysis on issues resulting from a lack of digitalization, but also ways how institutions collaborate with each other (see also Bogumil et al., 2018), suggesting that "integration" and "digitalization" should be considered as going hand in hand. Current (AI-based) assessment practices in the BAMF and (non-AI based) assessment in other relevant institutions, like job center or foreigners' offices, are thereby not only problematic in terms of (procedural) fairness but also for refugee's agency in all processes. This is reflected by the workshop participants' statement that support plays a permanently important role. Apart from that, existing issues, as touched upon in 2.1, by the "asylum lottery," are currently not addressed in the Digitalization Agenda, or tackled by AI in the very early stage in the asylum procedure, i.e., before the hearing.[15] Looking at the federal as well as the local level, the examples indicate that "legitimate claims," e.g., fair processes regarding asylum procedures, often cannot be, or only partially, met. Although German public administration is traditionally "very much rule-driven" (Kuhlmann et al., 2021), there are these discrepancies to be found. Rather, refugees are dependent to a considerable extent on the local context, individuals in public administration, and supporting infrastructures.

[15] Turning back time around five years before the BAMF published its Digitization agenda and its introduction of IDM-S-Tools, several organizations (e.g., all welfare organizations in Germany) compiled a 60-page memorandum asking for fair and adequate asylum procedures (Amnesty International et al., 2016). Among the issues criticized in this memorandum were, e.g., a lack of democratic procedures, transparency, and organizational issues.

The insights gained through empirical data, and discussion of these, indicate the complexity of decision-making, supported by technology, e.g., AI, on many different levels, and its normative dimensions. The concepts of legitimacy (of asylum bureaucracy) and agency (of refugees) have been very briefly issued here. The legitimacy of asylum bureaucracy (e.g., political assessment being influenced by changing geopolitical dynamics and public opinion) and agency of those who are affected by these decisions (assessment taking place during the asylum procedure and beyond) require further analysis (cf. Josipovic, 2023). The research participants supported in identifying "blind spots" and tensions of these, the (political) role of assessment, and practices of othering "*versus*" developing a sense of belonging and being able to do something/having agency in the different processes.

Conclusion

This chapter explored AI use in the asylum procedure, put changed social assessment practices in the context of changing politics/policies, and highlighted some important insights gained by empirical social research conducted with refugees and supporters. Generally, exploring perspectives with refugees, and supporters, on the one hand, were crucial to understand "blind spots" regarding assessment practices and which criteria are relevant in the procedure and beyond (after the BAMF's decision has taken place). Exploratory interviews as well as desktop research helped to identify that current assessment practices entail different types of stratification, such as regarding access to social service provision, or the labor market; besides, AI does not play a role yet in stakeholders' daily work (i.e., integration "on the ground"); however, the lack of digital workflow does. It became (quite) clear that current AI use in the context at hand is not really present in public discourse. This might, directly or indirectly, have a negative impact on AI development in the field, as the case of mobile data extraction has shown, i.e., law lagging behind technology development. Importantly, the temporal dimension of an assessment's implications on refugee's life in Germany, the tension of asylum bureaucracy legitimacy, and agency of refugees should be considered: on the one hand, the political, normative nature of an assessment in this particular context (i.e., due to changing circumstances in countries of origin and "host" societies and geopolitical dynamics), and, on the other hand, the implications beyond this assessment for the individual refugee (i.e., processes of integration). As extracted in the analysis, the current AI-based technology, DIAS, focuses on one important/crucial criterion, namely country of origin. While making processes more efficient, better, and fair, as proclaimed by the BAMF, is also in the interests of refugees, fairness appears as a quite narrow concept in the official documents. In the light of empirical data collected and analyzed, the (existing) tension between individual case deliberation and categorization of countries/regions is not solved by the DIAS. Furthermore, existing fairness-related issues, such as different decision-making regarding protection status among federal statutes, are not considered by the BAMF in its Digitalization Agenda 2022.

Last, there are some important limitations to mention regarding the analysis. On the one hand, the analysis of practices, norms, and values of decision-makers are based on document analysis only. On the other hand, a media analysis is not included in the chapter although it sheds light on important value discourses, especially on the "welcome culture," solidarity discourses, and the crucial role of civil society (organizations). For example, regarding different assessment practices of refugees from Ukraine and those from other countries (here: Syria, Afghanistan, and Iran), more recent discourses point to the possibility to learn from these different situations (2015/2016 and 2022) and to facilitate "federal structures and civil society" to create equal chances and participation possibilities for refugees (Wagner et al., 2023).

Acknowledgments Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

References

Ahmad, N. (2021). Refugees and algorithmic humanitarianism: Applying artificial intelligence to RSD procedures and immigration decisions and making global human rights obligations relevant to AI governance. *International Journal on Minority and Group Rights, 28*, 367–435.

Amelung, N., & Galis, V. (2023). Border control technologies: Introduction. *Science as Culture, 32*(3), 323–343. https://doi.org/10.1080/09505431.2023.2234932

Amnesty International, AWO Bundesverband, BAfF – Bundesweite Arbeitsgemeinschaft der psychosozialen Zentren für Flüchtlinge und Folteropfer, Caritas, Der Paritätische Gesamtverband, Deutscher Anwaltverein, Diakonie Deutschland, JRS – Jesuiten-Flüchtlingsdienst, Neue Richtervereinigung, Pro Asyl, RAV – Republikanischer Anwältinnen- und Anwälteverein/RechtsBeraterKonferenz. (2016). *Memorandum für faire und sorgfältige Asylverfahren in Deutschland. Standards zur Gewährleistung der asylrechtlichen Verfahrensgarantien*. Retrieved November 15, 2023, from https://www.proasyl.de/wp-content/uploads/2015/12/Memorandum-f%C3%BCr-faire-und-sorgf%C3%A4ltige-Asylverfahren-in-Deutschland-2016.pdf

Bauer, A., & Schreyer, F. (2019). Ausländerbehörden und Ungleichheit: Unklare Identität junger Geflüchteter und der Zugang zu Ausbildung. *Zeitschrift für Rechtssoziologie, 39*(1), 112–142. https://doi.org/10.1515/zfrs-2019-0006

Beduschi, A. (2021). International migration management in the age of artificial intelligence. *Migration Studies, 9*(3), 576–596. https://doi.org/10.1093/migration/mnaa003

Biselli, A. (2017). Syrien oder Ägypten? Software zur Dialektanalyse ist fehleranfällig und intransparent. *Netzpolitik.org*. Retrieved May 11, 2023, from https://netzpolitik.org/2017/syrien-oder-aegypten-software-zur-dialektanalyse-ist-fehleranfaellig-und-intransparent/#netzpolitik-pw

Bither, J., & Ziebarth, A. (2022). *Automatisierte Entscheidungsfindung in der Migrationspolitik: eine Navigationshilfe*. Migration Strategy Group on International Cooperation and Development. Retrieved from March 23, 2023, from https://www.bosch-stiftung.de/sites/default/files/publications/pdf/2022-04/Paper_Automatisierte%20Entscheidungsfindung%20Migrationspolitik.pdf

Bogumil, J., Burgi, M., Kuhlmann, S., Hafner, J., Heuberger, M., & Krönke, C. (2018). Bessere Verwaltung in der Migrations- und Integrationspolitik. *Handlungsempfehlungen*

für Verwaltungen und Gesetzgebung im förderalen System. Modernisierung des öffentlichen Sektors, Sonderband 49. Nomos.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Brücker, H., et al. (2023). Ukrainian refugees in Germany: Evidence from a large representative survey. *Comparative Population Studies, 48*, 395–424. https://doi.org/10.12765/CPoS-2023-16

Bundesamt für Migration und Flüchtlinge. (2018). Integriertes Identitätsmanagement - Plausibilisieren, Datenqualität und Sicherheitsaspekte. Einführung in das IDM-S Tool Auslesen von mobilen Datenträgern (AmD). *Schulung AVS-Mitarbeiter. [Training material]* Retrieved October 10, 2023, from https://fragdenstaat.de/dokumente/9650-schulung_avs_kurz/

Cambridge Dictionary. (2024). *Othering*. Retrieved May 2, 2024, from https://dictionary.cambridge.org/dictionary/english/othering

Dahmen, D., Koch, Miriam, L., Abal, D., & Polat, F. (2017). «Gut», «schlecht», «unklar» – Die «Bleibeperspektive» und ihre Folgen für die Integration von Geflüchteten. In *Einwanderungsland Deutschland: Bericht der Kommission "Perspektiven für eine zukunftsgerichtete und nachhaltige Flüchtlings- und Einwanderungspolitik" der Heinrich-Böll-Stiftung*. Heinrich-Böll-Stiftung.

Denzin, N. (2006). *Sociological methods: A sourcebook* (5th edn). Aldine Transaction. ISBN 978-0-202-30840-1.

Deutscher Bundestag. (2022, August *22*). *Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Clara Bünger, Nicole Gohlke, Gökay Akbulut, weiterer Abgeordneter und der Fraktion DIE LINKE. – Drucksache 20/3133 – Einsatz von Dialekterkennungssoftware im Bundesamt für Migration und Flüchtlinge*. Retrieved October 27, 2023, from https://dserver.bundestag.de/btd/20/032/2003238.pdf

Eule, T. G. (2017). Ausländerbehörden im dynamischen Feld der Migrationssteuerung. In C. Lahusen & S. Schneider (Eds.), *Asyl verwalten. Zur bürokratischen Bearbeitung eines gesellschaftlichen Problems* (pp. 175–194). Transcript.

Federal Office for Migration and Refugees. (2021a, April). *The stages of the German asylum procedure. An overview of the individual procedural steps and the legal basis*. Retrieved March 3, 2023, from https://www.bamf.de/SharedDocs/Anlagen/EN/AsylFluechtlingsschutz/Asylverfahren/das-deutsche-asylverfahren.pdf?__blob=publicationFile&v=12

Federal Office for Migration and Refugees. (2021b, December). *Digitisation Agenda 2022*. Retrieved May 12, 2023, from https://www.bamf.de/SharedDocs/Anlagen/EN/Digitalisierung/broschuere-digitalisierungsagenda-2022.pdf?__blob=publicationFile&v=5

Federal Office for Migration and Refugees. (2024). *The decision of the Federal Office*. Retrieved November 16, 2023, from https://www.bamf.de/EN/Themen/AsylFluechtlingsschutz/AblaufAsylverfahrens/Entscheidung/entscheidung-node.html

Flick, U. (2007). *Qualitative Sozialforschung - Eine Einführung* (erw. und akt. Neuausg.). Rowohlt.

Grote, J. (2018). The Changing Influx of Asylum Seekers in 2014–2016: Responses in Germany. Focussed Study by the German National Contact Point for the European Migration Network (EMN). Working Paper 79 of the Research Centre of the Federal Office for Migration and Refugees, 2nd revised edition, Nuremberg: Federal Office for Migration and Refugees.

Grote, J. (2021). *Accurate, timely, interoperable? Data management in the asylum procedure in Germany*. Study by the German National Contact Point for the European Migration Network (EMN). Working Paper 90 of the Research Centre of the Federal Office for Migration and Refugees, Nuremberg: Federal Office for Migration and Refugees.

Hahlen, J., & Kühn, H. (2016). Die Flüchtlingskrise als Verwaltungskrise – Beobachtungen zur Agilität des deutschen Verwaltungssystems. *Verwaltung und Management 22. Jahrgang, Heft, 3*, 157–168. https://doi.org/10.5771/0947-9856-2016-3-157

Hahn, H. (2020). *Digital identification systems and the right to privacy in the asylum context. An analysis of implementations in Germany*. Master-Thesis. Leuphana University.

HandbookGermany Website. (2024). Retrieved March 15, 2024, from https://handbookgermany.de/en

Hinger, S. (2020). Integration through disintegration? The distinction between deserving and undeserving refugees in national and local integration policies in Germany. In S. Hinger & R. Schweitzer (Eds.), *Politics of (Dis)Integration* (IMISCOE Research Series) (pp. 19–39). Springer. https://doi.org/10.1007/978-3-030-25089-8_2

IT-Planungsrat. (2018). *Koordinierungsprojekt Digitalisierung des Asylverfahrens. Zusammenfassung der Projektergebnisse*. Retrieved May 15, 2023, from hBps://www.it-planungsrat.de/SharedDocs/Retrieveds/DE/Entscheidungen/27_Sitzung/TOP13_Anlage2_DigAsyl.pdf?__blob=publica,onFile&v=2

Josipovic, I. (2023). What can data justice mean for asylum governance? The case of smartphone data extraction in Germany. *Journal of Refugee Studies, 36*(3), 534–551. https://doi.org/10.1093/jrs/fead049

Kinchin, N., & Mougouei, D. (2022). What Can Artificial Intelligence Do for Refugee Status Determination? A Proposal for Removing Subjective Fear. International Journal of Refugee Law, 34(3–4), 373–397. https://doi.org/10.1093/ijrl/eeac040.

Kosyakova, Y., & Brenzel, H. (2020). The role of length of asylum procedure and legal status in the labour market integration of refugees in Germany. *Soziale Welt, 71*(1/2), 123–159. https://doi.org/10.5771/0038-6073-2020-1-2-123

Kottmann, L. (2023). *Haben Geflüchtete aus der Ukraine eine Art "Sonderstatus", Frau Alabali-Radovan?* Retrieved May 20, 2024, from https://www.integrationsbeauftragte.de/ib-de/medien/presse/interviews/-haben-gefluechtete-aus-der-ukraine-eine-art-sonderstatus-frau-alabali-radovan%2D%2D2162250

Kuhlmann, S., Proeller, I., Schimanke, D., & Ziekow, J. (2021). German public administration: Background and key issues. In S. Kuhlmann, I. Proeller, D. Schimanke, & J. Ziekow (Eds.), *Public administration in Germany* (pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-030-53697-8_1

Kühn, B., & Heimann, C. (2021). *Hand in hand? Datenmanagement in der lokalen Integrationsarbeit Forschungsgruppe Migrationspolitik Bestandsaufnahme und erste Befunde*. Working Paper 01_2021. Migration Policy Research Group (MPRG) & Robert Bosch Stiftung (Hrsg.). Retrieved June 15, 2024, from https://www.uni-hildesheim.de/media/fb1/sozialwissenschaften/Forschungsfokus_Migrationspolitik/Startseite/MRPG_WP01_Datenmanagement.pdf

Levy, K. (2020). *Local public administration and social policy in Germany and China. A comparative report with special attention to the welfare mix and provision of social services for migrants*. LoGoSO Research Papers No. 5.

Löhr, K., Weinhardt, M., & Sieber, S. (2020). The "World Café" as a participatory method for collecting qualitative data. *International Journal of Qualitative Methods, 19*, 1–15. https://doi.org/10.1177/1609406920916976

Lulamae, J. (2022). Kontroverse Dialekterkennung: Das BAMF und sein Pilotprojekt. *AlgorithmWatch*. Retrieved October 27, 2023, from https://algorithmwatch.org/de/dialekterkennung-bamf/

Manning, J. (2017). In vivo coding. In J. Matthes (Ed.), *The international encyclopedia of communication research methods*. Wiley-Blackwell. Retrieved November 15, 2023, from doi:https://doi.org/10.1002/9781118901731.iecrm0270.

Mediendienst Integration. (2024). *Flüchtlinge aus der Ukraine.* Retrieved March 22, 2024, from https://mediendienst-integration.de/migration/flucht-asyl/ukrainische-fluechtlinge.html

Meyer, D., Philipp, J., & Wenzelburger, G. (2021). Die Migrationspolitik der deutschen Länder. Eine mehrdimensionale Analyse. *Zeitschrift für Vergleichende Politikwissenschaft, 15*, 1–38. https://doi.org/10.1007/s12286-020-00474-1

Molnar, P. (2020). *Technological testing grounds: Migration management experiments and reflections from the ground up. European Digital Rights Network & the Refugee Law Lab*. York University.

Muy, S. (2020). Fördern, Fordern und Verbieten. Widersprüche in der Asyl- und Integrationspolitik aus Sicht der Sozialen Arbeit. In R. Pioch, & K. Toens (Hrsg.), Studien zur Migrations- und Integrationspolitik. *Innovation und Legitimation in der Migrationspolitik. Politikwissenschaft, politische Praxis und Soziale Arbeit im Dialog* (S. 271–291). doi:https://doi.org/10.1007/978-3-658-30097-5_17.

Nalbandian, L. (2022). An eye for an 'I:' a critical assessment of artificial intelligence tools in migration and asylum management. *Comparative Migration Studies, 10*(1), 32. https://doi.org/10.1186/s40878-022-00305-0

Nationaler Normenkontrollrat. (2021, September). *Monitor Digitale Verwaltung*. Retrieved May 13, 2023, from https://www.normenkontrollrat.bund.de/Webs/NKR/SharedDocs/Retrieveds/DE/Positionspapiere/monitor-digitale-verwaltung-6.pdf?__blob=publicationFile&v=9

Ozkul, D. (2023). *Automating immigration and asylum: The uses of new technologies in migration and asylum governance in Europe*. Refugee Studies Centre, University of Oxford.

Palmiotto, F., & Ozkul, D. (2023). "Like handing my whole life over". The German federal administrative court's landmark ruling on mobile phone data extraction in asylum procedures. *Hertie-School.org*. Retrieved May 7, 2023, from https://www.hertie-school.org/en/news/detail/content/like-handing-my-whole-life-over-the-german-federal-administrative-courts-landmark-ruling-on-mobile-phone-data-extraction-in-asylum-procedures

Rambøll Management Consulting GmbH & Nationaler Normenkontrollrat Sekretariat. (2014). Lebenslagen von Asylbewerbern. *Vorschläge zur Verwaltungs- und Verfahrensvereinfachung*. Vorstudie. Herausgegeben von Robert Bosch Stiftung GmbH.

Raschke, M. (2023). Rechtliche Grundlagen der integration. In F. Bätge, K. Effing, K. Möltgen-Sicking, & T. Winter (Eds.), *Integration in Kommunen. Bedeutung, aktuelle Entwicklungen und Perspektiven aus Theorie und Praxis* (pp. 35–54). Kommunale Politik und Verwaltung. Springer VS.

Riedel, L., & Schneider, G. (2017). Dezentraler Asylvollzug diskriminiert: Anerkennungsquoten von Flüchtlingen im bundesdeutschen Vergleich, 2010-2015. *Politische Vierteljahresschrift, 58*. Jahrgang, 21–48. doi:https://doi.org/10.5771/0032-3470-2017-1-21.

Schammann, H. (2015). Wenn Variationen den Alltag bestimmen. Unterschiede lokaler Politikgestaltung in der Leistungsgewährung für Asylsuchende. *Zeitschrift für Vergleichende Politikwissenschaft, 9*, 161–182. https://doi.org/10.1007/s12286-015-0267-4

Schammann, H., & Gluns, D. (2021). *Migrationspolitik*. Nomos Verlag.

Schiele, H., Krummaker, S., Hoffmann, P., & Kowalski, R. (2022). The "research world café" as method of scientific enquiry: Combining rigor with relevance and speed. Journal of business research, 140, 280–296. https://doi.org/10.1016/j.jbusres.2021.10.075.

Schneider, G., & Riedel, L. (2017). The asylum lottery: Recognition rates vary strongly within Germany. *EU Migration Law Blog*. Retrieved April 14, 2023, from https://eumigrationlawblog.eu/the-asylum-lottery-recognition-rates-vary-strongly-within-germany/#more-1485

Sixtus, F., Kiziak, T., & Klingholz, R. (2019). *Von individuellen und institutionellen Hürden. Der lange Weg zur Arbeitsmarktintegration Geflüchteter*. Diskussionspapier 23. Berlin-Institut für Bevölkerung und Entwicklung.

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage.

Tangermann, J. (2017). *Documenting and establishing identity in the migration process. Challenges and practices in the German context*. Focused study by the German National Contact Point for the European Migration Network. Working Paper 76 of the Research Centre of the Federal Office for Migration and Refugees. Federal Office for Migration and Refugees.

Wagner, T., van den Berg, C., Sedding, M., Steinhilper, E., Hutter, S., Schwenken, H., & Zajak, S. (2023, May 2). *Engagement für Geflüchtete: Was bleibt von 2015? Bundeszentrale für politische Bildung*. Retrieved May 23, 2024, from https://www.bpb.de/themen/migration-integration/kurzdossiers/520529/engagement-fuer-gefluechtete-was-bleibt-von-2015/

Chapter 7
Social Assessment for the Targeted Subsidies Plan as a Social Service Provision in Iran: AI Application in the Targeted Subsidies Plan

Hassan Bashiri

Abstract As an additional case study of the AI FORA research project, the chapter delves into the intricate relationship between AI and public policy by investigating its application within the Iranian Targeted Subsidies Plan (TSP). The research methodology encompasses experimental techniques, involving data collection, document analysis, interviews, questionnaires, and quantitative data analysis. This chapter not only underscores the remarkable potential of AI in optimizing social services but also highlights challenges such as data access, privacy concerns, and governance issues. Findings reveal that the integration of AI in the TSP has led to substantial financial savings over the past decade. While the TSP initially succeeded in reducing poverty and narrowing the wealth gap, it ultimately fell short of its primary objective due to various factors, including economic instability. AI-driven algorithms have enhanced the accuracy and fairness of household eligibility assessments. Furthermore, the study demonstrates a remarkably high level of public acceptance (87%) and trust (79%) in AI's role within the TSP. In conclusion, the study showcases how AI has become a transformative force in data-driven policy-making within Iran's TSP, facilitated by the IWDB.

Introduction

Human society has become a socio-technical system using artificial intelligence technology (OECD, 2019). AI serves a dual role, acting as a powerful tool to enhance policies and decisions for citizens' improved quality of life, but also posing challenges as it can be exploited for criminal activities. Given the diversity of values, norms, technological advancements, economic models, and civil society sentiments across countries, attitudes toward the use of AI for public policy differ

H. Bashiri (✉)
Department of Computer Science, Hamedan University of Technology, Hamedan, Iran
e-mail: bashiri@hut.ac.ir

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*,
Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_7

substantially. Moreover, such attitudes may vary not only among countries but also within different social groups.

Increasingly, AI algorithms are being adopted by public administrations in various countries to determine the allocation of public services among citizens. This involves assessing individuals' profiles based on specific criteria to distinguish between eligible and ineligible recipients, addressing issues of legality, deservingness, and neediness (Furtado et al., 2015; Valle-Cruz et al., 2019). For instance, China's Social Credit System, launched in 2016, ranks citizens through continuous monitoring and influences access to public services (Creemers, 2018; Kshetri, 2020).

Therefore, it becomes imperative to carefully examine AI-based assessments and seek ways to enhance AI's performance in the realm of social assessment. Professional Participatory Technology Assessment innovations aim to create suitable institutional frameworks to meet these evolving needs. One such initiative is the "Artificial Intelligence for Assessment (AI-FORA)" an international project managed by the Social Simulation Laboratory of Innovation and Technology at the Johannes Gutenberg University of Mainz (*AI for Assessment*, 2020).

AI FORA's primary focus is the intricate interplay between AI technologies assessing society and society's assessment of AI technologies, with social values serving as a pivotal factor in this reciprocal relationship. The implications of AI FORA research extend far beyond mere technology, holding significant moral and social consequences for our global communities. Thus, addressing the question of "where are we going" necessitates scientific methodologies and participatory approaches (Ahrweiler, 2019). Crucially, the development of technology, particularly AI-based social assessments, must be grounded in societal values, requiring expertise, participation, and collaborative design across all social groups, facilitated by participatory tools.

In this chapter, we focus on exploring the application of AI in the context of social assessments, utilizing the Targeted Subsidies Plan (TSP) in Iran as a specific case study. Implemented by the Government of Iran since late 2010, the TSP aims to provide targeted government subsidies to low-income households. We delve into the use of artificial intelligence and data-driven decision-making in identifying eligible households for the TSP; we'll refer to it as "household decile ranking." This examination will shed light on its potential ramifications for social service provisions and its broader societal implications.

Research Objectives

This study contributes to the AI FORA project's goal of exploring the transformative potential of AI in tackling societal challenges by investigating its integration within Iran's Targeted Subsidies Plan. Examining the intricate relationship between AI and public policy, this research aims to:

- Evaluate the impact of AI technologies on targeted subsidy policies, service delivery mechanisms, and beneficiary outcomes within the TSP framework, considering the specific cultural values and norms that shape its implementation and reception.
- Assess stakeholder acceptance and perception of AI integration, including policymakers, implementers, and beneficiaries, with a particular focus on how their cultural backgrounds influence their attitudes and concerns toward this approach in social service provision.
- Analyze the effectiveness of AI-driven strategies in optimizing resource allocation, improving eligibility assessment accuracy, and enhancing overall program efficiency, paying close attention to how these strategies interact with and potentially adapt to the Iranian cultural context.
- Identify key challenges and opportunities associated with AI implementation in the TSP, particularly those related to data privacy, governance, equity, and access to technology, considering the unique cultural sensitivities and nuances present in Iran.
- Develop insights and recommendations for policymakers, practitioners, and stakeholders to leverage AI effectively in the design, implementation, and evaluation of social welfare policies and programs in Iran, ensuring that AI solutions are tailored to the specific cultural context and promote equitable and socially responsible outcomes.

By focusing on the Iranian TSP case, this study aims to contribute valuable empirical evidence and theoretical insights to the ongoing discourse on AI's role in social service provision, both nationally and globally.

Targeted Subsidies Plan

For decades, Iran has implemented a subsidy system where the level of benefit from government subsidies increases with higher consumption of goods and services. Paradoxically, this approach means that wealthier individuals tend to benefit more from these subsidies. As depicted in Fig. 7.1, the government provides subsidies for essential commodities and energy resources. This setup, however, creates a situation where the more affluent consumers receive greater support through subsidized access to basic goods and cheap energy. As a consequence, this hidden subsidy mechanism exacerbates social inequalities and discourages efforts to save energy and improve productivity in both consumption and production patterns.

The TSP represents the most extensive economic project in Iran's history, initiated by the government in late 2010. As per the parliament's decision, the government undertook responsibilities across various sectors of the economy by adjusting the prices of energy carriers to bring them closer to their actual cost. The revenue generated from these price adjustments was intended to fund the targeted subsidies plan.

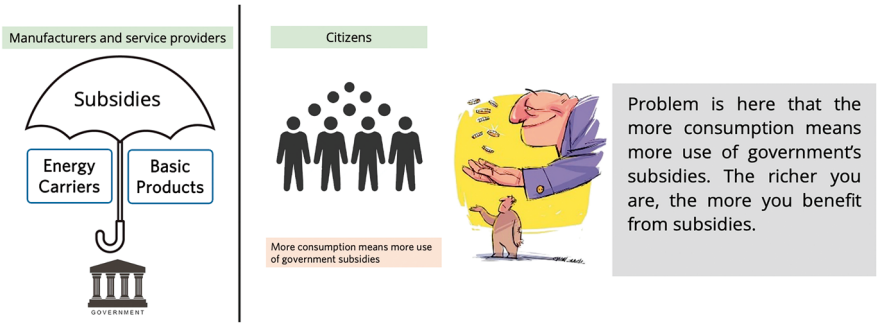

Fig. 7.1 For many years, the government has been providing subsidies for essential commodities and energy carriers (Author)

The rationale behind the TSP was to address the issue of wealthier individuals benefiting more from traditional subsidy allocation methods for goods. The plan aimed to remove subsidies from commodities and instead provide direct support to low-income households or struggling producers who faced rising energy prices.

To gain insight into the allocation of government financial resources and expenditures, the 2022 budget bill of the country includes a section dedicated to the Targeted Subsidies Plan, visually presented as an info-graphic in Fig. 7.2. The figure illustrates that a substantial portion of government revenues is directed toward providing cash subsidies to households and funding poverty reduction and public health initiatives. Conversely, the primary source of government revenues will stem from oil exports and increasing energy prices for domestic consumers.

Summary of the Law

The Iranian Parliament approved and communicated the Law on Targeted Subsidies in January 2010, with its implementation by the government commencing in December of the same year. Key points notified for execution in this law include:

- Price adjustment is envisioned over a 5-year period.
- At the end of the period, the average price of energy carriers (such as gasoline, diesel, fuel oil, kerosene, and other petroleum products) will be set to at least 90% of the average FOB (free on board) prices in the Persian Gulf.
- The average price of crude oil and gas condensate delivered to the country's refineries will be at least 95% of the FOB Persian Gulf prices.
- The average price of natural gas will be at least 75% of the average export price for natural gas.
- The average selling price of water and electricity will be based on their cost price.
- Domestic prices will not be affected by fluctuations in FOB Persian Gulf prices of energy carriers, up to 25%.

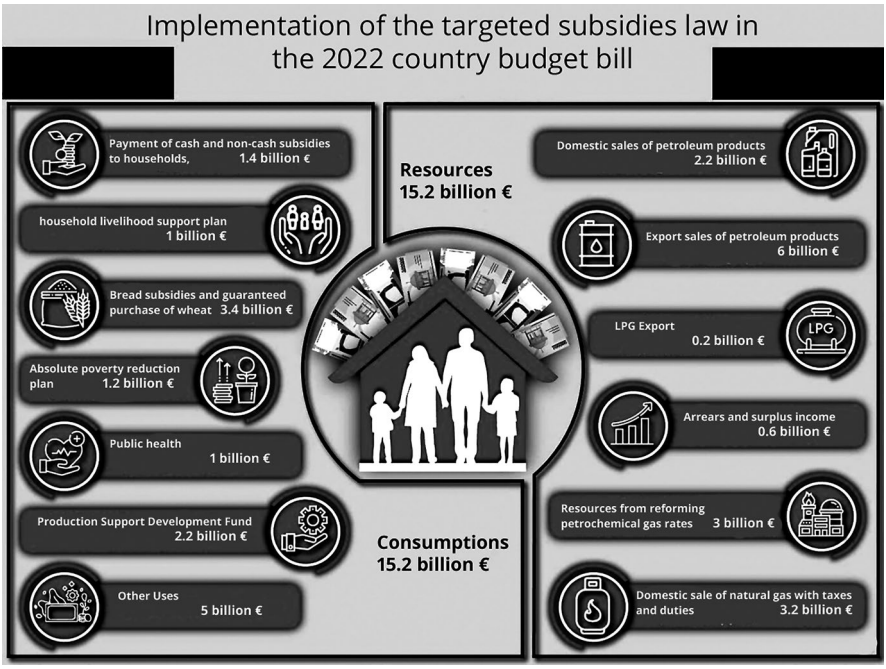

Fig. 7.2 The 2022 budget bill of the country outlines the financial resources and expenditures. (How to implement the law on targeted subsidies in the 1401 budget bill, 2022)

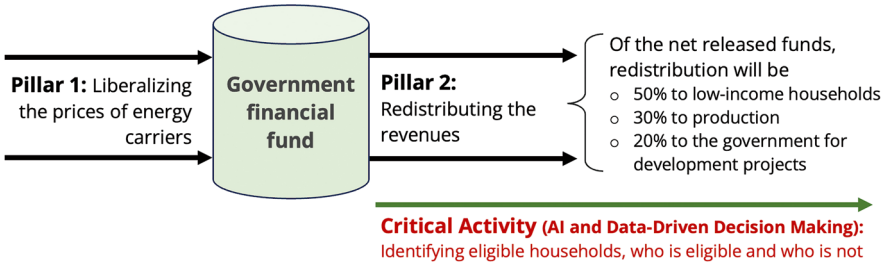

Fig. 7.3 Two pillars and one key activity in TSP (Author)

- Subsidies on commodities will be eliminated on goods and services such as wheat, rice, milk, sugar, postal services, airline services, and rail (passenger) services.
- Of the net released funds, 50% will be redistributed to low-income groups, 30% to production, and 20% will be allocated to the government for development projects.

The Targeted Subsidies Plan consists of two primary pillars and a critical activity (Fig. 7.3):

- **Pillar 1:** Phasing out subsidies and liberalizing the prices of energy carriers, such as gasoline, kerosene, electricity, water, and others.
- **Pillar 2:** Redistributing the revenues obtained from eliminating indirect subsidies and implementing price liberalization.
- **Critical activity:** Identifying eligible households for targeted assistance.

Cultural Context

To illustrate the cultural landscape of Iran and assess the potential for engagement of various stakeholders in AI-driven social service provision, we employed Hofstede's six cultural dimensions as a framework (Hoftede et al., 2010). Additionally, for comparative analysis, we utilized Hofstede's cultural framework tool, juxtaposing Iran with Germany, a focal point in the AI FORA research initiative (Hofstede, 2024). This comparative approach offers valuable insights into the cultural contexts shaping attitudes toward technology adoption and collaborative efforts in social service delivery. Figure 7.4 visually represents this comparative analysis, providing a comprehensive view of the cultural nuances influencing the feasibility of AI integration in social services within different socio-cultural contexts.

- **Power distance:** High (58) reflects a hierarchical society where decisions are largely *top-down*. This impacts technology adoption as central authorities play a strong role in implementation and direction, potentially limiting grassroots innovation and collaborative approaches.
- **Collectivism:** High (23) suggests strong group loyalty and responsibility. Social services might be delivered effectively through established community networks and family structures, but individual initiative might be less encouraged.
- **Uncertainty avoidance:** High (59) indicates a preference for clear rules and regulations. This could lead to rigid structures in technology use and social services, potentially hindering adaptation and flexibility.

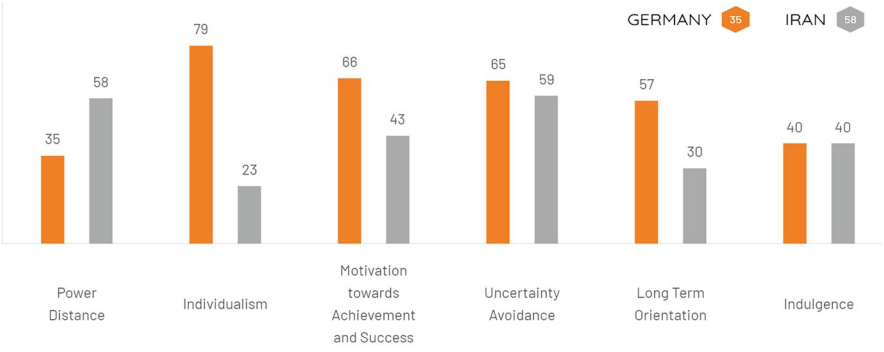

Fig. 7.4 Iran and Germany—Hofstede's six cultural dimensions as a framework (Hofstede, 2024)

- **Consensus:** Moderate (51) suggests a balance between top-down authority and collaborative decision-making. This could facilitate inclusive approaches to technology integration and social service design, but efficiency might be impacted by lengthy consensus-building processes.
- **Normative vs. pragmatic:** Strongly normative (30) highlights the importance of tradition and established practices. This might favor existing methods in social service delivery and resist rapid technological change.
- **Restraint vs. indulgence:** Restrained (40) reflects a focus on self-control and limited leisure time. Technology use might be primarily for practical purposes, and social services might be perceived as a duty rather than an individual entitlement.
- **Overall:** These dimensions paint a picture of a culture where collective well-being takes precedence over individual initiative, and established structures guide decision-making in technology and social services. While this fosters collaboration and social responsibility, it might also create challenges for flexibility, innovation, and bottom-up approaches.

Actors Network in TSP

In order to implement the Targeted Subsidies Plan effectively, an online system was developed to collect financial information from households. This system, established by the Statistics Center of Iran and overseen by the newly formed Targeted Subsidies Organization, served as a centralized platform for households to disclose their income and assets. Founded in 2010 according to Article 15 of the Targeted Subsidies Law, the Targeted Subsidies Organization operates as a state entity with a mandate to enhance the equitable distribution of government subsidies, mitigate inflationary pressures, and alleviate economic burdens on vulnerable segments of society. Presently, the organization operates under the auspices of the country's Plan and Budget Organization and assumes responsibility for disbursing subsidies among eligible households. The Ministry of Cooperation, Labor, and Social Welfare plays a pivotal role in the identification and categorization of households.

The primary objective of this initiative was to categorize households into distinct social deciles, to exclude the top three deciles characterized by higher income levels from subsidy entitlements while directing cash assistance toward the remaining deciles. However, challenges arose due to the unreliability of the data recorded in the household economic database. A significant number of households understated their income to qualify for government subsidies, thereby compromising the accuracy and integrity of the subsidy allocation process. This phenomenon is not uncommon, as similar issues regarding income disclosure have been observed in other developing countries (Doshmangir et al., 2015). Consequently, the Iranian parliament decided to provide all registered households with a uniform subsidy of 450,000 Rials during the initial phase, inclusive of 45,000 Rials designated for bread subsidy (equivalent to $42 at the time). Additionally, a subsidy was allocated for energy

consumption in industries. However, in 2022, in response to changes in economic conditions and currency devaluation, the government decided to revise the subsidy distribution framework. From 2022 onward, the subsidy amount per person for households in deciles 1–3 increased to 4,000,000 Rials (approximately $8), while for deciles 4–9, it rose to 3,000,000 Rials (around $6) per person. Notably, households falling within the tenth income decile do not receive subsidies.

Identifying eligible households and determining the redistribution of financial resources resulting from price adjustments have emerged as critical issues in TSP implementation (Bakhshoodeh, 2013). There is no perfect targeting method in practice. As for the redistribution of resources released from energy carrier and targeted goods price adjustments, officials initially emphasized the cash redistribution approach. They asserted that the calculation of cash subsidies had taken into account various factors, including household economic conditions, the repercussions of energy and commodity price adjustments on the prices of other goods, household expenditures, and income, and the influence of energy carrier price corrections on the production of goods and services.

In the government, two main actors are responsible for implementing TSP. The first is the Targeted Subsidies Organization, established within the government to collect household financial information and redistribute income. The second is the Deputy of Social Welfare from the Ministry of Cooperatives, Labor, and Social Welfare, which identifies subsidy beneficiaries and social deciles using Test Means. As depicted in Fig. 7.5, in accordance with the executive regulations of Article (7) of the Law on Targeted Subsidies, households first provide all their identity and financial information through the electronic system to identify vulnerable and target groups specified in Article (7) of the law. The Ministry of Cooperatives, Labor, and Social Welfare is responsible for identifying subsidized and vulnerable groups, as well as low-income groups residing in deprived areas of the country. The Targeted Subsidies Organization is also responsible for depositing subsidies into the head of the household's account.

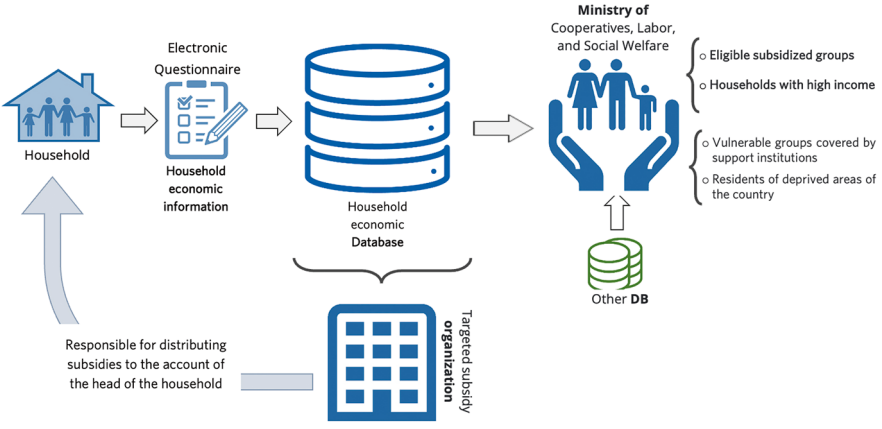

Fig. 7.5 The process of identifying eligible households and distributing subsidies (Author)

AI and the Identification of Eligible Households

Over time, economic conditions and currency devaluation, exacerbated by sanctions, resulted in a considerable reduction in the value of Iran's national currency. In response, the government implemented changes to the subsidy distribution framework. This adjustment underscores the government's endeavors to alleviate the impact of economic challenges on vulnerable demographics, albeit against a backdrop of currency devaluation and economic instability. Accurate identification of income deciles and the exclusion of high-income deciles became increasingly crucial in this context.

To identify eligible households, artificial intelligence played a more prominent role than in the past. Various parameters, such as income and assets, are now examined using AI without human intervention to determine entitlement to livelihood support. The regulation specifies that households with relatively high incomes, favorable financial indicators over the past 3 years, residential properties, employment, and multiple cars, or those who have taken more than three foreign trips, will be excluded from receiving livelihood support assistance.

The importance of using artificial intelligence in this identification process is emphasized in Note 1 of Article 3 of the regulation. The average financial indicators of the last 3 years or the average withdrawal of the 6 months preceding the determination of eligibility serve as the basis for identifying the income level and scoring the household. The eligibility of households for subsistence support and other benefits is determined automatically by AI and machines, without any human intervention in the results. Moreover, viewing and checking bank information is prohibited for individuals and officials involved in the process.

Iranian Welfare Database

The "Iranian Welfare Database (IWDB)" was initiated in February 2014 to establish the welfare-economic identity of individuals in Iran. This comprehensive database is constructed from a collection of 50 data sources, with the initial phase integrating 25 data sources to form the main structure of the database, consisting of over 60 data tables. More than 3 billion data records have been stored and integrated into this extensive database. Currently, the IWDB encompasses 221 information fields that directly and indirectly describe various aspects of citizens' identities and economic situations. The primary data sources utilized for this database include records from the Civil Registration Organization, Property and Deeds Registration Organization, Police, Social Security Organization, other pension funds, Central Bank, Tax Affairs Organization, Chamber of Commerce and Trades, Ministry of Education, Ministry of Health, and Staff Systems managed by the Ministry of Cooperatives, Labor, and Social Welfare.

The Ministry of Cooperatives, Labor, and Social Welfare gathers citizens' financial behavior information from the aforementioned databases. For instance, it receives data about foreign travel and car ownership from the Police database and information about citizens' exports and imports from the Chamber of Commerce and Trades. The process of integrating various data sources and creating a data warehouse is depicted in Fig. 7.6, with related public organizations also benefiting from the available services.

The data obtained and government policies serve as criteria for Test Means, which is used to determine household assets, economic deciles, and ten income deciles (ranging from the wealthiest to the poorest households). Identifying individuals within each income decile is essential for the government to achieve the objectives outlined in the TSP law, making the analysis of information from the IWDB a critical decision-making pillar.

The information in this database has played a pivotal role in implementing national programs across three distinct time periods:

1. In 2019, during the payment of fuel subsidies and the supporting living package, around 60 million people (18 million households) were identified for receiving this support package through the analysis of information from the IWDB.
2. During the outbreak of COVID-19, the database was instrumental in identifying beneficiaries for receiving bank loans to assist those affected by the pandemic. Different types of loans were transferred to the accounts of eligible citizens at different stages based on their individual needs.
3. The recent utilization of the database involved eliminating the subsidy for high-income deciles, in accordance with the parliament's law. The implementation of this measure required the identification of high-income deciles, which was accomplished with the help of information stored in the database.

Analysis of Dataset

To strike a balance between enabling the research community to access data from the "Iranian Information Database" for research purposes and safeguarding the

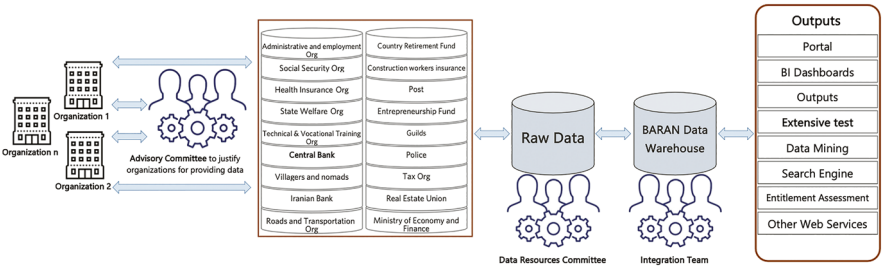

Fig. 7.6 Process of database integration from different data sources and services which are available for related public organizations (Author)

privacy of Iranian citizens to the greatest extent possible, the Ministry of Social Welfare has undertaken the task of creating 2% random samples derived from the IWDB. In preparing these samples, careful consideration has been given to the following key aspects:

- **Stratified sampling:** The sampling approach employed is stratified, meaning that the statistical characteristics and distribution of the created sample mirror those of the original population. This strategic method ensures that any analyses conducted on this statistical sample can be effectively extrapolated to the broader statistical population.
- **Privacy protection:** A paramount concern has been the protection of individuals' privacy. As such, none of the identifying features—such as national identification numbers, postal codes, mobile phone numbers, car chassis numbers, and similar data points that uniquely pertain to an actual individual—have been included in the sample file. These identification features, which hold no analytical significance, have been deliberately omitted to avoid any issues during researchers' analyses.

In this step of the research, we aimed to gain a deeper understanding of the TSP by analyzing the dataset from the IWDB. We utilized the Python programming language along with libraries such as Pandas, Numpy, and Matplotlib for our data analysis. A more precise comprehension of the data is crucial for the subsequent phase of our research, which involves agent-based modeling. This modeling relies on various indicators derived from data analysis, including the distribution of family sizes across the country, income distribution, and the relationship between income and expenditures. These insights aid in designing a targeted subsidies model that closely aligns with the actual behavior of the data.

The initial step of data analysis involved data preprocessing. During this phase, we thoroughly examined, cleaned, and prepared the data for subsequent analysis. This process encompassed tasks such as rectifying unclear or incomplete data, eliminating duplicate or anomalous data points, scaling the data, and converting it into usable formats. For instance, upon inspecting the data, we discovered that a negligible fraction, specifically 14,554 individuals out of a total of 1,048,576 (less than 0.01% of the records), lacked a date of birth. Since this field is essential for constructing the population pyramid of the country, we imputed the average age of individuals into these missing entries. To delve into the distribution of family sizes, we leveraged the "ParentID" feature to ascertain family sizes and visualize the results. Figure 7.7 illustrates that family size distribution in Iran closely follows a normal distribution, with the majority of families consisting of four members.

Determining household income was based on the total income of households. We utilized density charts to effectively depict income distribution across the country. Figure 7.8 presents a representation of income distribution in the country, demonstrating a power distribution pattern that can be inferred.

Furthermore, we conducted an analysis of the Lorenz curve for household income distribution. The Lorenz curve serves as a valuable tool for scrutinizing income distribution, enabling economists to evaluate financial policies and compare

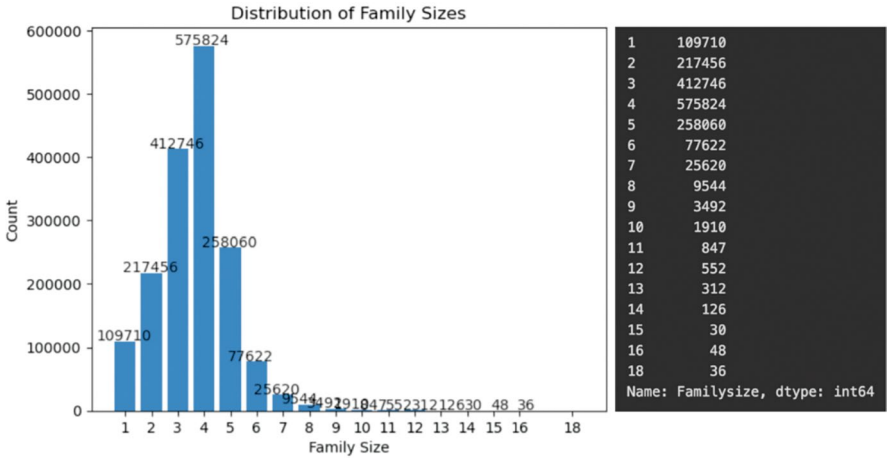

Fig. 7.7 Family size distribution

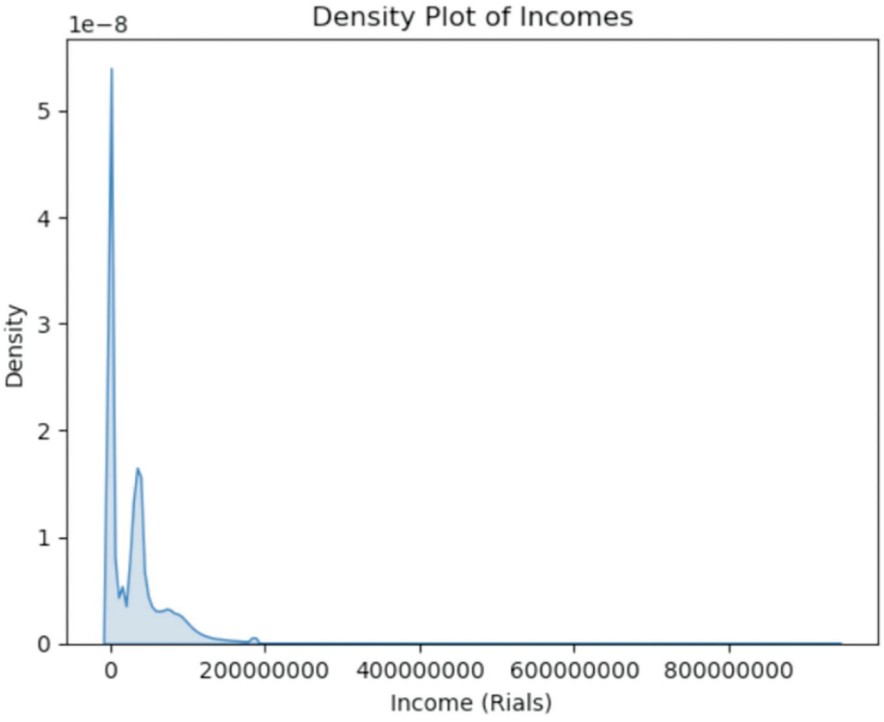

Fig. 7.8 Income distribution

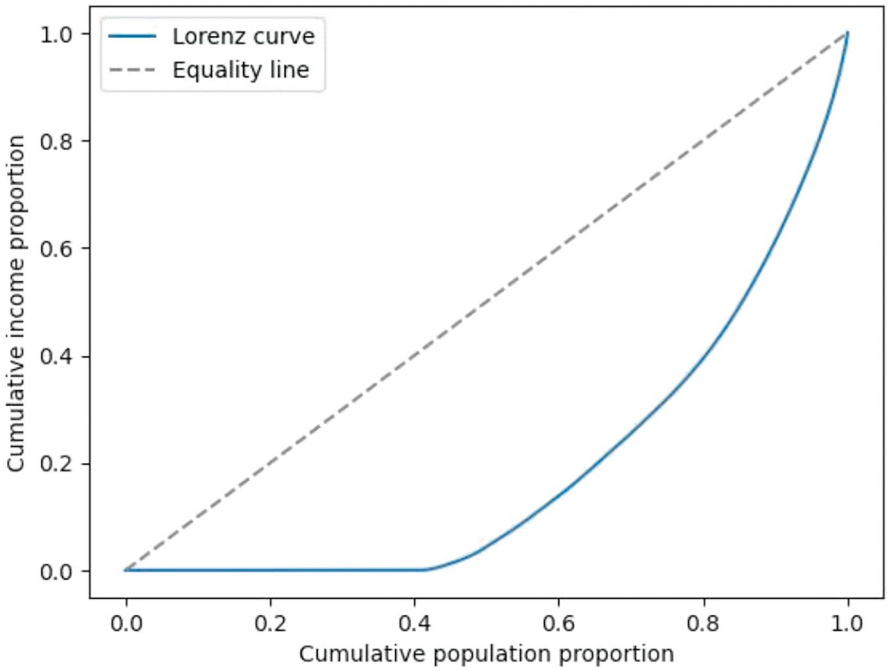

Fig. 7.9 Lorenz curve of income distribution

income disparities across societies and time periods. This curve represents a cumulative abundance curve, with the equilibrium line, denoting income equality, as the reference point. The greater the deviation from this equilibrium line, the more skewed the income distribution (Fig. 7.9). Our analysis indicates that between 2017 and 2021, income distribution became increasingly unequal.

Lastly, we explored the relationship between expenditure and income in 2017. Our findings revealed a positive correlation: as income increased, so did expenditures (Fig. 7.10). This pattern persisted in subsequent years, including 2018, 2019, and 2020.

Eligibility of Households Using Test Means

To create a comprehensive test, various types of data available in the IWDB were utilized, with some of these data points combined to generate new indices. The new index is calculated and evaluated using a linear regression model. In the most recent implementation of Test Means, the following quantities were used:

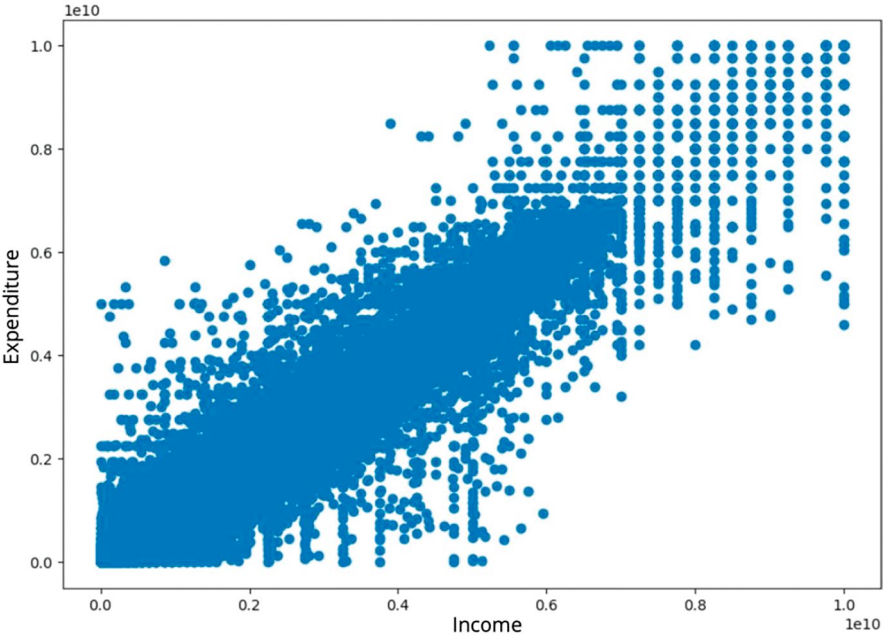

Fig. 7.10 Relationship between income and expenditures of households in 2017

A) **Initial quantities:** These are raw and unchanged data points, including:

- Family size
- Gender of the head of the household
- Number of vehicles
- Total car value
- Declared property number
- Number of foreign trips

B) **Composite quantities:** These quantities are created by combining several other data points, including:

- Being a doctor
- Being a public manager
- Being employed, encompassing various professions such as those in social security, doctors, public managers, lawyers, faculty members, individual employers, legal employers, exchange offices, employees, and guilds
- Age of the head of the household, categorized as the young head (under 25 years old), middle-aged head (between 25 and 54 years old), and elderly head (over 54 years old)
- Average age of the household, mapped to tags like young household (under 29 years old), middle-aged household (29–46 years old), and elderly household (over 46 years old).

- Number of people of working age
- Weighted dimension of the household: It reflects the distribution of household members based on age groups—child (under 11 years), adolescent (between 11 and 17 years), and adult (over 17 years).
- Province and region of residence
- Mean housing value

The decimalization process, as overseen by the Deputy of Social Welfare within the Ministry of Labor, Cooperative and Social Welfare, is meticulously detailed. In compliance with Article 29 of the 6th Development Plan, all executive government organizations are mandated to submit accurate and verifiable registration data biannually to the Iranian Welfare Database.

This process entails the categorization of households and individuals based on comprehensive income and asset data. Furthermore, each income decile is subdivided into 10 equal segments, known as percentiles, to precisely ascertain the income bracket of an individual, thereby enhancing the accuracy of income measurement.

For instance, data released by the Social Welfare Department in September 2023 illustrates that individuals with an annual income below 300 million Rials (approximately $750[1]) are categorized within the lowest income decile, placing them in the bottom 10% of the societal income distribution. Conversely, individuals earning between 1160 and 1410 million Rials annually (approximately $2900 to $3525) fall within the fifth decile. Those in the highest income bracket, or the 10th decile, earn upward of 2800 million Rials annually (approximately $7000). Refining the classification into 100 parts, rather than 10, identifies the uppermost income percentile as those earning over 4290 million Rials annually (about $10,750). Individuals with a monthly income exceeding 350 million Rials ($875) are thus considered to be in the top percentile, surpassing 99% of the population in terms of income (Delazimi Farideh, Amraei Mojtaba, 2023).

Iran's population, estimated at 85 million, sees approximately 78 million individuals receiving cash subsidies. The distribution of these subsidies is tiered, with 30% of the population in the lowest three deciles receiving 4,000,000 Rials monthly (about $10), and 60% in the fourth to ninth deciles receiving 3,000,000 Rials monthly (about $7.5). The top 10% of earners do not qualify for subsidies. Notably, the income of households in the 10th decile is nearly tenfold that of those in the first decile. The Ministry of Cooperation, Labor, and Social Welfare's statistical yearbook reveals that the monthly expenses of the 10th decile households are six times those of the first decile. When comparing the 9th decile, considered the societal middle class, to the 10th decile, there is a 40% income disparity, with a corresponding 33% increase in expenses. Collectively, the first through third deciles comprise

[1] Based on the 2022 average exchange rate. The data used for the decimation of 2023 is related to the three years ending in 2022, which is why we used the average dollar conversion rate in 2022 to calculate the approximate annual income in dollars.

25 million people, indicating a significant portion of the population within these income brackets (Delazimi Farideh, Amraei Mojtaba, 2023).

Welfare Atlas Based on IWDB

The Welfare Atlas based on the IWDB has been prepared for over a hundred cities. By using the national code and six digits of the postal code, the location of each person has been determined, allowing for the depiction of poverty, prosperity, and inequality distribution within cities. The welfare maps provide valuable insights into various aspects of a city, including the distribution of the poor population in neighborhoods, the residence of female heads of households, households without insurance, households with disabled members, and households without fixed monthly income.

The atlas presents 24 indicators of poverty and social welfare for each province and city, facilitating comparisons between different regions and the national average. Each neighborhood's economic decile is estimated and depicted on city maps, leading to interesting findings. For example, cities like Ahvaz or Mashhad show a concentrated population of poverty, forming poverty clusters, while cities like Isfahan or Tabriz exhibit a more scattered distribution of the poor population. Additionally, in Tehran, the majority of the population falls into the top 4 income deciles, whereas in cities like Iranshahr, the majority belong to the bottom 4 income deciles.

Further examination of the data from the atlas reveals additional insights. Certain cities, such as Urmia, Khomeinishahr, Ilam, and Mashhad, exhibit a clear division between rich and poor areas. For instance, the rich families of Urmia reside in the south, while lower decile families live in the north. Conversely, Khomeinishahr and Ilam show the opposite pattern, and in Mashhad, the upper deciles are concentrated in the west, while the lower deciles are in the east. In other cities, such as Ahvaz, Fars, Qom, and Kermanshah, the rich and poor neighborhoods are arranged in a checkerboard pattern next to each other (MCLS, 2021).

For a visual representation, Fig. 7.11 displays the welfare atlas of Mashhad City in Razavi Khorasan province. The information provided in this atlas offers valuable insights into the socioeconomic landscape of Iranian cities and aids policymakers in making informed decisions to address inequalities and improve living standards for all citizens.

Data Mining

Following the integration and evaluation of raw data, various analyses were conducted on the initial data using data mining methods. Some notable examples of data mining outputs include:

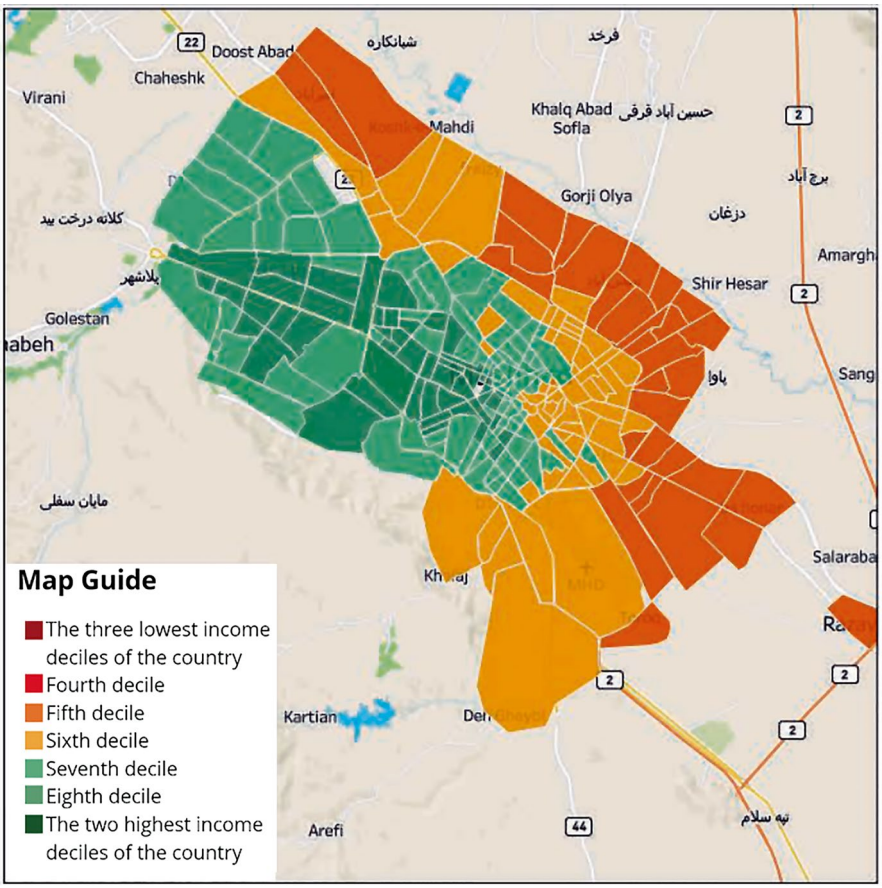

Fig. 7.11 Welfare atlas of Mashhad, distribution of the residence of households in different deciles in Mashhad city (MCLS, 2021)

- Collaboration with the Tax Affairs Organization (TAO) to identify 300,000 wealthy households without tax records: Based on Test Means and assets of Iranian households, those without tax records were identified and introduced to TAO.
- Analysis of the status of outstanding students (elites) in the national entrance exam for the National Elite Foundation: The National Elite Foundation provided employment and residence status data of the first 20,000 entrance exam students in different years to the ministry. This analysis helped ascertain accurate immigration and other elite-related information for the first time.
- Collaboration with the Ministry of Science, Research, and Technology to analyze the labor market of graduates: The Ministry of Science provided the national code of graduates from two universities, which was analyzed by aligning it with social security data.

AI Level of Acceptance and the Level of Trust

One of the primary focuses of the AI FORA research project is the concern about the growing delegation of decision-making authority, particularly in social services, to machines and artificial intelligence. Hence, the level of acceptance and trust in this decision-making approach emerges as a critical parameter in the implementation of such programs. The acceptance and level of trust in artificial intelligence vary among different societies, particularly in cases where AI is involved in social provisions and decision-making processes. While the Targeted Subsidies Plan initially ignored this aspect, at the beginning it was tried to do the decimation based on the self-declaration of the income conditions. However over time, due to the economic pressure of implementing the plan and the devaluation of the national currency, the importance of identifying high-income people increased, and the government sought more accurate, low-error, and scalable methods. Considering Iran's cultural and social context, initial inquiries regarding trust and acceptance levels toward AI-driven decile classification might have yielded unfavorable responses. However, gradual improvements in accuracy, coupled with governmental policies allowing households to contest their income classification, bolstered trust in the utilization of artificial intelligence algorithms for household classification over time.

In Iran, the targeted subsidies plan has been in place for over a decade, and artificial intelligence technology has been widely used in algorithms and eligibility testing. In 2020, the Council of Ministers approved the use of artificial intelligence exclusively for household entitlement assessment, without any human intervention. This AI-driven system makes eligibility decisions based on various data points, including financial status, assets, income, foreign trips, loans, employment status, and retirement.

A digital questionnaire was crafted to gauge the degree of acceptance and trust in artificial intelligence concerning subsidy targeting. The questionnaire's link was disseminated across various social networks, targeting families and individuals aged 18 and above. Participants were prompted to respond to two queries. The questionnaire remained accessible for 1 month, allowing stakeholders to provide feedback on their levels of trust and acceptance regarding artificial intelligence.

Question 1: How acceptable is it to you that AI algorithms make decisions about household eligibility based on various data points?

- Option A: I find it acceptable, and I believe it is more accurate than traditional methods used by experts.
- Option B: I find it acceptable, but I acknowledge that it may have errors similar to traditional methods.
- Option C: I do not find it acceptable, and I believe this task should be handled by experts.

The majority of respondents (87%) chose options A and B, indicating a high level of acceptance of artificial intelligence in the targeted subsidies plan.

Question 2: How much trust do you have in artificial intelligence algorithms regarding household eligibility in the targeted subsidies plan?

- Option A: Very much (I fully believe that human error does not influence household eligibility determination).
- Option B: A lot (I trust AI, recognizing that there may be biases and errors that can be improved over time).
- Option C: Average (not much different from usual methods).
- Option D: Low (I have low trust in AI algorithms due to potential errors and biases associated with technology).
- Option E: Very little (I do not trust AI at all, fearing deliberate data manipulation to the detriment of households).

According to the results, 79% of participants expressed high levels of trust in artificial intelligence, selecting options A and B.

Overall, the survey included 420 participants, and the findings, presented in Figs. 7.12 and 7.13, reveal a significant level of acceptance (87%) and trust (79%) in the use of artificial intelligence for the targeted subsidies plan. This data provides valuable insights into public perception and attitudes toward AI implementation in social policy decisions.

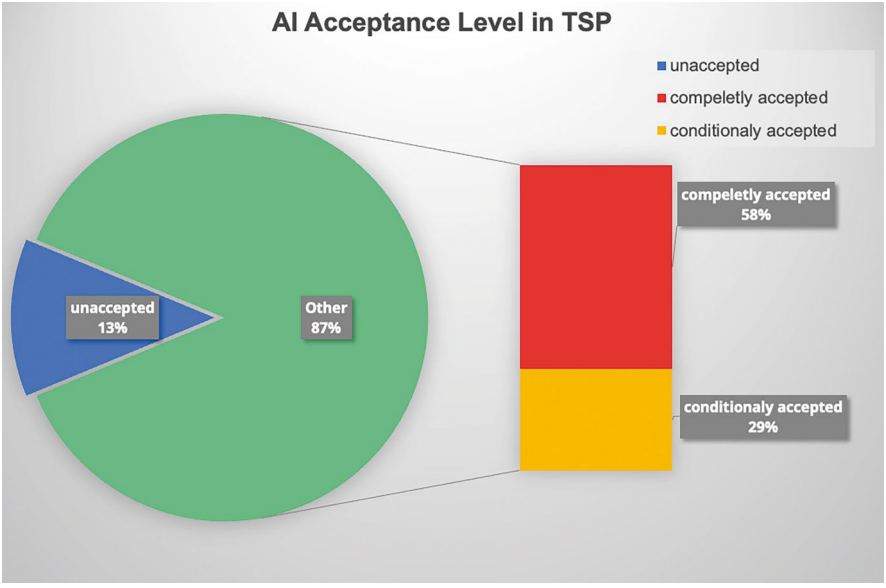

Fig. 7.12 AI acceptance level in the Targeted Subsidies Plan

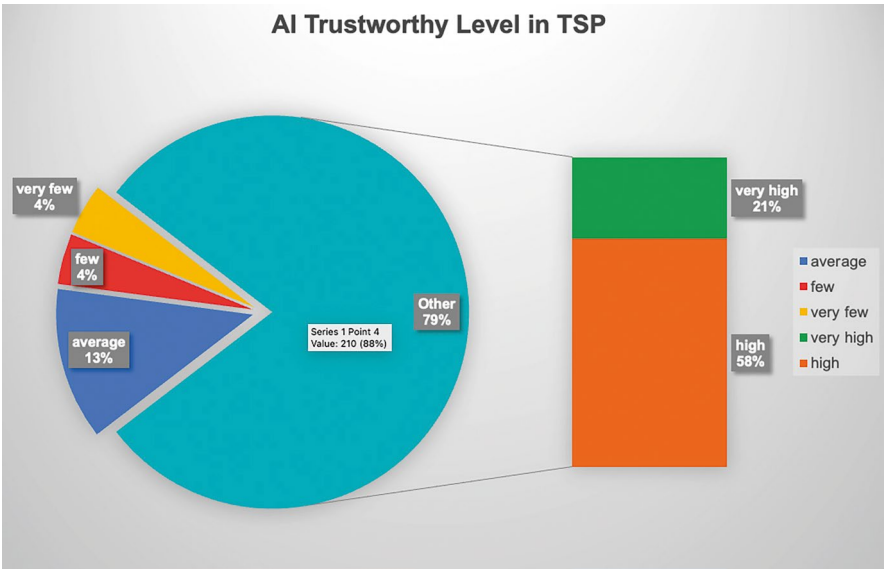

Fig. 7.13 AI trustworthy level in the Targeted Subsidies Plan

Conclusion

The integration of artificial intelligence (AI) into the targeted subsidies plan has been a transformative step in data-driven policy and decision-making in Iran. The establishment of the IWDB in 2014 has proven to be the cornerstone of this plan, enabling various projects such as welfare atlas preparation, identification of eligible households for subsidies, and improvement of distribution mechanisms. The utilization of AI-based data processing and decision-making models has streamlined the identification of deserving recipients and resulted in substantial financial savings, amounting to 36,000 billion Rials (about 76 million dollars with 2023 currency rate) over the last 5 years.

AI's capabilities in problem-solving and data analysis have made it a driving force in the realm of social services. While acceptance and trust in AI's use for household deciles ranking have been generally positive, identifying individuals with AI algorithms, as demonstrated in the "Woman, Life, Freedom" movement, poses significant challenges, raising questions about its application in social services and its impact on acceptance and trust.

The management of the IWDB under the deputy of social welfare in the Ministry of Labor, Cooperation, and Social Welfare has encountered obstacles due to other actors' limited roles in providing timely and continuous information. Transferring IWDB's infrastructure to the Central Bank has been suggested as a solution (based on the interview with experts), given its financial leverage and access to banking network information, highlighting the importance of considering technical aspects when defining the roles of various actors in the network.

Data-driven decision-making has become increasingly crucial, emphasizing the value of data. However, organizational data privacy concerns have hindered regular and up-to-date data provision, necessitating improved legal mechanisms for data access within the targeted subsidies plan. Technical solutions for automatic data collection at specific intervals can enhance data availability and realize informed decision-making.

Despite the targeted subsidies plan's appropriate targeting and AI-based identification of low-income households, it has not led to a reduction in poverty in the country. Economic inflation, intensified international sanctions, and devaluation of the national currency have impacted plan implementation. Addressing socioeconomic challenges requires a holistic approach that combines technological progress with effective governance ideas and political efforts.

In conclusion, the Targeted Subsidies Plan in Iran demonstrates the immense potential of AI in social services, underpinned by the invaluable IWDB. However, to achieve sustainable and effective outcomes, continued efforts are required to address societal acceptance, data access, and comprehensive governance. By striking a balance between technology and policy, AI can continue to play a pivotal role in data-driven decision-making and the advancement of public welfare in Iran. In the second phase of our research, we will construct an agent-based model informed by our insights gained from the analysis of TSP in Iran. Subsequently, we will explore future scenarios involving the implementation of various policy options.

Acknowledgments Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

References

Ahrweiler, P. (2019). *AI for assessment (AI-FORA)*. Volkswagenstiftun.

AI for Assessment. (2020). *TISSS lab*. Retrieved 10 Jan from ai-fora.de

Bakhshoodeh, M. (2013). Proxy means tests for targeting subsidies scheme in Iran. *Iranian Journal of Economic Studies, 2*(2), 25–46.

Creemers, R. (2018). *China's Social Credit System: An evolving practice of control*. Available at SSRN 3175792.

Delazimi, F., & Amraei Mojtaba, E. Z. (2023). *Statistical yearbook of the ministry of cooperation*. Labor and Social Welfare.

Doshmangir, L., Doshmangir, P., Abolhassani, N., Moshiri, E., & Jafari, M. (2015). Effects of targeted subsidies policy on health behavior in Iranian Households: A qualitative study. *Iranian Journal of Public Health, 44*(4), 570.

Furtado, B. A., Sakowski, P. A. M., & Tóvolli, M. H. (2015). *A complexity approach for public policies*.

Hofstede, G. (2024). *Hofstede country comparison tool*. Hofstede-Insights. https://www.hofstede-insights.com/country-comparison-tool

Hoftede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind: Intercultural cooperation and its importance for survival*. McGraw-Hill.

How to implement the law on targeted subsidies in the 1401 budget bill. (2022). IRNA. Retrieved 20 Dec from irna.ir

Kshetri, N. (2020). China's social credit system: Data, algorithms and implications. *IT Professional, 22*(2), 14–18.

MCLS. (2021). *Ministry of Cooperative Labour and Social Welfare, Project Reports*. https://www.mcls.gov.ir/fa/news/246322/اطلس-نقشه-رفاهی-یکصد-شهر-ایران-منتشر-شد

OECD. (2019). *Artificial intelligence in society*. OECD Publishing. https://doi.org/10.1787/eedfee77-en

Valle-Cruz, D., Alejandro Ruvalcaba-Gomez, E., Sandoval-Almazan, R., & Ignacio Criado, J. (2019). A review of artificial intelligence in government and its potential from a public policy perspective. *Proceedings of the 20th annual international conference on digital government research*.

Chapter 8
Social Assessment and Cultural Resistance: The Public Distribution System in Tamil Nadu, India

Sumathi Srinivasalu, Manjubarkavi Selladurai, Shobana Sharma, Gunanithi Perumal, Muniraj Mathaiyan, Ashly Ann Jo, and Ebin Deni Raj

Abstract This chapter describes and analyses the technology-based social assessment process in the Indian Public Distribution System (PDS) in Tamil Nadu and its impact on vulnerable communities. After introducing the history, characteristics, and policies of the PDS, it presents the current actors, stakeholders, and the processes involved, especially those of socially assessing beneficiaries for receiving rations and the role of technology. It investigates the performance, characteristics, gaps and barriers of the existing PDS in the high-performing state Tamil Nadu, the role "cultural resistance" plays in everyday social assessment practices in Fair Price Shops, the effects of the current practices on vulnerable groups, and how these groups envisage improvements for a more desired system. The chapter uses two data collection approaches to address its research questions: A quantitative household analysis focusing on household demographics, ration utilisation, savings due to rationed items, and overall satisfaction with the PDS, and a Safe Spaces workshop with participants from vulnerable communities to gather qualitative data about the role of culture and context in daily assessment practices. The use of biometric data and distribution algorithms can benefit from insights into these socio-cultural dynamics: decision-makers have to take them into account for system improvements which are yet to adapt to the AI framework.

S. Srinivasalu (✉) · M. Selladurai · S. Sharma · G. Perumal · M. Mathaiyan · A. A. Jo
Department of Anthropology, University of Madras, Chepauk, Chennai, India

E. D. Raj
Indian Institute of Information Technology, Kottayam, Kerala, India
e-mail: ebindeniraj@iiitkottayam.ac.in

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*,
Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_8

Introduction: History of the Public Distribution System (PDS) of India

In 2022, India's population was approx. 1.375 billion people, which has grown steadily over the years with an average annual growth rate of 2.2%. They are leading to high pressure on resources and high demand for consumer goods. A gap between availability and demand for goods and services results in the persistence of poverty. Hunger alleviation and poverty eradication are the twin objectives of the Public Distribution System (PDS) in India (Gulati et al., 2012; Mallik et al., 2017; Shivakumara, 2022).

The Public Distribution System (PDS) is intended to provide essential goods and services, mainly food items to everybody, especially "the poorest of the poor" at a reasonable cost contributing to general social welfare (Bhattacharya et al., 2017). The central and state governments share the responsibility to provide food grains to the identified beneficiaries. The central government procures food grains from farmers at a minimum support price (MSP) and sells them to state governments at central issue prices**.**

Transporting food grains from the godowns to each Fair Price Shop (FPS or ration shop) is the responsibility of the central government. The beneficiary buys the food grains at the lower central issue price which is the responsibility of the state government. Many states further subsidise the price of food grains before selling it to the beneficiaries. The effectiveness of the PDS largely depends on adequate policy decisions regarding the operational and organisational aspects of the system**.** In India, subsidies to the poor have been defined as the unrecovered cost of social and economic services delivered by the government. The PDS plays a pivotal role in the food economy and security.

To deal with the supply and distribution of essential commodities, the Food and Price Control Department was established in 1942 under British rule, a dedicated committee organised the "Grow more food" campaign, and all exports were stopped. The Department took control over the procurement of food grains to create a central food reserve and distribution network through the so-called Fair Price Shops (FPS) in deficit areas. In 1948, the FPS network was part of the overall policy of fighting against price rises. Rationing was introduced in the mid-1950s with the opening of more FPS throughout the country to ensure an equitable distribution of essential commodities. Before the 1960s, distribution through the PDS was generally dependent on imports of food grains. It was expanded as a response to the food shortages of the time, and subsequently, the Government of India set up the Agriculture Prices Commission and the Food Corporation of India to improve domestic procurement and storage of food grains for PDS. By the 1970s, PDS had evolved into an all-India scheme for the distribution of subsidised food.

In 1997, the government launched the Targeted Public Distribution System (TPDS) intending to provide subsidised food, and now also fuel, to the poor through the network of ration shops. In 2013, the Parliament enacted the National Food Security Act (NFSA) and made it a legal entitlement to poor households. The Act

illustrated the working of the food supply chain from the farmer to the beneficiary, identified challenges to the implementation of TPDS (Targeted Public Distribution System), and discussed alternatives to reform it. The state-wide variations in the implementation of TPDS discussed changes to the existing system by the Act. At the national level, the poverty line had been estimated periodically, and remedial actions were targeted with alternative approaches. However, at the state level, development promises and policies mostly focused on financial allotment strategies and population groups targeted, rather than on implementation processes, follow-ups, or assessment procedures.

Artificial Intelligence in the Indian Context: Policy Execution with Technology

As part of the digital India programme under TPDS (Targeted Public Distribution System), beneficiaries were divided into two categories: households below the poverty line (BPL) and households above the poverty line (APL)**.** Antyodaya Anna Yojana (AAY) was aimed at making TPDS (Targeted Public Distribution System) the tool to reduce hunger among the poorest segments of the BPL population. Out of maximum coverage of 81.35 crore (crore = 10 mill. equaling 100 lakh in the Indian numbering system) around 80.60 crore persons have been covered under NFSA at present for receiving highly subsidised food grains. This population count is equivalent to the populations of all the major countries of Europe put together.

The identification of beneficiaries by states and Union Territories (UT) is a continuous and complex process. Presently, the use of AI technology involves citizen identification for the exclusion of ineligible/fake/duplicate ration cards and updates due to death, migration, etc. It also involves new additions due to birth, new definitions of "family", and new policies regarding household models. The inclusion of more formerly left-out households was also addressed (Standing Committee on Food, Consumer Affairs and Public Distribution, 2021).

Tamil Nadu State Government is currently a pioneer for an improved PDS based on data analytics and AI (George & McKay, 2019; Vinayagamoorthi & Uma, 2014). In June 2019, the Central Government of India launched a "One-Nation-One-Ration Card Scheme" that was taken up by the State Government of Tamil Nadu as an "Intra-State Probability Scheme" to be implemented in the districts of Tirunelveli and Tuticorin from February 2020 onward on a pilot basis. The COVID-19 crisis brought forth the necessity of holding it in abeyance as the situation was that the ration cards were only valid in their local Fair Price Shops and many of the migrants were left stranded mid-way; this traumatic experience pushed the implementation of the ONORC scheme for the future.

The PDS actors and their interactions can be mapped out from two perspectives: the procurement and the distribution process. In the procurement and storage process, five main groups of actors are involved: farmers, FCI/state government as

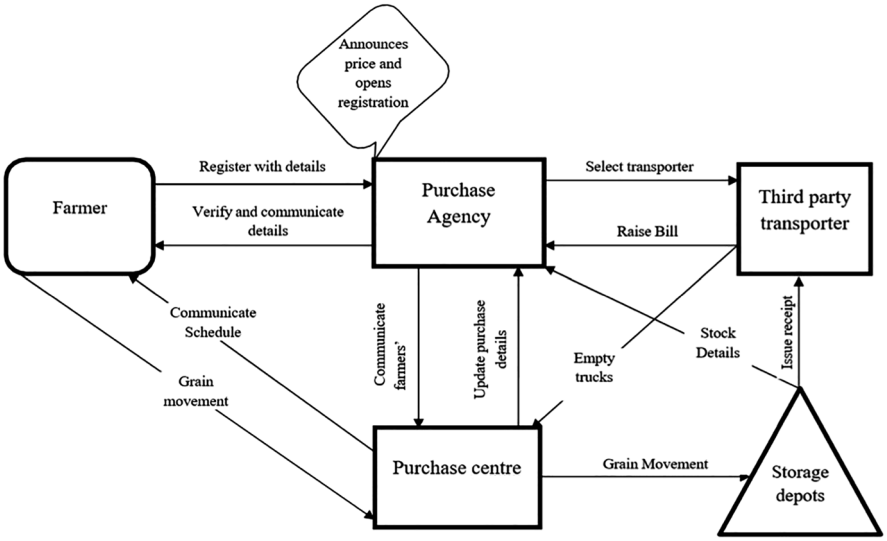

Fig. 8.1 PDS procurement network

purchase agencies, purchase centres, third-party transporters, and storage depots. Figure 8.1 depicts the grain procurement process, which can serve as an example for supply networks.

In the distribution process, nine main groups of actors can be identified: The Ministry of Consumer Affairs, Food and Public Distribution (MCAFPD), the State Government Food and Civil Supply Department (SFACSD), the District Food and Civil Supply Department (DFACSD), the FCI Head Office (FCIHO), the FCI Regional Office/sales office (FCIRO), the FCI Division/District Office (FCIDO), again the storage depots, the Fair Price Shops (FPS), and the cardholders (Fig. 8.2).

At the national government level, the Ministry of Consumer Affairs, Food and Public Distribution (MCAFPD) is responsible for the state-wise allocation of goods based on their assessment of updates on cards received from the state governments. To illustrate the size of funds: The Ministry received the second-highest budgetary allocation among all the ministries in 2019–2020. The major responsibility of it was spreading awareness among consumers about their rights, protecting their interests, implementing standards, and preventing black marketing. The focus is on issues related to consumers. Table 8.1 shows an overview of fund allocation to the Ministry to illustrate size:

The state government's District Food and Civil Supply Department (DFACSD) allocates cards and stocks to their Fair Price Shops. They also issue cards to the cardholders who can then present them at the FPS and receive their goods for free or against subsidised payment.

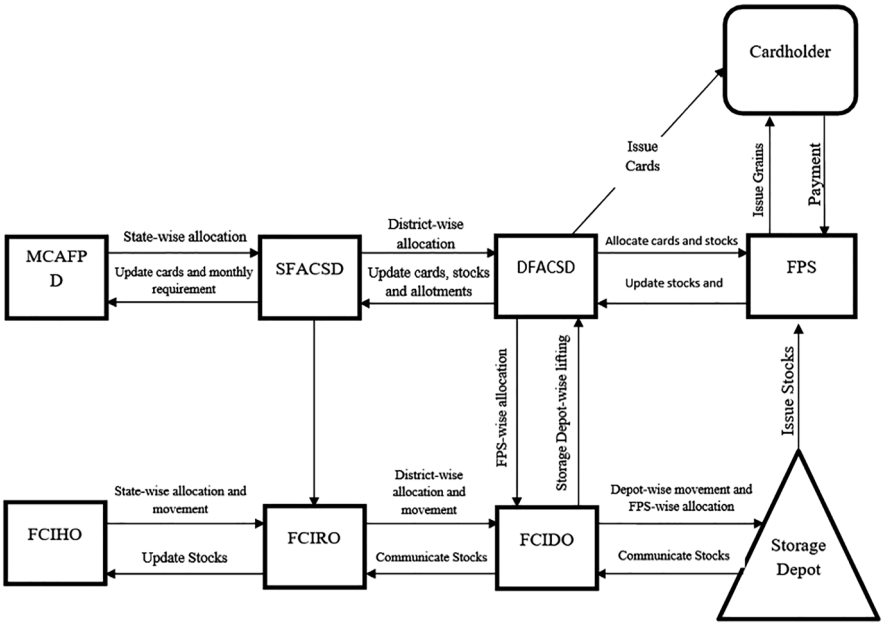

Fig. 8.2 PDS distribution network

Table 8.1 Fund allocation by Govt of India

Departments	2017–18 Actual	2018–19 Revised	2019–20 Budgeted	% change in 2019–20 over 2018–19
Food and public distribution	1,05,865 crore/USD 12,76,90,12,840	1,77,874 USD 21,45,44,50,384	1,92,240 USD 23,18,72,19,840	8.1%
Consumer affairs	3713 USD 44,78,47,208	1782 USD 2,14,93,771	2272 USD 27,40,39,552	27.5%
Total	1,09,578	1,79,655	1,94,513	8.3%

Sources: Expenditure Budget, Union Budget 2019–20; PRS

Inception of Digitisation and Technology with a Cultural Context

The PDS was a manual and labour-intensive dispensation using paper-based ledgers and ration cards with handwritten entries. The emergence of office automation by the government moved the whole process to digital and automated systems that used smart cards and mobile technology. The technological transformation was aimed at producing more efficiency and providing new options for social welfare, but, at the same time, implied the risk of re-establishing existing inequalities dignified by the allegedly "unbiased" machine. Hence, the policymakers introduced a

new identification system which linked the family card to the central citizen management system. This new AI-driven process came to be known as the Aadhaar card.

AADHAAR Card

Under the new Digitisation policy of the Indian government using AI technology in 2009, Aadhaar cards were introduced for all the citizens of the country. Aadhaar is a 12-digit unique identity for every Indian individual, including children and infants, enables identification for every resident Indian, and establishes the uniqueness of every individual based on demographic and biometric information. In 2009, the Unique Identification Authority of India (UIDAI) was set up to implement Aadhaar cards.

The Family Card

A Family Card is the official document issued by the respective state governments. The card enables eligible households to buy food grains at subsidised rates under the NFSA (National Food Security Act). The introduction of digital technologies forced many states in India to introduce online portals for Family Card application, renewal, and management. The card serves as proof of identity, residence, and eligibility for receiving government benefits. It categorises households into different categories based on their economic status, such as Below the Poverty Line (BPL) or Above the Poverty Line (APL), where each category is entitled to a specific quantity of subsidised food grains and commodities. In Tamil Nadu, five different cards are currently in use: Priority Household Cards (PHH), PHHA Cards (Antodaya Anna Yojana, AAY scheme) for top priority cases, and three types of non-priority household cards (NPHH)—non-priority household cards meaning only sugar as commodity (NPHH-S) and meaning no commodity at all (NPHH-NC). The majority of the rural and half of the urban population are entitled to receive highly subsidised food grains under two categories of beneficiaries—Antodaya Anna Yojana (AAY) households (PHHA) and Priority Households (PHH).

The social assessment using technology takes place in the Fair Price Shops where cardholders present their cards and biometric information is evaluated by the electronic card readers for the allocation of goods and services. The system runs with the support of two portals—Integrated Management of Public Distribution System (IM-PDS) and Annavitran, which host all the relevant data. The data management systems collate the data provided by the cardholder, and this includes demographic details including household categorisations. There are provisions for disclosure of records related to these operations. Beneficiaries' lists are placed in the public domain/portals for enhanced transparency by the government at the centre.

The PDS policy was one of the successful social welfare policies implemented by the central government as it could reach the last mile person in any part of the

Fig. 8.3 Model of smart ration card issued by Tamil Nadu civil supplies and consumer protection department, CC BY-SA 3.0 Source: https://commons.wikimedia.org/wiki/File:Tamil_nadu_smart_ration_card_specimen.jpg (accessed 17/11/2024)

country. The digitisation process became a highly populist campaign as the ease of use by the end-user was extremely simple and gave them a sense of empowerment. The central and the state governments who are equal stakeholders in the process use this as a political tool.

The example below shows a man as the head of the family. However, complying with gender equality requirements, the eldest woman of a beneficiary household (18 years or above) is also considered as "Head of Family" to issue cards. In general, there are provisions for women's empowerment such as grievance redressal mechanisms through State Food Commissions, Vigilance Committees, and other organisations at different levels (Ghabru et al., 2017). However, by 2022, the integration of Aadhaar cards with Family Cards will significantly reduce pilferage resulting from duplication (Fig. 8.3).

The Relevance of Culture in the PDS

Indian society is highly stratified by caste structure. For the chapter, we choose to use the term "community" instead of "caste", because the vulnerable social groups refer to and identify themselves as "communities" rather than as "castes" since the term "caste" is a European import. A central cultural aspect is the acute awareness among individuals of their position on the social hierarchy, as well as their ability to discern whether others may rank higher or lower. Social assessment in India works with this type of stratification based on the affirmative action policy. In various everyday situations, individuals of dominant communities are accorded precedence, and this informal "priority ranking" is ingrained in the minds of all, spanning from

ration shop attendants to customers. For instance, the dominant community individuals receive prompt information about the arrival of goods at the ration shop and are offered superior quality items in terms of freshness or source. This was evidenced sharply in the case study results (section "Household Analysis of PDS Performance Using Technology").

The Affirmative Action Policy of India operates nationally, categorising individuals into four groups: upper caste/others, comprising approximately 2–3% of the population; other backward castes (OBC), representing about 75% of the population and dominant castes on the social status hierarchy; scheduled castes (SC), constituting around 20% of the population; and scheduled tribes (ST), making up about 2% of the total population. The latter two categories are particularly recognised as vulnerable groups.

The social assessment scenario for vulnerable individuals, particularly those belonging to Scheduled Castes (SC), at Fair Price Shops for accessing social services is characterised by contradiction. According to affirmative action principles, they are deemed eligible for preferential access to goods and services. However, they encounter "cultural resistance" in the form of negative discrimination based on traditional stratification and segregation within society, where they are relegated to a "low priority" status. This clash between positive and negative discrimination undermines efforts towards fairness upheld by the welfare state. Moreover, it poses challenges for utilising AI technology in goods allocation, as the typical variables employed for positive discrimination are nullified.

The Indian social fabric can be characterised by dual aspects: the political aspect being, a democratic, welfare-oriented nation on the one hand, and the cultural aspect where deeply stratified social structures and inter-community relationships are based on traditional hierarchical patterns. This duality manifests itself in high inequalities. The State of Tamil Nadu has experienced significant polarisation since the 1960s, stemming from the self-respect movement led by the Dravidian parties. This movement impacted the categorisation of communities and castes for political purposes, resulting in a highly polarised society with distinct divisions between dominant and subjugated groups.

Household Analysis of PDS Performance Using Technology

Members of the vulnerable communities, stakeholders of the FPS administration, key informants of non-governmental organisations, and political stakeholders were interviewed with a questionnaire about using the technology. The questionnaire to the stakeholders was broad-based in the following contexts.

1. The knowledge and awareness created by the technology.
2. Its impact on the marginalised communities and their worldview.
3. Success/drawbacks in AI-based programme implementation.
4. Equal rights to welfare measures have been addressed/not addressed.

5. Benefits due to this distribution method and any further improvements be identified.
6. The opinion of the community members on the impact of entitlement aspects since it bridges social inequalities and affects the success of the AI programme.

A standardised household questionnaire was developed and carried out as a door-to-door inquiry by trained social researchers. In Tamil Nadu, there are 39 districts. The PDS is used by 70,053649 beneficiaries of which 69,652,487 have Aadhaar entries wherein they have linked their family card and Aadhaar card, and 22,393,868 have connected their mobile phones to the family cards and records of the same are collated by the central server. Villages, and households in villages, were chosen by a random sampling technique. The household analysis, derived from a comprehensive survey dataset, scrutinises the performance of the PDS across various dimensions, focusing on household demographics, ration utilisation, savings due to rationed items, and overall satisfaction with the PDS.

PDS ration shops in local districts of different states were visited to understand processes after participatory observation. To find out about the popularity of PDS and how the system has been used by policy, a content analysis of two Indian newspapers was conducted during 2020–2024. The *Dinamalar* is the Tamil Nadu daily newspaper in Tamil language, and the *Times of India* is an English-language national daily newspaper. A corpus-based approach to frequency analysis systematically extracted action words as search terms and analysed language use within the linguistic context of the given items. The analysis showed that the PDS is a discussion point every day in the newspapers (Economic and Political Weekly, 2023). Reporting words such as "corruption," "technology", "smart card", and "quality" had high frequencies in both newspapers. News reports frequently employ speech reporting, maintaining an objective stance toward the events being reported. However, they also utilise speech act reporting to introduce reported speech, thereby imbuing the narrative with subjectivity. There is a tendency for these reports to incorporate mental reporting more extensively to convey the opinions and attitudes of common people or authorities, especially in the Indian context.

The analysis is based on survey data encompassing responses from 2020 households across multiple villages. The survey includes detailed questions about household demographics, the quantities of rationed items received, monthly expenditures on food, and savings attributed to PDS. Statistical methods and visual data representation techniques were employed to analyse and interpret the survey data. The survey data presents a comprehensive overview of the demographic and economic landscape of households participating in the PDS, revealing significant insights into their socio-economic status and the impact of PDS on their daily lives. The variables looked at were the age of respondents with an average age of approximately 39.7 years, indicating a broad demographic engagement with the PDS. On average, households report about 1.5 earners, highlighting the multi-contributor model prevalent in these communities. The financial aspects covered in the survey show an average monthly household income of ₹12,191, which varies significantly across the spectrum. This variation underscores the economic diversity among the

beneficiaries of the PDS. The expenditure on food consumption averages ₹3914 monthly, with some households spending as much as ₹40,000, reflecting the varying degrees of financial commitment towards nutrition. The average household size is reported to be about 3.5 individuals, suggesting a mix of nuclear and joint family setups among the respondents (Fig. 8.4).

In terms of rationed food items, households receive an average of 22 kg of rice and approximately 1 kg of toor dal, which are staples in the Indian diet. The data also reveals that households manage to save an average of ₹880 monthly through ration items, with savings extending up to ₹15,000 for some, showcasing the significant relief the PDS provides in their food expenditure. The statistics of the data are provided in Table 8.2:

This bar plot compares key financial statistics, including average monthly income, average monthly expenditure on food, and average savings with ration items. It illustrates the disparity between income and expenditure, highlighting the relatively high average income compared to the lower expenditure and savings, underscoring potential areas for financial optimisation and budgeting strategies. The majority of the respondents belong to the Most Backward Class (MBC) and Backward Class (BC), with a significant representation of Scheduled Tribes (ST). This demographic distribution underscores the PDS's role in supporting socially and economically marginalised communities. Moreover, the data revealed a predominance of female respondents, highlighting the critical role of women in household food management and their interaction with the PDS (Fig. 8.5).

The demographic overview of respondents highlights family types, gender, marital status, and community distribution. The survey indicated that nuclear families

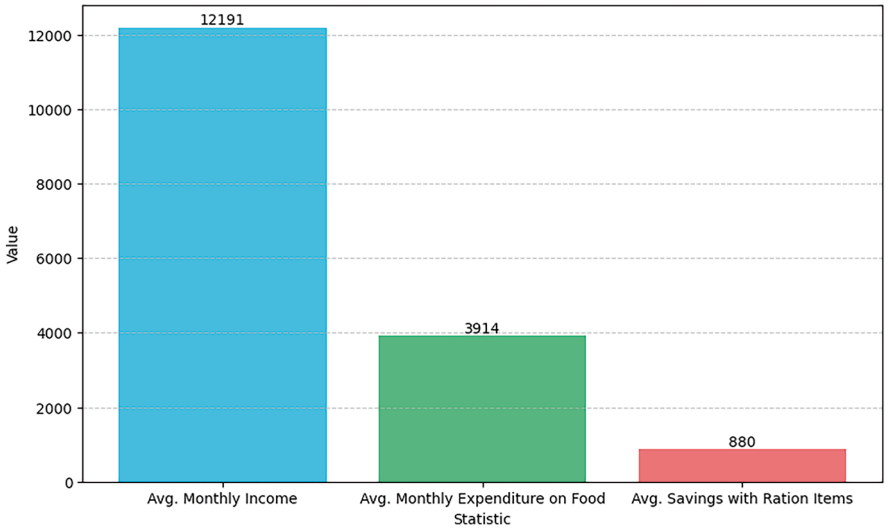

Fig. 8.4 Key financial statistics comparison table of key variables. Source: results from household survey

Table 8.2 Data on income and age correlation

Variables	Description
Age range	12–83 years
Average age	39.7 years
Average number of earning persons	1.5 earners per household
Average monthly income	₹12,191
Income range	₹0 to ₹75,000
Average monthly expenditure on food	₹3914
Expenditure range on food	Up to ₹40,000
Average household size	3.5 individuals
Average quantity of rice obtained	22 kg per household
Average quantity of toor dal obtained	1 kg per household
Average savings with ration items	₹880 per month
Maximum savings with ration items	Up to ₹15,000

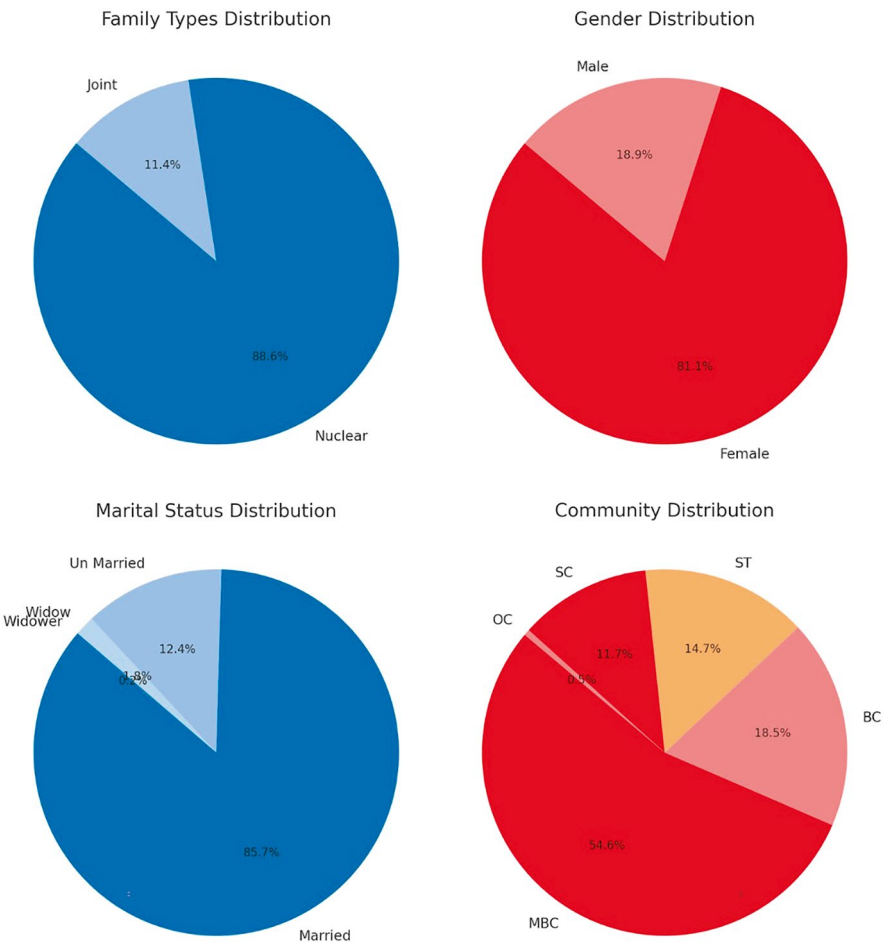

Fig. 8.5 Demographic overview of respondents

are the predominant household type among the respondents which is an evolving change from the joint family system that existed even during the 1970s and 1980s. The average quantity of rice obtained per household was approximately 22 kg, with other essential items like wheat, sugar, and kerosene also being accessed through the PDS. These savings play a pivotal role in enhancing the food security of beneficiary households.

Figure 8.6 compares total household income, monthly expenditure on food, and savings attributed to ration items across varying household sizes. It reveals insights into how household size influences financial dynamics, showcasing potential correlations between size, income, expenditure, and savings. The data suggests that larger households may exhibit higher expenses and lower savings relative to smaller ones, prompting further investigation into the intricate interplay between household size and financial behaviours.

The analysis underscores the necessity for policy reforms in the Public Distribution System as its role is pivotal in ensuring food security across diverse household demographics. Furthermore, the changing Indian household landscape, with a marked increase in nuclear families, calls for an optimisation of ration allocations to better meet their specific nutritional requirements. This adjustment would cater not only to the quantity but also to the variety of food items distributed, ensuring a balanced diet for smaller family units.

Additionally, the analysis highlights the importance of targeted community support, especially for communities identified as most vulnerable, such as the Most Backward Classes (MBC), Backward Classes (BC), and Scheduled castes and Tribes (SC&ST). Tailored interventions for these groups could enhance the equity

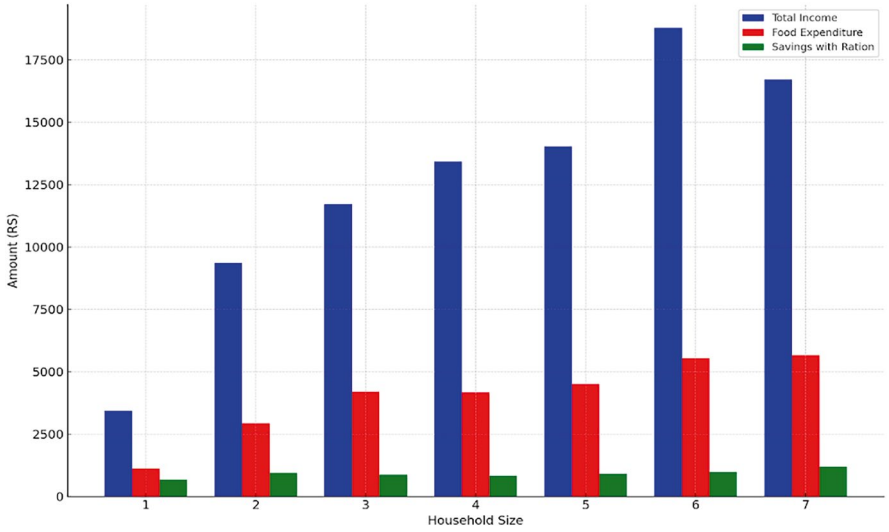

Fig. 8.6 Comparison between total income, food expenditure, and savings with ration items across different household sizes

and effectiveness of the PDS. There is a pressing need for strategies aimed at improving savings through the PDS for all households, particularly those at the lower end of the income spectrum. The household analysis of PDS performance provides valuable insights into its role in ensuring food security among marginalised communities. While the PDS demonstrates a significant positive impact, there is a need for continuous improvement to address the diverse needs of its beneficiaries more effectively. Tailored policies that consider beneficiary households' demographic and socio-economic dynamics can enhance the PDS's role as a critical safety net in India's social welfare landscape.

Safe Spaces Workshop with Vulnerable Communities

In February 2023, a stakeholder workshop convened at Saccidananda Ashram Shantivanam in Tamil Nadu was established as a Safe Space for vulnerable communities to freely express their opinions on PDS gaps, barriers, and fairness issues without fear of reprisal. Shantivanam has a longstanding commitment to community development in neighbouring villages, with the pandemic crisis fostering closer ties between vulnerable groups and safe spaces.

The structure of the workshop was a free flow of conversation with minimal guidance from the moderators who intervened only when the checklist details were not adhered to. The materials used in the workshop were as per the instruments available in the FPS. The agenda was constructed in a way that it was for gathering information rather than as a controlled discussion as in the European context.

The majority of workshop participants hailed from the Scheduled Caste (SC) background, comprising 18 women, while 7 represented the Most Backward Class (MBC). The SC community, characterised by a submissive demeanour, contrasted with the assertive and dominant traits of the MBC. During the workshop, which was conducted in the local Tamil language, this focus group engaged in discussions on the PDS structure and operations based on their personal experiences. Topics included biometrics and the functionality of family cards (smart cards). The workshop incorporated elements of gamification through role-play

7. At Fair Price Shops
8. An on-site visit to a local Fair Price Shop at Thannirpalli

The moderated focus group discussions were recorded with the informed consent of participants, transcribed and translated.

Participants were enthusiastic about the AI technology in PDS schemes, viewing it as a tool to enhance management efficiency. All participants held Family Cards and had regular access to Fair Price Shops, which operated at least 6 days a week, serving both urban and rural areas. Monthly ration distributions enabled families to plan their food budgets accordingly. Community cooking preferences favoured fresh/raw ingredients and items sourced from wet markets.

Workshop discussions primarily focused on quantity and quality issues, with common complaints regarding the poor quality of rice, a staple food in Tamil Nadu. Concerns were raised about faulty weighing scales and under-portioning, with beneficiaries emphasising entitlements as necessities rather than mere handouts. The workshop shed light on the undue priority given to certain castes due to their purchasing power, exacerbating disparities in access to essential items. The quantity given to the household is insufficient as it lasts for only 20 days and the shortage has been made up from other sources, which is highly expensive.

While biometric data usage was seen as a deterrent to fraud and benefit losses, challenges arose from malfunctioning biometric machines, particularly for beneficiaries with worn-out thumbprints due to manual labour. Instances of illegal PDS goods smuggling and diversion of subsidised items to the black market by ration shop workers were reported regularly. The mock role-play session, conducted as a form of engagement, aimed to address objections raised by participants regarding the term "gamification", which carried strong connotations related to survival and prosperity, particularly concerning food items in their cultural context. This perception which has strong cultural connotations is diametrically opposite to the European culture where gaming influences social scenarios. The analysis of workshop contributions reinforced the notion that participants possessed a profound understanding of technology and its advantageous impact on their lifestyle, effectively mitigating prevailing socio-economic and politico-cultural biases. Of particular interest was the widespread recognition of technology's benefits. However, despite this awareness, existing categorisations have failed to eradicate discrimination, which continues to persist as a stark reality.

The workshop analysis also uncovered that the Scheduled Tribe communities, lacking functional ties with caste groups and receiving goods from the Fair Price Shops, are largely unaware of the purported discrimination. This lack of awareness stems from their homogeneity and relatively isolated habitats, situated far from mainstream communities. Rag pickers, destitute individuals, migrant workers, the disabled, and the homeless, who lack familial or household ties, are issued Family Cards as part of the social inclusion process, without distinction from regular Family Cards. However, traditional cultural norms do not apply to these individuals, as they lack a collective identity based on caste or community. Consequently, their priority for receiving goods from Fair Price Shops is minimal due to their lack of housing infrastructure, leaving them with no space for storage or cooking. Some resort to purchasing these goods and exchanging them for prepared meals at private shops.

The findings substantiated several challenges within the current system. The PDS is plagued by numerous shortcomings, primarily attributed to the human interface, where corruption thrives. These issues encompass disparities in the quantity and quality of goods, extensive waiting times, sluggish processing speeds, irregularities in food grain procurement, diversion of goods before reaching Fair Price Shops, biased selection of warehouses for storage, and incidents of material theft. Furthermore, challenges persist with outdated biometric machines and firmware compatibility issues, leading to undue delays in cardholder identification. This primarily affects daily wage labourers, who, frustrated by the prolonged process, often

encounter friction with FPS attendants. Another issue predominantly faced by women, particularly those engaged in agricultural labour, is the difficulty in providing thumb impressions during biometric verification. The manual labour they undertake often renders their fingerprints invisible under infrared lights, making the verification process laborious and time-consuming.

Conclusion

While national-level policy decisions aim to foster inclusivity and empower the underprivileged, the execution by implementing agencies and stakeholders often reflects their adherence to parochial customs. Therefore, it is imperative to consider cultural connotations when formulating policies, ensuring they resonate effectively with diverse communities.

The analysis of PDS performance reveals key insights into its role in ensuring food security among marginalised communities in India. Emphasising the importance of tailoring PDS services to meet the needs of the communities, we need to acknowledge female respondents play a critical role in household food management and interact significantly with the PDS. The analysis also suggests correlations between household size and financial dynamics, with larger households exhibiting higher expenses and lower savings relative to smaller ones. Policy recommendations include enhancing gender sensitivity within the PDS, optimising ration allocations for smaller family units, implementing tailored interventions for vulnerable communities, and developing strategies to improve savings, particularly for low-income households. Overall, these insights underscore the need for continuous improvement in the PDS.

Culture plays a pivotal role in influencing consumption patterns, social norms, and community dynamics, all of which directly impact the functioning of the PDS. Additionally, diverse regional contexts, including economic disparities, geographical challenges, and demographic variations, further complicate the distribution of public services. Culture profoundly influences food preferences, dietary habits, and traditional practices across India's diverse communities. Regional cuisines, religious dietary restrictions, and cultural celebrations contribute to unique consumption patterns that must be considered in PDS implementation. Moreover, social hierarchies and caste dynamics deeply ingrained in Indian society influence access to and treatment within the PDS. Marginalised communities often face discrimination and exclusion when accessing public services, including the PDS. Addressing these systemic inequalities requires a nuanced understanding of cultural norms and power dynamics embedded within local contexts.

Given the complexity of cultural and contextual factors shaping the PDS, the adoption of contextualised, participatory, responsive, and dynamic artificial intelligence (AI) for social assessment becomes crucial. AI systems should be tailored to specific cultural, linguistic, and socio-economic contexts, ensuring relevance and effectiveness in diverse settings. By incorporating local knowledge and insights,

contextualised AI can better understand and address the unique challenges faced across India. Participatory AI involves engaging stakeholders, including beneficiaries, local leaders, and government officials, in the design and implementation of AI-driven solutions. This participatory approach ensures that AI technologies are responsive to the needs and priorities of the communities they serve, fostering ownership and sustainability. Furthermore, responsive AI systems adapt and evolve in real time based on feedback and changing circumstances, enabling agile and effective responses to emerging challenges. In the dynamic context of the PDS, where factors such as market fluctuations, climate variability, and policy changes constantly impact service delivery, responsive AI can facilitate timely and informed decision-making.

In conclusion, the impact of culture and context on the Indian PDS underscores the need for contextualised, participatory, responsive, and dynamic AI for social assessment in public service distribution. By harnessing AI technologies sensitive to local realities and inclusive of diverse voices, India can enhance the effectiveness, efficiency, and equity of its public service delivery systems, ultimately advancing social welfare and inclusive development which can be replicated in any of the European nations.

From the survey data presented, a community-based vulnerability prediction model tailored to the Indian context can be developed. The model will incorporate principles of fairness, accountability, and transparency to ensure equitable outcomes. It will also explore the pivotal role of deep learning in the development of community-specific vulnerability prediction models in the Indian scenario which can be used in the future for other countries. This will allow for the creation of highly accurate models, revolutionising the approach to community-based interventions. The information from the vulnerability prediction model will be used for prototyping AI social assessment systems, which involves creating a synthetic population, generating synthetic outcomes based on improved assessment rules, and training a neural network, thus evaluating the effectiveness of the AI system. Results will be communicated to policymakers and technology providers for "better AI".

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

References

Bhattacharya, S., Falcao, V. L., & Puri, R. (2017). The public distribution system in India: Policy evolution and program delivery trends. In S. Bhattacharya, V. L. Falcao, & R. Puri (Eds.), *The 1.5 billion people question: Food, vouchers, or cash transfers?* (pp. 43–105). The World Bank. https://doi.org/10.1596/978-1-4648-1087-9_ch2

George, N. A., & McKay, F. H. (2019). The public distribution system and food security in India. *International Journal of Environmental Research and Public Health, 16*. https://doi.org/10.3390/ijerph16173221

Ghabru, M., Devi, G., & Rathod, N. (2017). Public distribution system in India: Key issues and challenges. *Indian Journal of Economics and Development, 13*, 747. https://doi.org/10.5958/2322-0430.2017.00240.2

Gulati, A., Gujral, J., & Nandakumar, T. (2012). *National food security bill: Challenges & options*. https://prsindia.org/files/bills_acts/bills_parliament/2011/CACP_Report_on_Food_Security_Bill.pdf

Mallik, R., Chaudhury, S. K., & Sarkar, S. (2017). *Public distribution system - A strategy for food security of India*. S.K. Book Agency.

Seventeenth Lok Sabha, M. of C. A., Food and Public Distribution. (2021). *Twelfth report.* Standing Committee on Food, Consumer Affairs and Public Distribution. https://loksabhadocs.nic.in/lsscommittee/Consumer%20Affairs,%20Food%20and%20Public%20Distribution/17_Food_Consumer_Affairs_AndPublicDistribution_21.pdf

Shivakumara, B. S. (2022). Public distribution system in India – An overview. *African Journal of Accounting, Auditing and Finance, 8*(1).

Trust, S. (2023). *Economic and Political Weekly*, *58*(36). https://www.epw.in/journal/2023/36

Vinayagamoorthi, G., & Uma, K. (2014). Public distribution system in Tamil Nadu. *Shanlax International Journal of Commerce, 2*(2), 64–68.

Chapter 9
The Role of AI in Effective Social Protection Delivery: A Focus on National Cash Transfer Program

Emmanuel Ejim-Eze

Abstract Nigerian Social Protection coverage is limited by booming population, political economy, sociocultural structure, corruption, infrastructure, etc. This makes the Nigerian case unique and highly germane to the overall focus of AI FORA. The Nigerian state is the unit of analysis in this study. This study obtained data from interviews and focus group discussions with SP experts who provided the author with a deep understanding of context dependencies of actors within the Nigerian SP environment. The research strategy adopted made room for understanding social, economic, political, and environmental pressures that drove the discourse on social assessment of technology use in SP in Nigeria. Several works of literature were reviewed and used as sources of secondary data on supply and demand sides of Nigerian SP. Triangulation of data helped to create robust data for this study. Findings from this study show that AI and related technologies were useful in producing high-resolution poverty maps for both Nigerian urban and rural areas, improving the targeting of social safety net program (SSNEP) beneficiaries. General impact assessment dominates genuine social assessment of SP and limits development of robust SP interventions in Nigeria. The future use of AI should extend the technology beyond targeting beneficiaries to include other parts of the SSNEP delivery chain.

E. Ejim-Eze (✉)
Department of Science Policy and Innovation Studies, National Centre for Technology Management, Head Quarters, Obafemi Awolowo University, Ile-Ife, Nigeria

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_9

Introduction

The very essence of social protection (SP) is to shrink or prevent abject poverty and its vulnerability throughout the human life cycle (OECD, 2019). This conforms to the definition of SP[1] by the United Nations and United Nations Economic Network. SP can be in the form of cash transfers, health insurance schemes, social security and social protection floors, pensions, disability benefits, and income security for children and their families; they also include labor market policies such as maternity, paternity, and parental leave (OECD, 2019; ILO, 2018). However, this chapter adopts the definition of Berrou et al. (2021), which states that SP includes all the collective mechanisms and provisions that allow a society to protect itself against the effects of social risks such as sickness, old age, disability, unemployment, and social exclusion.

SP is also viewed as part of human rights and contributes to social justice and economic development. Its realization is key for the attainment of 2030 Sustainable Development Agenda, notably SDG targets 1.3, moving toward universal social protection systems, including floors: leaving no one behind. However, national expenditures on SP for children are very low, equating to only 1.1% of GDP on average, compared to 7% of GDP spent on pensions (ILO, 2021). The rights of children to SP seem not to catch serious attention. The paradox here is that the regions of the world with the largest share of children in the population and the greatest need for SP unfortunately have some of the lowest coverage and expenditure rates for children. A good example is in the sub-Saharan Africa (SSA) with about 0.4% GDP expenditure on children. In Nigeria for instance, whereas 58.7% of adults were poor in 2021, and 67.5% of children under the age of 18, and fully 70.1% of children under 5 were poor (NBS, 2022). Also, SP coverage remains low with significant gender gaps in the African region; only 3.9% of women enjoy comprehensive legal coverage compared to 10.8% of men, a reflection of vast informal labor markets with women concentrated in the most vulnerable forms of informal employment (UN Women, 2022). Breakdown of social assistance budgets and fiscal spending on SP by international organizations such as the UNDP and World Bank revealed that less than 2% of national GDP is allocated to SP in some African states. Unfortunately, it was also noted that states with higher income commonly expend fewer resources on SP (UNDP, 2019; World Bank, 2019). Crude oil-rich countries in SSA such as Angola, Gabon, and Nigeria are known to spend less on SP (Devereux, 2019; Shadare, 2019). It is, however, unfortunate that Nigerians whether as children, as women, or as men are highly vulnerable to poverty as the country is

[1] SP refers to the set of policies and programs aimed at preventing or protecting all people against poverty, vulnerability, and social exclusion throughout their life cycles, with a particular emphasis on vulnerable groups—https://www.un.org/sites/un2.un.org/files/2021/04/a-tb_on_social_protection.pdf

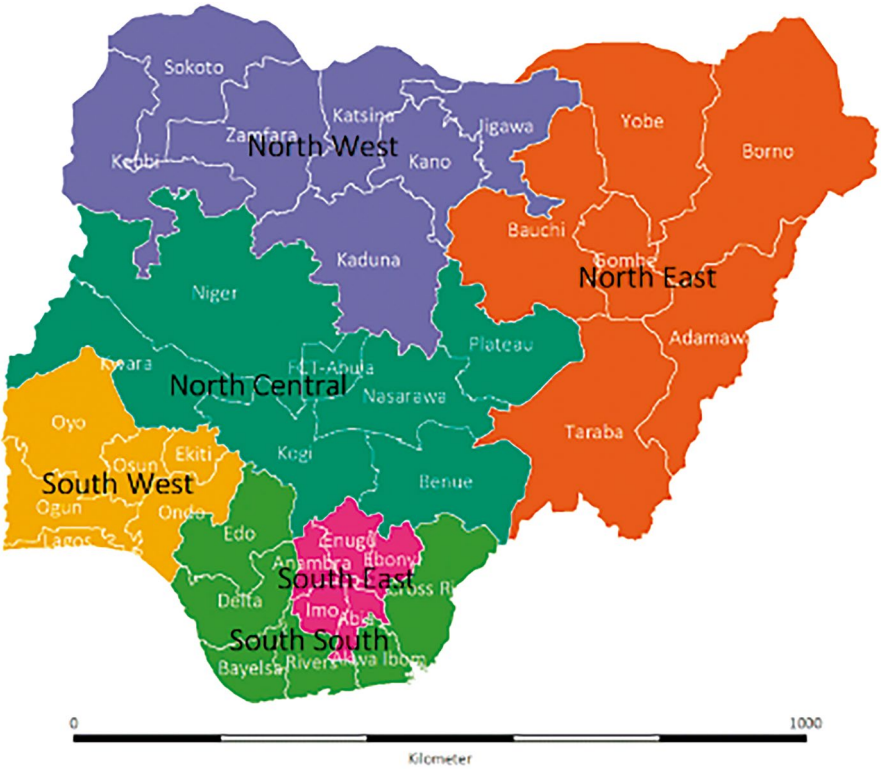

Fig. 9.1 Map of Nigeria showing boundaries of six geopolitical zones, 36 states and the Federal Capital Territory (FCT—Abuja). Source: Map was obtained from Wong et al. (2018)

the most populous (about 228,148,359 in 2024)[2] in Africa with poor governance structure to re-distribute the national commonwealth from the huge endowed natural resources. Figure 9.1 shows the geopolitical zones and states in Nigeria and the poverty and vulnerability map of Nigeria (Fig. 9.2).

Findings from the previous three waves of General Household Surveys (2011–2016) buttress the fact about the high degree of susceptibility of Nigerian population to poverty at varying points. Nigeria is currently the country with the highest number of extremely poor people in the world (90 million),[3] and is unable to provide social protection coverage for the majority of the poor.

According to 2018 figures of the World Bank, Nigeria had 86.9 millions of its population living in extreme poverty—the largest globally (World Bank, 2018).

This number has increased with high inflation in 2024. Nigeria had overtaken India and assumed the position of the poverty capital of the world (Odey & Sambe,

[2] As of Friday, April 26, 2024—https://www.worldometers.info/world-population/nigeria-population/

[3] Nigeria overtakes India in extreme poverty ranking, CNN (June 26, 2018).

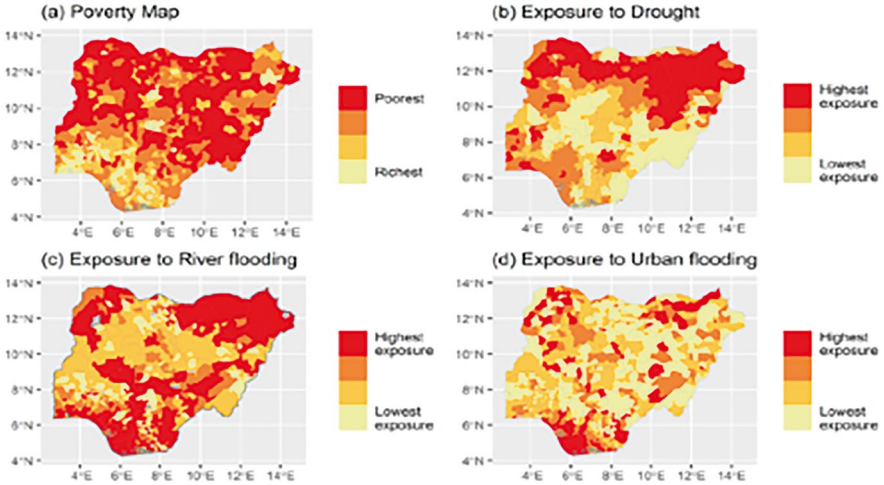

Fig. 9.2 Poverty and vulnerability map of Nigeria. Source: Chi, Fang, Chatterjee, and Blumenstock (2021) and 2018 DHS for wealth estimates; Sutanudjaja et al. (2018) for drought exposure data; ThinkHazard! Database (ThinkHazard! 2021), based on the output of the UF-GLOBAL-FATHOM Global flood model, for flooding data; and World Bank estimates as cited in World Bank (2022)

2019). 40% of Nigerians were extremely poor prior to the COVID-19 pandemic; the pandemic has also led to more poverty with the country feeling more of the economic impact of the pandemic (Lain & Vishwanath, 2021). Nigeria was struggling to survive the 2016 economic recession when it was hit by one more recession in 2020 due to the COVID-19 pandemic. Whereas the COVID-19 pandemic control measures instituted by the then Nigerian government contained the pandemic, those measures shattered people's livelihoods, social well-being, harmony, and human and industrial relationships (NBS, 2022). How could the Nigerian government totally shut down major cities and business activities in the country during the COVID-19 pandemic with ensuing economic consequences? Urban poverty increased with the pandemic. Nigeria also fares poorly in multidimensional poverty index or measures, with high inequality varying across the six geopolitical regions.[4] 53.3% of Nigerians were categorized as multi-dimensionally poor in 2017. Considering data from the 2018/19 Nigerian Living Standard Survey (NLSS) of the National Bureau of Statistics (NBS), official monetary poverty in 2019 was measured at 40.1%—meaning that 82.9 million Nigerians had real per capita expenditure below the poverty line of Naira 137,430 per year (or Naira 376.50 per day) and were therefore considered poor (NBS, 2019). This number has increased with about 65.64 million people aged above 18 being multi-dimensionally poor in 2021 (NBS, 2021).

Beyond the issue of high population and its associated impact on social protection coverage, Nigerian political and socio-economic structures seem neither to

[4] Based on a 2017 report by the Oxford Poverty and Human Development Initiative (OPHI).

serve the needs of its federating units nor entire citizens of the nation (not minding their ethnic group). Nigeria has a wide diversity of over 250 ethnic groups and dialects; however, it is divided largely into North and South. The presidency and political power is rotated between the North and South (see discussion in the background section), but this is largely contested as the politicians use this power sharing framework to rotate power among themselves. National integration has been an uphill task since independence. The structuring of Nigeria affects social service provision in the country. There is also a general belief that most of the poor in Nigeria are in the rural areas and many of them are located in the Northern area. The belief in the geo-spatial location and prevalence of poverty in some known locations has influenced the focus of humanitarian and social protection programs in Nigeria. The bias in the prevalence of poverty in northern Nigeria is controversial as there is a paucity of updated and robust data on poverty in Nigeria. The last Nigerian population census data is not available to the public. This has stalled the likelihood to pool census and survey data to generate poverty figures at the fine geographical scale required to inform social protection program targeting (Gething & Molini, 2015). There were revelations during the COVID-19 pandemic showing that the urban poor in major cities and their nearby slums are as vulnerable as other poor population elsewhere (such as poor in the rural areas) in Nigeria. Most of these poor people in the urban areas whether in southern or northern Nigeria live on daily earnings in an economy that is largely informal and with most jobs within the service sector. This must have warranted the obvious use of AI in targeting the poor, who are beneficiaries of the social protection program in Nigeria during the COVID-19 pandemic. The previous national social registry for SPP was no longer adequate to tackle poverty in Nigeria.

The COVID-19 pandemic was a game changer even though the political opposition in Nigeria and the press had been shouting out that the government social protection programs were associated with massive corruption leading to loss of government funds, money laundering, political patronage, and vote buying (Chenge & Oigbochie, 2023, Ojo, 2008). It was the COVID-19 pandemic that was seen to have made the government to resort to the use of AI and other high technologies in social protection delivery in Nigeria. AI was used to provide better satellite imagery-based poverty mapping that was used to target beneficiaries of the SPP instead of the community-based targeting used prior to the COVID-19 pandemic.

History, Political Economy, and Social Context of Social Protection in Nigeria

History of SPP in Nigeria

Social protection (SP) is not totally new in Nigeria and Africa as a whole (Slater & McCord, 2009). It is stated (Bastagli et al., 2016; Beegle et al., 2018) that in the 1920s the first cash transfer took place in South Africa and then in Namibia and

Botswana. From a historical point of view, social welfare has evolved in Nigeria from services mainly provided by the extended family in the pre-colonial period which differed significantly from today's social protection program (SPP) provided by the state (Okoye, 2013; Berrou et al. 2021). However, the kind of social services provided then varied from one ethnic group to the other. Religion and ethnic diversity worked together to create disparities in the social service delivery system, as evidenced by the Yorubas' extended family structure in the South West of Nigeria, the egalitarian Igbo system in the South East, and the Islamic Hausa/Fulani in the north (Olowu, 1980 as cited in Nwoba, 2015). The legacy of colonialism included a faster pace of urbanization and industrialization, which weakened the extended family structure that had previously protected the supply of social services in many traditional societies (Nwoba, 2011). A greater emphasis on child labor (children under 14 years old) was placed in the 1940 colonial social service statute.

The free education services for the disabled were highlighted in the 1945 legislation, and in 1955 a register for the disabled was established. The 1942 act prioritized the creation of boy's clubs as a means of preventing delinquency. The traditional values that served as the foundation for older social services were further abandoned by the missionaries' activities, particularly in the area of education advancement as a tool for social transformation and a requirement for religious conversion (Nwoba, 2011). However, the missionaries served as the first pillar of Nigeria's early social service delivery system, particularly with regard to the majority of the country's citizens who lived outside of Lagos and needed health care and education.

African countries implemented a range of SPPs after independence, including the provision of free health care and pensions for government employees, as well as food and agricultural subsidies to farmers. Nevertheless, the introduction of structural adjustment programs (SAPs) in the late 1980s and early 1990s caused the reduction of domestic expenditure on SPP and scaling down or forcing the total collapse of the program.

Some SPPs have been implemented by successive Nigerian governments since 1960 to date. See Fig. 9.3 stating the timeline. Programs such as Operation Feed the Nation, the Green Revolution, the Family Economic Advancement Program (FEAP) in 1992, National Directorate of Employment (NDE) in 1986, community banks, directorate of food and rural roads infrastructure (DFFRI) in 1986, better life for rural women in 1987, Education Trust Fund (ETF), 1993, Petroleum Trust Fund (PTF) established in 1994, and Pension Reform Project, 1994. The return to democracy in 1999 saw the introduction of the National Poverty Alleviation Program (NAPEP),[5] Subsidy Reinvestment Scheme (SURE-P), and Growth Enhancement

[5] NAPEP had Youth Empowerment Scheme and National Resource development and Conservation Scheme components. Jonathan administration introduced Subsidy Reinvestment Scheme (SURE -P), with the aim of reducing poverty comprised of the Mass transit Scheme, Vocational Training Scheme, and Community Service/Women and Youth Employment (CSWYE).

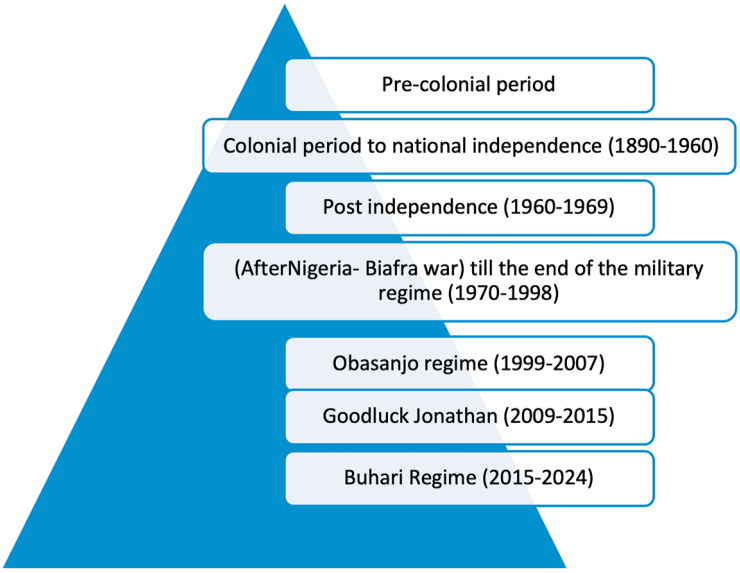

Fig. 9.3 Social protection program timeline in Nigeria

Support Scheme (GES) that used electronic wallet to distribute farm inputs (seeds and fertilizers).

However, it is important to note the role played by the implementation of the structural adjustment programs (SAP) in 1985. SAP as implemented in Nigeria led to massive disruption of SPP and the entire social policy space (Gboyega et al., 2011; Ayoade et al., 2014). SAP had an indelible impact on the social fabric of the Nigerian state with the central structure having tensions with the other levels of government on revenue sharing formulas. This shrank the SPP space until the Goodluck Jonathan regime and then notably the Buhari government took SP to the present level.

The administration of President Buhari in Nigeria launched the National Social Investment Program (NSIP) in 2016 to tackle the country's high rate of poverty and vulnerability. NSIP comprises four major components: the N-Power designed to assist young graduates to acquire and develop lifelong skills; the Conditional Cash Transfer (CCT) for the support of those within the lowest poverty brackets; the Government Enterprise and Empowerment Program (GEEP), which is a micro-lending intervention for traders, farmers, women, etc.; and the National Home Grown School Feeding Program (NHGSF), which aims to deliver school meals to young children.

Political Economy and Social Context SPP in Nigeria

An analysis of political economics is necessary to understand the trajectory of SP and CTs in Nigeria. Social outcomes are controlled by politics because the political economy approach emphasizes social, political, and economic structures and connections that may be, and frequently are, outside the purview of the institutions or individuals they affect (Bambra et al., 2005). According to Krueger (1993), the political economy approach produces patterns of social outcomes through the structures, values, and priorities of political and economic systems; as a result, social inequality varies among countries and is characterized by factors such as privilege, power, and property.

Lord Frederick Lugard was governor of both the Northern Nigeria Protectorate and the Colony and Protectorate of Southern Nigeria and on January 1, 1914, signed a document amalgamating the two to form the Colony and Protectorate of Nigeria (Steinberg, 1965). The South had the Eastern and Western regions. The northern part due to its larger land mass was given more federating units that amounted to more political power in the central or federal government and house of parliament. The political opposition and clamor for independence from British rule were mainly from the Southern Nigeria. The colonial master Britain influenced the emergence of the political economic structure that subsists in Nigeria to date. Geopolitical zones were created after some political agreements, but each zone does not have equal states as seen in Table 9.1. Nigeria is socially and culturally diverse.

Nigeria's geopolitical structure and religious bipolarity give a space for a unique vibrant cultural and socially diverse nation, but its more than 250 ethnic nationalities have remained a source of intercultural tension, political power tussle, state capture, corruption, etc.

Nigeria is well endowed with vast human and natural resources that should warrant sustainable economic growth and development of the country. Nigeria is a low middle-income country with high dependence on oil revenues, even though there has also been growth in the non-oil economy in recent years (Holmes et al., 2012a). Huge crude oil and natural gas reserves of over 37[6] billion barrels and 187 trillion standard cubic feet, respectively, are deposited in the southern part of Nigeria. The major source of revenue to the country is from the exportation of crude oil and natural gas. Nigeria is resource dependent and its economy extractive. Extractive

Table 9.1 Six geopolitical zones in Nigeria and states within each zone

S/n	Zones	States
1	North Central	Benue, Kogi, Kwara, Nasarawa, Niger, Plateau, and Federal Capital Territory, Abuja
2	North East	Adamawa, Bauchi, Borno, Gombe, Taraba, and Yobe
3	North West	Jigawa, Kaduna, Kano, Katsina, Kebbi, Sokoto, and Zamfara
4	South East	Abia, Anambra, Ebonyi, Enugu, and Imo
5	South-South	Akwa Ibom, Bayelsa, Cross River, Delta, Edo, and Rivers
6	South West	Ekiti, Lagos, Ogun, Ondo, Osun, and Oyo

[6] https://www.nigerianstat.gov.ng/pdfuploads/NIGERIAN%20ECONOMY.pdf

economies have the characteristics of easy state capture, massive corruption, weakness of the state, and poor economic development. Agriculture with low productivity dominates the most of the Nigerian economy, and northern Nigeria is mostly involved in agricultural production of the country. There has been a loss of manufacturing since SAP in the early 1980s; mass employment in agriculture and then service sectors could partly explain the issue of a wealth country and poor people in Nigeria.

Population figures seem to comply with the landmass rather than the actual number of people. The data from the last Nigerian National population census of 2006 is yet to be publicly available (Bertoni et al., 2016; FAO, 2015).[7] So the population figures and census data are issues of controversy. Of the southern regions, the southeast is the most affected by poverty. States and LGAs in this region show poverty rates that are consistently over 40%. In LGAs in the state of Enugu, poverty rates of 85% and over are observed. In general, as it appears from the maps, poverty is concentrated in a belt stretching from Zamfara State and covering most of the Northwestern and Eastern States. High poverty concentration is particularly visible in a number of LGA's clusters in Katsina and Kano States; these are also densely populated areas. Poverty pockets are present along the Eastern part of Nigeria bordering Cameroon; these are mainly Northeastern States with a few belonging to the Southeastern region, notably Enugu and Ebonyi.

The British handed over power in 1960 to Nigeria but did not mind how it could affect the social well-being and sustainable development of the country. The Nigeria-Biafra war took place just 6 years after independence and this attests to the deep-rooted ethnic interests and division in Nigeria. This history and sociopolitical background could easily give someone a hint of the political undertones that are often associated with public service delivery in Nigeria. Most often than not the heads of government are accused of serving ethnic rather than national interests (Shadare, 2020). Politics in Nigeria has been branded as patron–client relations informed by ethnic identity across political parties (Joseph, 2014). Political patrons continually provide material benefits to clients to remain in power and enrich themselves from the corruption system. The political background and ethnic divisions have made it hard to curb corruption, which is pervasive in most public service delivery systems in Nigeria. Political–client relations and political patronage have crippled sustainable economic growth but remain sources of the informal SP in Nigeria. These things disempower citizens and drive poverty in Nigeria.

The northern Nigeria is dominated by Islamic religion and the south is majorly made up of Christians. There are religious issues in Nigeria although Nigeria is officially a secular state, but religion also affects policies/programs of government and to a large extent the balance of political power of the state.

Today, Nigeria is structured such that there are six geopolitical zones that constitute the whole north and south division. There are three geopolitical zones each in the north and south of Nigeria. Three geopolitical zones out of the six are yet to

[7] https://data.apps.fao.org/catalog/dataset/54753faa-7f30-4841-ba8d-9b3685183f37/resource/beea75bc-afbf-41d3-9bdf-cd068acffb6b/download/developing-an-updated-poverty-map-for-nigeria.pdf

produce the president of country for over 64 years of independence and rotation of presidency. The rotation seems to favor the politicians from the North West of Nigeria and now the South West of Nigeria. There is continued agitation for cessations even 60 years after the bloody Nigeria-Biafra civil war. There are 36 states in the federal republic but unfortunately the states are not equally divided among the six geopolitical zones. North West and North East geopolitical zones have seven states, whereas South East has just five states. This affects the distribution of public goods among the ethnic tribes. The structuring of the federation into states and creation of states not only affect the number of representatives to the senate and national house of assembly but are seen as marginalization. The patterns of poverty seem to vary by geographic location and are predisposed by social, cultural, religious customs, and the predominance of conflict and instability (Holmes et al., 2012a). However, the general assumption that poverty in Nigeria is geographic (Northern area) is really a political paradox as the politicians of the Northern extraction had ruled the country more than those in the south for the 64 years of national independence. The distribution of public goods in Nigeria had neither been equitable nor fair to take care of the poor. The poor in the north and south are left out in the distribution of national wealth. Poverty and vulnerability had been aggravated by food, fuel, and financial (Triple F) crisis (Holmes et al., 2012a), yet hunger and poverty have been weaponized against Nigerians by the recent removal of subsidy on petrol and electricity costs and associated high inflation. The impacts of recent subsidy removals and high inflation are concurrent with the global impacts of the Russia–Ukraine war just after the impacts of the COVID-19 pandemic (World Economic Forum, 2023).

The removal of the subsidies has further contracted the economic space and widened income disparity in Nigeria today. Income inequality is one of the multidimensions of poverty in Nigeria: poverty and vulnerability are also highly influenced by social and other factors such as age, gender, and beyond geography and ethnicity (Holmes et al., 2012a) as highlighted earlier in the chapter.

The AI FORA Nigerian Case Study

Case Study Focus

The case study focuses on the national cash transfer (NCT) program also known as Household Uplifting Program (HUP). Cash transfer programs targeted on poor and vulnerable households are the most rapidly growing type of social safety net programs (SSNP[8]). It is one of the SSNPs implemented by the Nigerian government with a focus on addressing key social concerns, deficiencies in capacity, and lack of

[8] Social safety net program (SSNP) is a collection of public measures intended to assist individuals, households, and communities in managing risks in order to reduce vulnerability, improve consumptions smoothing, and enhance equity while contributing to economic development in a participatory manner.

investment in human capital, especially among our poorest citizens. NCT was launched in September, 2016. NCT has three components which are base cash transfer, top-up based on state selected conditions, and livelihood support. Beneficiaries of NCT were mined from the National Social Register (NSR), comprising State Social Registers (SR) of poor and vulnerable households. According to NASSCO (2019), the NCT program is made to support development goals and priorities by providing beneficiary households with timely and easily accessible cash transfers. It aims to accomplish the following specific outcomes:

- Boost household consumption
- Increase the use of health and nutrition services
- Increase school enrolment and attendance
- Boost environmental management and sanitation
- Promote the financial and asset acquisition of the households
- Encourage recipients in sustainable livelihood

According to the immediate past Vice President Prof Yemi Osibanjo, the COVID-19 cash transfer project was aimed to lift the urban poor highly impacted by the COVID-19 pandemic out of poverty. Table 9.2 captures the typical social safety net program components as described by the International Monetary Fund.

The description in Table 9.2 as captured by IMF seems to describe the national safety net program that is implemented in Nigeria.

Table 9.2 Description of social safety net as captured by the IMF (2022)

Social safety nets (SSNs) comprise a broad array of non-contributory transfer programs designed to protect households from poverty and destitution. While the programs that compose SSNs can be classified in different ways (see Grosh et al., 2008; Beegle et al., 2018; World Bank, 2018), these programs are widely understood to include: • Cash transfer programs. These programs offer periodic monetary transfers with the objective of providing regular and predictable income support to beneficiaries. This category covers a wide range of programs, including poverty-targeted cash transfers; family, children, and orphan allowances (including benefits for vulnerable children); and non-contributory funeral grants and burial allowances. This includes both conditional and unconditional cash transfer programs. • Social pensions. These are regular cash transfers provided exclusively to the elderly which, unlike contributory pensions, do not require prior contributions; social pensions may be universal or targeted to the poor and vulnerable. • School feeding programs. These programs supply meals or snacks for children at school, including take-home food rations for children's families. • Nutrition programs. These include programs providing food stamps and vouchers, food distribution programs, and other nutritional programs (excluding school feeding programs). • Fee waivers and targeted subsidies. This encompasses a wide variety of targeted (as opposed to universal) subsidies focused on the poor and vulnerable, including health insurance exemptions and reduced medical fees, education fee waivers, housing subsidies and allowances, utility and electricity subsidies and allowances, agricultural input subsidies, and transportation benefits • Emergency programs. These include programs providing emergency support in cash and in-kind to individuals in case of emergency or in response to shocks (including support to refugees and returning migrants). Transfers are usually temporary. Other programs. These include other non-contributory programs intended to target the poor or vulnerable, such as programs distributing school supplies, tax exemptions, social care services, and other programs not included in the other categories.

Key Research Questions

The case study attempted to answer the following research questions:

- How exactly are NCT targeted beneficiaries assessed for receiving cash in the SPP?
- How would the six geopolitical zones represent different cultural contexts with different assumptions, values, etc.?
- How do beneficiaries/people perceive the NCT program?
- What are the technologies (AI technologies included) used/planned and the focus of goals of AI technologies used in NCT in Nigeria?
- How do culture and values around social well-being create motivation for the NCT?
- What are assumptions, values, and norms related to what AI fairness means in the context of cash transfer and what are causes/sources of inequities, including AI-related issues?
- How do culture and social values affect the future role of increased AI use in NCT in Nigeria?

Empirical Social Research in AI FORA

The concept of social policy was used to capture the approach of NCT and strategy used in the implementation of the intervention in Nigeria. Nigeria seems to use a residual welfare or humanitarian assistance approach, or framework dominantly used by the World Bank. Most of the secondary data were extracted from literature review and document analysis, but the primary data were extracted from interviews and focus group discussions.

Actors and Stakeholders Involved in the NCT (Mapping of Actors of the NCT)

The mapping of actors of the SPP in Nigeria has been done before by Holmes et al. (2012a). SP is cross-cutting requiring a number of ministries, departments, and agencies (MDAs). It also involves donors, non-governmental organizations, and civil society that provide assistance for the funding, development, and implementation of new programs. The success of policies and programs of social protection necessitates proper coordination, both horizontally between MDAs and other agencies and vertically between the federal and state levels, under the direction of a single ministry.

The actors in the NCT program are the Federal government of Nigeria,[9] the World Bank, service providers, community members, beneficiaries, the National Social Safety Net Coordinating Office[10] (NASSCO), and National Cash Transfer Office (NCTO). NCTO's mandate is to deliver the targeted cash transfer across the country; the actual implementation happens at the state level through the State Cash Transfer Unit (SCTU). SCTU manages and coordinates the targeted cash transfer and livelihood intervention. Each local government area establishes a cash transfer team to implement activities at the community level. Other stakeholders include frontline health workers, pregnant women, school management and school children enrollees, policymakers, and program managers in the conditional cash transfer programs. Table 9.3 captures the category of these actors.

Table 9.3 Categories of NCT stakeholders

S/No	Stakeholder types	Stakeholders
1	Government actors/Internal stakeholders	Relevant internal stakeholders: NASSCO and NCTO staff, SCTU and SOCU, Federal Ministry of Humanitarian Affairs, Disaster Management and Social Development (FMHDSD), Federal Ministry of Finance, Budget and National Planning, other MDAs—Education, Women Affairs, Labour and Productivity, Youth and Sports, Agriculture, Health, Communication and Digital Economy, National Identity Management Commission, National Orientation Agency, National Primary Health Care Management Agency, National Health Insurance Scheme, National Directorate of Employment, etc.
2	Beneficiaries	1. Poor and vulnerable Nigerians identified by communities 2. Target communities; their local leaders, traditional rulers; local government authorities 3. State and federal authorities (governors, commissioners/ministers of relevant ministries working with the poor and vulnerable, state and federal lawmakers) 4. Opinion shapers, political influencers, development experts (champions and adversaries) 5. Nigerian media, bloggers, social media influencers 6. Civil society groups, faith-based organizations
3	Multilateral organizations and UN agencies	World Bank and IMF followed by UN agencies (UNICEF, FAO, WFP, WHO)
4	Bilateral organizations	USAID, DFID, GIF (Deutsche Gesellschaft für Internationale Zusammenarbeit), EU

[9] Ezenwaka, U., Manzano, A., Onyedinma, C., Ogbozor, P., Agbawodikeizu, U., Etiaba, E., Ensor, T., Onwujekwe, O., Ebenso, B., Uzochukwu, B., and Mirzoev, T. (2021). Influence of Conditional Cash Transfers on the Uptake of Maternal and Child Health Services in Nigeria: Insights from a Mixed-Methods Study. Frontiers in Public Health, https://doi.org/10.3389/fpubh.2021.670534

[10] This was established in 2016 by the Government of Nigeria in partnership with the World Bank to strengthen social safety nets and social protection system in Nigeria as a core strategy to help end extreme poverty and to promote shared prosperity. The core mandate of NASSCO therefore is to lay a strong foundation of rigorous and reliable evidence of poor and vulnerable households in Nigeria, by building a National Social Register (NSR), as well as coordinate, refine, and integrate the social safety net programs into social protection systems while ensuring policy coherence.

Current Social Assessment Practice for SP Schemes in Nigeria

There have been impact evaluations of the cash transfer programs in Nigeria and most sub-Saharan African countries. The Transfer Project[11] (a research project) is one of such assessments in Africa and project claims to have spawned thorough evidence on the impacts of cash transfers in sub-Saharan Africa and hence supported their expansion. The research project produced a broad array of positive impacts of cash transfer programs from longitudinal mixed-methods impact evaluations across government-implemented unconditional cash transfer programs in ten SSA countries.[12] However, there were mixed and unintended impacts.

The impacts of CCT programs in Nigeria have been assessed over time to have generated some positive outcomes on food security, malnutrition, momentary reduction of poverty and inequality levels, significant increase in child survival, maternal health, and increase of health-seeking behavior of beneficiaries (Obeten & Isokon, 2018; Onwujekwe et al., 2020; Eluwa et al., 2023). Holmes et al. (2012a, b) assessed SP in Nigeria and carried out an extensive evaluation of the impact of cash transfer programs in Nigeria. They also carried out a micro-simulation targeting analysis.

An interesting social assessment of SP as now being conducted in Nigeria is the Accountability to Affected Populations (AAP) assessments. One was conducted on target communities in local government areas of Borno State, North East Nigeria. AAP is all about the inclusion of affected people in program design, impact evaluation, and monitoring systems. It means stakeholders' inclusion, inspiring participation, and empowering program beneficiaries to state their needs and to have their voices heard. It is also seen as a right-based approach[13] whereby donors and aid-based organization recognize the dignity, safety, and rights of people they serve, rather than treat them as passive "aid recipients." These right holders are acknowledged as primary stakeholders.

The AAP assessment in Borno State captured the perceptions of beneficiaries around five key themes, which are awareness of humanitarian service delivery, fairness/inclusion of the humanitarian response, feedback modalities within the humanitarian response, relevance of humanitarian interventions, and respect of affected populations by humanitarian service providers. The assessment also explored aspects of protection concerns and barriers to accessing the interventions. Awareness

[11] The Transfer Project is a collaboration among UNICEF (Innocent, Regional and Country Offices), the Food and Agriculture Organization of the United Nations, the University of North Carolina at Chapel Hill, national governments, and researchers.

[12] Impact evaluations from projects in Ethiopia, Ghana, Kenya, Lesotho, Madagascar, Malawi, South Africa, Tanzania, Zambia, and Zimbabwe, then Burkina Faso and Mozambique, and previous countries with completed evaluations (Ethiopia, Ghana, and Tanzania).

[13] https://www.nutritioncluster.net/sites/nutritioncluster.com/files/202010/Accountability_to_Affected_Populations_a_handbook_for_UNICEF_and_partners.pdf

varied between the community leaders and members and indicated a challenge for SP implementation to carefully balance stakeholder management of the traditional authority structure in communities while addressing the beneficiaries' desire for direct communication with SP providers. Beneficiaries' perceptions of fairness on the selection of target beneficiaries and assistance distribution within their settlements showed that people believed SP providers were responsible for selecting beneficiaries and that community leaders purposively selected their own family or friends to receive assistance. This is a perception of favoritism and power play. The vast majority of beneficiaries preferred direct communication and sharing of information with SP providers as they share a perception that community leaders hold back information and prioritize their own family or friends over them.

However, beneficiaries who received assistance 6 months prior to the assessment were satisfied with the relevance and appropriateness of support to their needs, while others reported less satisfaction with offered assistance base on value and amount provided, poor targeting, delay, and irrelevant assistance. This is indicative of the necessity for more tailored programming and skill empowerment to engage in productive activities. They also perceived main barriers to the program as communication between communities SP providers and then targeting of beneficiaries.

The incorporation of the less measurable social and political variables in social assessment of SP, coupled with measures of unintended positive and negative impacts, should be integrated in current assessments of SP schemes. Other factors to be considered in such assessments are the effects of the SP on social relations between beneficiaries and non-beneficiaries, between recipient households and non-recipient households, and then the potential for empowerment of beneficiaries and the communities at large.

The NCT in Nigeria runs under the SSNP and seems to be following the World Bank's social risk management framework (SRMF[14]) for SP, which was over time the dominant underpinnings for framing SP and used to structure SP as SSNP. However, SMRF framework got criticism for the lack of extensive conception of vulnerability, chronic poverty, and social risks. Social inclusion, cohesion, and stability were included as positive externalities of SRMF. The current social assessment practice for cash transfer in Nigeria is mainly influenced by the implicit and explicit objectives of the national policies as institutionalized by the government and supported by its development partners (World Bank and IMF). The World Bank approach to SP seems to favor the humanitarian, welfare-based approach as seen in its loans and logistics for SSNPs as seen mainly in NCT in Nigeria. The present assessment in Nigeria should rather follow the International Labour organization (ILO, 2019) SP conception, which is structured to maintain dignity, ensure human rights, foster inclusion in economic growth, and support policies for human

[14] SRMF is an analytical tool to categorize alternative policies and measures for addressing livelihood risks.

capital development and empowerment of the poor to participate in productive livelihoods. According to Morlachetti (2016) and Sepúlveda and Nyst (2012), states should take up overall and primary responsibility for structuring and sustaining SP systems. Realizations of the right to social security spill over to economic and other sociocultural rights, which comprises food security, shelter, clothing, health, and education, which are necessary for achieving human dignity.

Cash transfers can be top-ups with material impact, coupled with social and political processes that have social and political impacts (MacAuslan & Riemenschneider, 2011). Most assessment and evaluations of cash transfers in current multilateral organizations development corridor do not focus on the social and political impacts of cash transfers. There is a strong case for considering the nature of cash transfers, decision-making processes (stakeholders/beneficiaries involvement), and institutional structures that support these programs, especially as cash transfers (such as conditional cash transfers—CCTs) were initially intended as momentary, compensatory, or emergency-based schemes (Bastagli, 2011; FAO, 2018; UNDP, 2019). NCT in Nigeria has followed the trend and features of earlier CCTs implemented in Nigeria. Today NCTs in Nigeria are delivered to citizens in the form of unconditional cash transfer (UCTs), yet they share all the features of CCTs in the duration of the program. The social assessment of the NCT should improve to accommodate present-day realities but that may happen when the objectives are to provide SP to citizens with a human rights approach and realize human dignity and overall well-being of citizens.

Any social assistance program is embedded within a web of social relationships that is molded in turn by strongly held social norms, beliefs, and values (MacAuslan & Riemenschneider, 2011). These influence how it is perceived by different stakeholders, how it operates in practice, scope for its reform, and ultimately its effect on material well-being as well. The targeting process of NCT in Nigeria should be assessed to see if there is a probability of strengthening the symbolic role of the state government, market group leaders, and community leaders, as they are involved in community-based targeting process used in NCT. Most times these leaders may double as political party community leadership in the locality. What is the proportion of NCT beneficiary households reporting reduction on receiving assistance from other households and other members of the community or organizations? This may probably reflect a perception that these households need less assistance with the often-short NCT program. Beneficiaries of NCT are individuals embedded within varying institutions at different levels of households, communities, and national politics and social traditions. Funds, power, and know-how flow through these institutions. Cash transfer programmes influence these flows at each part of the delivery chain: stakeholder engagement, targeting, enrollment, payment, and monitoring and evaluation. Social assessments should include issues of social relations and symbolic roles of the stakeholders involved in the providing of the safety nets.

Context and Culture Matters for Social Assessment of Beneficiaries of NCT in Nigeria

The context for the social assessment of beneficiaries of NCT in Nigeria is such that poverty is prevalent and beneficiaries must be selected for the interventions based on some proximity means testing. Those assessed to be more vulnerable than others are to be targeted and selected for cash transfers. This exposes the beneficiaries to some social relationship risks or conflict with community members who are not selected or non-beneficiaries. Beneficiaries have to be assessed on their satisfaction and social well-being after the program. Will the beneficiaries lose friends or community assistance after the cash transfer intervention which often lasts for few months in the case of NCT?

According to UNPF (2008), the concept of culture implies the totality of factors that affect perceptions, understanding, behavior, and responses of human beings on issues. Culture cannot be quantified, but is dynamic and inescapable. Africans tend to be more communal with strong family or social ties in their communities or villages. Targeting of beneficiaries of NCT often creates social conflict between households that were enrolled or selected for the cash transfer and those excluded. This affects the social assessment of beneficiaries as they will report resentment toward treatments from non-beneficiaries. There is also generally a rent-seeking behavior from state coordination groups or community leaders in Nigerian society. The culture of rent seeking is known to affect the selection and targeting of beneficiaries of social assistance. NCT beneficiaries will often not want to share part of their cash benefits with the middle men or community leaders involved in the targeting process. When beneficiaries share their benefits, it will affect their perception of social assistance. Some of these beneficiaries will perceive the cash transfer as political patronage and may be forced to give more loyalty to those middle men or community leaders. Increase in human dignity and total well-being of the beneficiaries should be the focus of assessment.

The culture in Nigeria and Africa as a whole assigns some responsibilities to gender roles and there are household traditions or institution that guides female or women participation in social activities or engagements. Selecting women (especially the married women) or girls as targeted beneficiaries does not make them not to submit to the heads of the households who would make the ultimate decision or give permission for the possible use of the cash transferred to the female beneficiaries. Most times cash transfers targeted to women beneficiaries are meant for the household and that context should be considered in the design and targeting of beneficiaries of NCT.

Conditional cash transfer programs should be assessed based on culturally informed beliefs and practices, ensuring high-quality services and consistent availability for vulnerable populations. Theodore et al. (2019) and Habicht and Pelto (2019) established that sociocultural factors, which includes social norms, traditional beliefs, and patriarchal family organization, contribute substantial measures in approval, implementation, and impacts of the SP.

Bastagli et al. (2016) reported community perceptions of humanitarian assistance in targeted local government areas (LGAs) of Borno State, North East Nigeria. It was an assessment of beneficiaries' perception of social assistance provided to them.

The Use of AI or Other Technologies in Implementation of NCT and in Which Parts of NCT Delivery Is AI (If Any) Deployed?

The adoption of AI in SP in Nigeria SSNP is in its early stage. The utilization of AI and machine learning in the SSNP delivery chain in Nigeria has just been limited to mining of data from the social registry and targeting of beneficiary for the cash transfer schemes. AI has been extensively used in targeting social protection programs in Nigeria by constructing high-resolution poverty maps from satellite imagery to help augment community base targeting of beneficiaries of cash transfers during the COVID-19 pandemic. Development of an updated poverty map was based on spatial statistical modeling approaches that could support improved geographical targeting of poverty alleviation programs in Nigeria (Bertoni et al., 2016; FAO, 2015). According to Stockman et al. (2022), machine learning and artificial intelligence could also help identify clients at greatest risk of treatment interruption in cash transfer programs in Nigeria, by simply combining temporal and location data of the beneficiaries.

What Are the Problems and Barriers of the Current System?

NCT in Nigeria may have a robust design but faces implementation challenges (Oduenyi et al., 2019). The current system has some known barriers and challenges. They are coverage and eligibility issues and sociocultural issues such as gender, age, and polygamy/family sizes. Others are the duration of scheme or participation in the intervention, the value of the cash transfer, and lack of complementary services. A significant challenge with cash transfer (as opposed to other types of social protection instrument is that the value of the cash is very low compared with the need of households (especially in the context of increasing prices and variations in state-level provision of services). Oduenyi et al. (2019) conducted this study for the CCT pilot program under the Subsidy Reinvestment and Empowerment Program on Maternal and Child Health (SURE-P MCH) in Nigeria. This barrier has remained with the recent UCT in recent years. The value of the cash is low compared to the high inflation in Nigeria.

It is also notable that Nigerians are still entrenched in their past culture with strong attachment to cash payments, security concerns of digital theft, and infrastructure deficits (Soyemi et al., 2015). All these affect digitalization and shift to the use of AI in its SSNP and cash transfer. Other problems identified in cash transfer schemes in Nigeria include the issue of diversion of funds by beneficiaries, inappropriate delineation of exit and entry periods, and errors of beneficiary exclusion and inclusion (Paul, 2022). Another significant challenge is the issue of the NCT biting more than it can chew; the support is meager and deficient to graduate citizens from the multidimensional poverty in Nigeria. The paltry value of the cash transfer is exacerbated by the large family size of average households in polygamous households in Northern Nigeria. The scheme also seems to support households for a short duration like 3 months. This is likely to skew the perception of the

public on the entire intervention on even the highest or best of technologies used. Infrastructure gaps also affect the deployment of high-end technologies in the program or transfer of cash to people in remote or rural areas. Nigerian banks control cash transfers and may add to cost of cash transfers. Mobile money and cash transfer seem to be more diversified in eastern Africa than in Nigeria and west of Africa. Geographical location is also a barrier as it limits connectivity and Internet/mobile network connections. Some people or ethnic groups in Nigeria are migratory or engaged in nomadic lives. You cannot target them as households as they migrate in different seasons of the year. Some such as the fishing or remote coastal communities are forced to move or migrate during flooding seasons. Migration due to climate change (as seen during high flooding in southern coastal communities or during drought in semi-arid northern Nigeria) is a barrier to effective targeting of poverty and effectiveness of interventions. There are also the barriers caused by the lack of national identification or national identification system, not having a government-recognized form of identification, and illiteracy of significant portion of the population, especially the marginalized people. Data management, data quality, and issues of personal data privacy as enshrined in national data privacy remain to be a barrier to NCT program in Nigeria. Integration of data for SP into or public use is also a possible barrier that may make up fears of people offering their data for SP purposes.

Target Vulnerable Group of the NCT in Nigeria

Basically, HUP or NCT is targeted at the poor and vulnerable households in rural areas across different states in Nigeria especially in the northern part of the country. However, COVID-19 impacted the Nigerian economy and the disruption of livelihoods already impacted by two quick recessions resulted in the emergence of population of urban poor in big cities in Nigeria. These are the targets of the NCT in Nigeria.

Discussion of Results, Conclusion, and Outlook

Discussion

The Nigerian state should invest in social inclusion and sustainable growth by fostering social relationships and social empowerment.[15] National accountability to rights holders will win over the citizens' trust and loyalty. It is possible for states to

[15] Social empowerment looks like organizations and institutions that help marginalized groups of people gain the resources to be empowered, such as material assets, good health, education, social belonging, self-esteem, self-confidence, and economic opportunity.

revitalize their strained social contracts[16] (Krainov et al., 2021; Razavi et al., 2020) and improve their developmental status when it socially enfranchises its citizens.

The goal of building empowerment structures is to give the poor access to resources that give them a sense of identity and dignity, social inclusion, and well-being and offer them the opportunity for social inclusion as citizens will not feel disenfranchised by middlemen or rent-seeking community leaders. There is a high chance of people in Nigeria trusting the AI tools or algorithm rather than their local leaders (about whom they already have reservations for being biased) as seen in the SP assessment in Borno State in Nigeria as cited above. There are concrete public concerns about the arbitrariness of AI use and human bias in AI algorithm; it is salient to ask about if AI will be free of some form of bias. However, the relevant question should be to know if AI can function more effectively than human-based system. The issue of fairness and equity in government decision-making and politics in Nigeria and elsewhere brings more demand for a shift to AI, which will at least be more consistent than politicians or other humans as public service providers. It is easier to do damage control for a biased AI algorithm rather than to engage in productive activities. AI can support the reduction of errors of inclusion and exclusion of eligible beneficiaries or targeted recipients of government SP. Unlike in the earlier episodes where SP schemes are perceived to be tools of political patronage, vote buying by government, and favoritism by community leaders, AI will help build self-confidence in the SP system as citizens can be assured of their rights to have access to government intervention. This will help reduce social conflict and resentment and increase social cohesion even with fewer resources to provide universal SP. The use of AI in SP delivery will help build a to eliminate genuinely imbedded human biases.

The transformation of the SP delivery chain will help build empowerment structures which will address structural causes of poverty and vulnerability. AI could disrupt the old order and radically enhance efficiency, effectiveness, and quality of social service delivery as seen in education, agriculture, health care, and other sectors (Bullock, 2019; Samoili et al., 2020). AI could assist in all stages of the SP delivery chain including assessment of prospective beneficiaries' eligibility, decisions on enrolment decisions, offering payments or other assistance, and managing and monitoring delivery of social assistance.

Nigeria is yet to deploy AI in the entire SP delivery chain and it is not proper to speculate on how AI will change procedures of assessment but I have stated above how AI will increase the changes of empowering the citizens to assess SP interventions and how it will reduce the rent-seeking behavior of the middle men or community leadership and also increase fairness and equity as it relates to selecting citizens irrespective of their gender, location/ethnicity, religion, social status, closeness to power brokers, etc.

[16] A social contract can be defined as an implicit agreement between all members of a society—whether defined in terms of government and citizens, labor and capital, or different population groups—to cooperate for their mutual benefit and respect each other's rights and obligations (ILO 2016).

The SP delivery chain provides the functional anchor for the delivery systems framework. Generally, all forms of SP go through and have common implementation stages along their delivery chain. These delivery chains basically have outreach stage, assessment, enrolment, provision or provision and payment stage, and then monitoring and evaluation stages. AI tools can be effectively used in each stage of the delivery chain to boost performance.

Conclusion and Outlook

The majority of Nigeria's huge population is vulnerable to varied dimensions of poverty. Social assessment will assist in spotting real challenges, factors driving vulnerability and poverty rates, and barriers to design effective interventions. AI use in SP will fix targeting issues contributing to social exclusion, deprivation of rights to SP, and marginalizing majority of the Nigerian populace. AI-assisted SP in Nigeria will help facilitate the possibility and chances for citizens to find the means of livelihood or employment. Such an AI-assisted SP system will resemble or share features of universal social protection. AI can be an enabler of effective targeting of beneficiaries of SP but government must have the political will to take the responsibility of keeping social contract of providing protection to its citizens whether in formal or informal employment irrespective of their geographic locations or ethnicity. Government should as a matter of policy ensure inclusion of all stakeholders in developing the delivery chain for SP intervention. Then it can deploy AI and related technologies in all stages or part of the SP delivery chain. The commitment of the state to keep to its social contract will change the perception of the citizens and trust in all means (including technologies such as AI) to be used to deliver SP to them.

This case study contributes massively to the AI FORA project's objective in that it presented a context where poverty is pervasive, and there is an opportunity to use AI to make meaningful progress in SP delivery in the country. This also opens an opportunity for social assessment for the use of AI in SP delivery in Nigeria. This study makes recommendations for "better AI" or responsible AI terms in the context of massive, failed human system in the management of SP. It also recommends the inclusion of beneficiaries of SP as participants and stakeholders in the design, monitoring, and evaluation of SP in Nigeria. There is also a suggestion to carry out more elaborate social assessment of SP rather than evaluations with economic metrics or econometrics.

The outlook of this case study will be to model a new national SP policy framework with AI strategy and carry out a micro-simulation targeting analysis of an AI-enabled social cash transfer program in Nigeria. Further research also includes engaging in stakeholder-actors network mapping and analysis using an open-source program.

There is a plan to call for a research dissemination workshop to share the research results, which will open up opportunities for possible uptake of recommendations from the study in SP policymaking, implementation of SPP in Nigeria, and

assessments and impact evaluation of SP in Nigeria. The research dissemination workshop will invite representatives of multilateral organizations in Nigeria, the Ministry of Humanitarian Affairs, bilateral organizations, non-governmental organizations, disaster management and social development, other cognate ministries, National Social Safety Net Coordinating Office, civil society organizations, other NCO stakeholders, AI FORA research groups, etc.

Acknowledgments Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

References

Ayoade, J. A. A., Akinsanya, A. A., & Ojo, O. J. B. [Eds.] (2014). *Nigeria: Descent into anarchy and collapse*? University Press of America.

Bambra, C., Fox, D., & Scott-Samuel, A. (2005). Towards a politics of health. *Health Promotion International, 20*(2), 187–193.

Bastagli, F. (2011). Conditional cash transfers as a tool of social policy. *Economic & Political Weekly, 46*(21), 61.

Bastagli, F., Hagen-Zanker, J., Harman, L., Barca, V., Sturge, G., Schmidt, T., & Pellerano, L. (2016). *Cash transfers: What does the evidence say? A rigorous review of programme impact and the role of design and implementation features*. ODI.

Beegle, K., Coudouel, A., & Monsalve, E. (2018). *Overview: Realizing the full potential of social safety nets in Africa*. World Bank. https://doi.org/10.1596/978-1-4648-1164-7_ov. Accessed 30 April 2024.

Berrou, J.-P., Piveteau, A., Deguilhem, T., Delpy, L., & Gondard-Delcroix, C. (2021). *Who drives if no-one governs? A social network analysis of social protection policy in Madagascar*. [Research Report].

Bertoni, E., Clementi, F., Molini, V., Schettino, F., & Teraoka, H. (2016). *Poverty work program: poverty reduction in Nigeria in the last decade*. Osun State, Nigeria.

Bullock, J. B. (2019). Artificial intelligence, discretion, and bureaucracy. *The American Review of Public Administration, 49*(7), 751–761.

Chenge, A. A., & Oigbochie, A. E. (2023). Cash and ballots: Unearthing the nexus between conditional cash transfers (CCTS) and voter's electoral behavior in Nigeria. *International Journal of Political Science and Governance, 5*(2), 94–100.

Devereux, S. (2019). A brief history of social protection's brief history in Africa. *Presentation at the 11th NordWel Summer School, CRC 1342, Global Dynamics of Social Policy*. University of Bremen.

Eluwa, T. F., Eluwa, G. I., Iorwa, A., Daini, B. O., Abdullahi, K., Balogun, M., & Lawal, A. (2023). Impact of unconditional cash transfers on household livelihood outcomes in Nigeria. *Journal of Social Policy*, 1–16.

FAO (2018). *FAO and Cash+. How to maximize the impacts of cash transfers*. Rome: FAO.

FAO (2015). *Food Outlook: Biannual Report on Global Food Markets*. Rome: FAO.

Gboyega, A., Søreide, T., Mihn-Le, T., & Shukla, G. P. (2011). *Political Economy of the petroleum Sector in Nigeria*. Policy Research Working Paper 5779. Africa Region. Washington: The World Bank.

Gething, P. W., & Molini, V. (2015). *Developing an updated poverty map for Nigeria*. Unpublished working paper. World Bank.

Grosh, M. E., Del Ninno, C., Tesliuc, E., & Ouerghi, A. (2008). *For protection and promotion: The design and implementation of effective safety nets*. Washington, DC: The World Bank.

Habicht, J. P., & Pelto, G. H. (2019). Program impact pathways and contexts: A commentary on theoretical issues and research applications to support the EsIAN component of Mexico's conditional cash transfer program. *The Journal of Nutrition, 149*(Supplement_1), 2332S–2340S.

Holmes, R., Akinrimisi B, Morgan, J., & Buck, R. (2012a). *Social protection in Nigeria: Mapping programs and their effectiveness*. The Overseas Development Institute, Accessible: http://socialprotection.org/discover/publications/social-protection-nigeria-mapping-programs-and-their-effectiveness.

Holmes, R., Samson, M., Magoronga, W., Akinrimisi, B., & Morgan, J. (2012b) *The potential for cash transfers in Nigeria*. ODI/UNICEF Nigeria.

ILO (2016). *Inclusive business practices in Africa's extractive industries, ILO Policy Notes*. International Labour Organization, Geneva. http://www.ilo.org/global/topics/employment-promotion/multinationalenterprises/WCMS_449662/lang--en/index.htm

ILO. (2018). *Innovative approaches for ensuring universal social protection for the future of work*. ILO.

ILO. (2019). *Work for a brighter future. Global commission on the future of work*. ILO.

ILO. (2021). *World social protection report 2020–22: Regional companion report for Central and Eastern Europe and Central Asia*. ILO.

Joseph, R. A. (2014). *Democracy and prebendal politics in Nigeria* (Vol. 56). Cambridge University Press.

Krainov, G. N., Rudneva, S. E., & Fedyakin, A. V. (2021). International labour organization and the future of the world of work. *European Proceedings of Social and Behavioural Sciences*. https://doi.org/10.15405/epsbs.2021.11.118

Krueger, A. O. (1993). *Political economy of policy reform in developing countries*. MIT Press.

Lain, J., & Vishwanath, T. (2021). *The COVID-19 crisis in Nigeria: What's happening to welfare?* New data call for expanded social protection in Africa's most populous country. Retrieved August 13, 2021, from https://blogs.worldbank.org/en/africacan/COVID-19-crisis-nigeria-whats-happening-welfare-new-data-call-expanded-social-protection

MacAuslan, I., & Riemenschneider, N. (2011). Richer but resented: What do cash transfers do to social relations? *IDS Bulletin, 42*(6), 60–66.

Morlachetti, A. (2016). *The rights to social protection and adequate food: Human rights-based frameworks for social protection in the context of realizing the right to food and the need for legal underpinnings*. FAO. http://www.fao.org/3/a-i5321e.pdf

National Social Safety-Net Coordinating Office (NASSCO). (2019). *Program implementation manual*. (Third Review). https://nassp.gov.ng/national-cash-transfer-program/

NBS (2019). *Poverty and Inequality in Nigeria 2019: Executive Summary*. https://www.nigerianstat.gov.ng/elibrary/read/1092

NBS (2021). *Nigeria Multidimensional Poverty Index Survey 2021*. https://microdata.nigerianstat.gov.ng/index.php/catalog/71

Nigerian Bureau of Statistics –NBS. (2022). https://nigerianstat.gov.ng/elibrary/read/1241254

Nwoba, M. O. E. (2011). *Rudiments of Social Welfare Administration, the Nigerian experience*. Abakaliki De-oasis Communication Publisher 204.

Nwoba, M. O. E. (2015). An evaluation of social services delivered by local governments in Nigeria: A study of Ebonyi State Local Governments Administration (1996-2012). *Journal of Policy and Development Studies, 9*(3), 142–152.

Obeten, U. B., & Isokon, B. E. (2018). Assessment of conditional cash transfer scheme and poverty alleviation among rural poor in Cross River State, Nigeria. *Advances in Social Sciences Research Journal, 5*(8).

Odey, S. A., & Sambe, N. (2019). Assessment of the contribution of N-Power Program to Youth Empowerment in Cross River State, Nigeria. *International Journal of Sociology and Anthropology Research, 5*(4), 1–13.

Oduenyi, C., Ordu, V., & Okoli, U. (2019). Assessing the operational effectiveness of a maternal and child health (MCH) conditional cash transfer pilot programme in Nigeria. *BMC Pregnancy and Childbirth, 19*, 1–12.

OECD. (2019). *Enabling women's economic empowerment new approaches to unpaid care work in developing countries*. Accessed November 29, 2023. https://www.oecd-ilibrary.org/sites/c8eb9246-en/index.html?itemId=/content/component/c8eb9246-en

Ojo, E. O. (2008). Vote buying in Nigeria. In V. Adetula (Ed.), *Money and politics in Nigeria* (pp. 109–122). IFES-Nigeria.

Okoye, U. O. (2013). Trends and challenges of social work practice in Nigeria. In V. E. Cree (Ed.), *Becoming a social worker: Global narratives* (pp. 149–157). Routledge, Taylor and Francis Group.

Onwujekwe, O., Ensor, T., Ogbozor, P., Okeke, C., Ezenwaka, U., Hicks, J. P., & Mirzoev, T. (2020). Was the maternal health cash transfer programme in Nigeria sustainable and cost-effective? *Frontiers in Public Health, 8*, 582072.

Paul, C. (2022). Conditional Cash Transfer (CCT) and national development in Nigeria: Emerging pitfalls and pathways to results. *Journal of Enterprise and Development (JED), 4*(1), 60–76.

Razavi, S., Behrendt, C., Bierbaum, M., Orton, I., & Tessier, L. (2020). Reinvigorating the social contract and strengthening social cohesion: Social protection responses to COVID-19. *International Social Security Review, 73*(3), 55–80.

Samoili, S., Cobo, M. L., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). *AI Watch. Defining artificial intelligence. Towards an operational definition and taxonomy of artificial intelligence*. EUR30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi: https://doi.org/10.2760/382730, JRC118163.

Sepúlveda, M., & Nyst, C. (2012). *The human rights approach to social protection* (pp. 1–72). Ministry for Foreign Affairs of Finland.

Shadare, G. A. (2019). Transformation of social transfer programmes in Nigeria – a political settlement explanation. Working paper published by IPC-IG/Socialprotection.org. Available at: https://www.socialprotection.org/discover/publications/nigerias-social-protection-cusp-transformation

Shadare, G. A. (2020). Conditional cash transfers in Nigeria–an exploratory study (Doctoral dissertation, University of Sheffield).

Slater, R., & McCord, A. (2009). *Social Protection, Rural Development and Food Security: Issues paper on the role of social protection in rural development*. Overseas Development Institute.

Soyemi, J., Soyemi, O. B., & Hammed, M. (2015). Nigeria cashless culture: The open issues. *Nigeria Cashless Culture: The Open Issues, 4*(4), 51–56.

Steinberg, S. H. (1965). Federation of Nigeria. In: S. H. Steinberg (eds) *The statesman's yearbook. The statesman's yearbook* (pp. 491–499, p. 6). Palgrave Macmillan, . Retrieved June 17, 2024, from https://link.springer.com/chapter/10.1057/9780230270954_15#citeas

Stockman, J., Friedman, J., Sundberg, J., & Harris, E. (2022). Predictive analytics using machine learning to identify ART clients at health system level at greatest risk of treatment interruption in Mozambique and Nigeria. JAIDS Journal of Acquired Immune Deficiency Syndromes, 10-1097.

Théodore, F. L., Bonvecchio Arenas, A., García-Guerra, A., García, I. B., Alvarado, R., Rawlinson, C. J., Neufeld, L. M. & Pelto, G. H., (2019). Sociocultural influences on poor nutrition and program utilization of Mexico's conditional cash transfer program. *The Journal of Nutrition, 149*(Supplement_1), 2290S–2301S.

UNDP (2019). The State of Social Assistance in Africa. New York: United Nations Development Programme. New York: UNDP.

United Nations Population Fund- UNPF. (2008). *Culture matters_II Lessons from a legacy of engaging faith-based organizations*. p. 6. Retrieved June 17, 2024, from https://www.unfpa.org/sites/default/files/pub-pdf/Culture_Matter_II.pdf

UN Women (2022). Putting gender equality at the centre of social protection strategies in sub-Saharan Africa: How far have we come? POLICY BRIEF NO. 24.

Wong, K. L., Radovich, E., Owolabi, O. O., Campbell, O. M., Brady, O. J., Lynch, C. A., & Benova, L. (2018). Why not? Understanding the spatial clustering of private facility-based delivery and financial reasons for homebirths in Nigeria. *BMC Health Services Research, 18*, 1–12.

World Bank. (2018). The State of Social Safety Nets 2018. Washington, DC: World Bank.

World Bank. (2019). World Development Report 2019. The changing nature of work. Washington, DC: World Bank.

World Bank. (2022). *A better future for all Nigerians: Nigeria poverty assessment 2022*. Retrieved April 6, 2024, from https://reliefweb.int/report/nigeria/nigeria-poverty-assessment-2022-better-future-all-nigerians

World Economic Forum. (2023). The global risks report 2023. *World Economic Forum*. https://www.weforum.org/reports/global-risks-report-2023

Chapter 10
Potential of Artificial Intelligence in the Assessment of System of Social Integration of Veterans of the Russian-Ukrainian War

Oleksandr Khyzhniak and Jesús M. Siqueiros García

Abstract War veterans in modern Ukraine are a particularly socially vulnerable group whose social integration requires new approaches, including the involvement of contemporary information and communication technologies and artificial intelligence (AI). Based on the use of "Diia", an AI online application that helps Ukrainians receive social services, the study explores the potential of AI in social work with war veterans and the conditions for its effective application. The author identifies three areas of social work with war veterans: social rehabilitation, social integration, and social assistance. Indicators to measure this potential have been developed, which include the ability of the social work system to remove outdated approaches, forms, and types that cannot be digitized; social workers' proficiency in digital technologies; the ability to implement these technologies in service delivery; and maintaining constructive social work through controlled AI application by professional staff (primarily state and municipal services). Our methods included Participatory Systems Mapping and gamification while working with Ukrainian war refugees. Ukraine is becoming a country of veterans due to prolonged war. It has been proved that the digitalization of veteran services and the use of AI in social work with veterans should be controlled processes with a transparent algorithm and control system.

O. Khyzhniak (✉)
Centre for International Cooperation, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: o.khyzhniak@vu.nl

J. M. Siqueiros García
IIMAS Unidad Mérida, Universidad Nacional Autónoma de México, Ucú, Yucatán, México
e-mail: jmario.siqueiros@iimas.unam.mx

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_10

Ukraine Case Study Introduction and Focus

Introduction to the Current Situation in Ukraine

In Ukraine, among the numerous social groups that are deprived of the opportunity to fully integrate into community life due to their vulnerable situation, war veterans occupy a special place as defenders of the state and Ukrainian society as a result of the direct and indirect use of the armed forces by the Russian Federation against the sovereignty and territorial integrity of Ukraine as an independent state and UN member. Ukraine has been suffering from Russian aggression since 2014. First, the Anti-Terrorist Operation (ATO) was conducted—a set of military and special organizational and legal measures of Ukrainian law enforcement agencies aimed at counteracting the activities of illegal Russian and pro-Russian armed groups in the war in eastern Ukraine; the operation lasted from 14 April 2014 to 30 April 2018. The ATO was followed by the Joint Forces Operation (JFO) in eastern Ukraine, a set of military and special organizational and legal measures by Ukrainian law enforcement agencies that lasted from 30 April 2018 to 24 February 2022, aimed at countering the activities of illegal Russian and pro-Russian armed groups in the war in eastern Ukraine. Veterans of the war of that period are now being added to the veterans of the war as a result of the full-scale Russian aggression that began on 24 February 2022 and continues to this day. Estimates of the number of war veterans still vary. Thus, according to the Ministry of Economy, the number of veterans after the war will reach approximately 1.7 million, while the Ministry of Veterans calls the probable figure of five million, which includes veterans, members of their families, and families of fallen soldiers (Veterans, 2024).

For the state social protection system, war veterans are a special group whose social integration has required new approaches since 2014. And in the digital age, it requires the use of modern information and communication technologies and artificial intelligence.

Ukraine is the country with active and positive digitalization path: in 2019, the new Ministry of Digital Transformation team established four strategic goals through 2024 for building a digital state:

- 100% of government services are online.
- Six million Ukrainians participate in the digital skills development programme; IT represents 10% share of the country's GDP.
- 95% of the transport infrastructure and settlements are covered with high-speed Internet.

Due to war, according to different evaluations, Ukraine's population decreased by 7.3 million (−16.8%) between 2022 and 2023 with the total population of 36.07 million in January 2023. In 2023, there were 28.57 million Internet users in Ukraine at the start of 2023, when Internet penetration stood at 79.2%. Ukraine was home to 26.70 million social media users in January 2023, equating to 74.0% of the total population. A total of 55.88 million cellular mobile connections were active in

Ukraine in early 2023, with this figure equivalent to 154.9% of the total population (Digital 2023: Ukraine, n.d.). So, we could state that despite war digitalization process didn't stop, but rather provided new opportunities for the recovery and organization of life in a new condition.

In such situations over the past 2 years, digitalization has emerged as the focal point of Ukraine's agenda and a top priority for the government.

The purpose of the study is to unlock the potential of artificial intelligence in social work with war veterans and create conditions for its effective use.

The key issues considered by the author include artificial intelligence and services for war veterans, medical services for war veterans, veteran's assistant as an innovation in social support for war veterans, risks of using artificial intelligence in social work with war veterans, independent centres of forensic diagnostics in the system of service provision to war veterans as an innovation, and social responsibility of society for war veterans.

Theoretical basis. We define the potential of artificial intelligence as its ability to bring the system of social work with war veterans to a new level. We are talking about the following triad (Fig. 10.1).

We propose to fill the elements that make up the identified triad that was developed by the authors with the content of services that help meet the needs of war veterans.

- Social rehabilitation—medical services (treatment, prosthetics)
- Social integration—educational services (professional training, advanced training), employment, social support for veterans' families
- Social assistance—provision of the full range of benefits and types of assistance provided by law

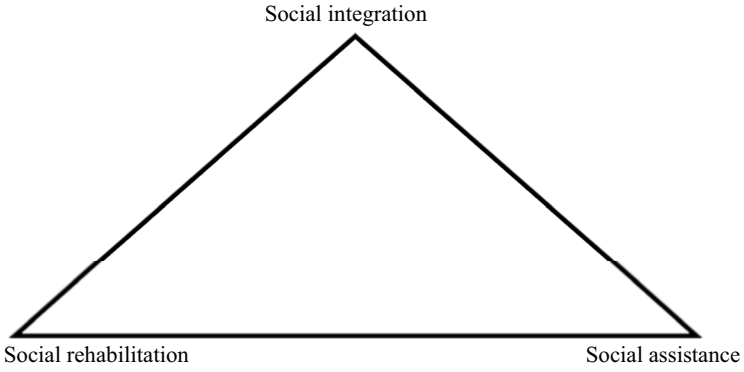

Fig. 10.1 Triad of areas of social assistance to war veterans

Case Study Research Focus and Questions

Ukrainian case studies look at social assessment practices with value reference frameworks investigating the context-dependent service provision selectivity of the current welfare systems in the situation of war. Primarily we focus on one of the newly emerged groups as war veterans. We study all possibilities and consequences of AI use supporting practices, but with the specific focus on employability potential. As the key AI tool for analyses we took online digital application Diia (State Enterprise DIIA, n.d.). In this section, we will describe general requests for artificial intelligence in services for war veterans; based on existing empirical studies offer the generalized "War Veteran Image" in Ukraine; focus on the challenges of medical assistance (in broader sense, including mental health); describe current veteran (ex-combatants) assistance requests and key services that are provided; and present the possibilities of digital application Diia as an AI tool in providing social services with focus on implementing the new approach for employability-related issues for veterans (ex-combatants).

Requests for Artificial Intelligence in Services for War Veterans

Surveys conducted in Ukraine since 2014 (Аналіз системи соціального захисту ветеранів та військовослужбовців, 2022), with the support of international organizations, allow us to assess the dynamics of war veterans' needs and the state of their satisfaction. At the same time, it is advisable to consider the dynamics of the needs of war veterans and the value differences between the perceptions of the problems and needs of veterans by members of the armed forces, ordinary citizens, and veterans themselves, depending on their length of service.

To implement these areas, it is worth applying the innovative potential of AI, which we propose to measure through the following indicators: (1) the ability of the system of social work with war veterans to get rid of outdated approaches, forms, and types that cannot be digitized; (2) the ability to perceive new information and communication technologies of social work, which implies that social workers are proficient in digital technologies; (3) the ability to implement these technologies in the practice of service delivery; (4) the ability to retain the constructive in social work due to the manageability of the process of AI application by professional staff of the relevant services (primarily state as well as municipal at the level of territorial communities where a particular veteran recipient of the post-employment benefits lives).

We would like to add that the above indicators will be useful in determining the effectiveness of the use of artificial intelligence in social work with war veterans.

Social protection of veterans and disabled veterans of the Russian-Ukrainian war of 2014–2023 is provided by various official mechanisms, from the legislative level to the civilian level. Even on 11 March 2020, the Ministry of Veterans Affairs of Ukraine was established.

However, the effectiveness of this work with veterans and disabled veterans of the Russian-Ukrainian war of 2014–2023 is determined by the efforts of many central government agencies and local governments. These are dozens of institutions: 11 ministries, 18 services and committees at the central level, and relevant structures at the local level. Their work is largely regulated by the current legislation of Ukraine, which currently includes more than 12,000 laws. Each law requires the executive branch to comply with 2–3 to 20–30 bylaws of different territorial levels—resolutions, orders, instructions, etc.

Out of the total number of laws in Ukraine, almost half of them relate to the social protection of veterans and disabled veterans of the Russian-Ukrainian war of 2014–2023. These are issues related to the processing of various documents confirming the status of a participant, veteran, or disabled person, housing and land, veteran entrepreneurship; treatment and rehabilitation, social and professional adaptation, guarantees and benefits, veteran development centres, sports, paralympic competitions, etc.

It is difficult for an ordinary official in Ukraine, either alone or together with like-minded people, to make sense of this maelstrom of information that is growing every month—there will be mistakes, partial and/or complete non-fulfilment, conflict of laws, corruption, and dissatisfaction of Ukraine's defenders and their relatives with the results of social rehabilitation.

Conclusion. At this stage, artificial intelligence (hereinafter referred to as AI) can objectively and effectively eliminate the consequences of the human factor. In Ukraine, it is not yet widely used in this context.

"The War Veteran Image" in Ukraine

Research on the problems of veterans and military personnel has been conducted in Ukraine since 2014 (Аналіз системи соціального захисту ветеранів та військовослужбовців, 2022). However, their authors noted the difficulties of evaluating the quality and accessibility of services, as there were no regulatory grounds for the introduction of a veteran's electronic cabinet for accessing public services online, and regulations for the introduction of an electronic veteran's certificate were still under development. The lack of official statistical information was also a limitation (Аналіз системи соціального захисту ветеранів та військовослужбовців, 2022). Under these conditions, there was no application of AI in social work with war veterans.

The author used the materials of sociological surveys as empirical information, namely:

- The twentieth nationwide survey. The image of veterans in Ukrainian society (14–16 January 2023). The survey was conducted by the Sociological Group Rating on the initiative of the Ukrainian Veterans Fund of the Ministry of Veterans Affairs of Ukraine on 14–16 January 2023. The survey was conducted among the population of Ukraine aged 18 and older in all oblasts, except for the temporarily

occupied territories of Crimea and Donbas, as well as the territories where there was no Ukrainian mobile coverage at the time of the survey. The results are weighted using the latest data from the State Statistics Service of Ukraine. The sample is representative in terms of age, gender, and settlement type. Sample population: 1000 respondents. Survey method: CATI (Computer-Assisted Telephone Interviews). The error of representativeness of the survey with a confidence level of 0.95: no more than 3.1% ("Образ ветеранів в українському суспільстві).

- The second anonymous online survey among veterans and active military personnel "Portrait of a Veteran. Veterans' Needs" was conducted on 6–12 February 2023, and involved the collection of primary data on the needs of current and future female and male veterans. The tool was an anonymous online survey, so there were no sampling requirements. The survey involved 1247 respondents aged 18 and over (Друге анонімне онлайн-опитування серед ветеранів та діючих військовослужбовців «Портрет ветерана. Блок "Потреби ветеранів", 2023, n.d.).
- The nationwide survey "Image of Veterans in Ukrainian Society" was conducted by the Ukrainian Veterans Fund of the Ministry of Veterans and the Sociological Group "Rating". Experts conducted a telephone survey of 1000 respondents on 5–7 September 2023. The sample is representative in terms of age, gender, and type of settlement. The error of representativeness of the survey with a confidence level of 0.95: no more than 3.1% (18% of Ukrainians believe that society does not respect veterans, 2023).
- Assessment of access to and satisfaction with medical care in Ukraine by ex-combatants of the ATO/JFO. The study was conducted by Social Consulting with the assistance of the United Nations Development Programme (UNDP) under the UN Recovery and Peacebuilding Programme, with financial support from the European Union and the Government of the Kingdom of the Netherlands, as well as with the expert input of the Ministry of Veterans Affairs of Ukraine, the Ukrainian Healthcare Centre, and the International Charitable Foundation "Ukrainian Foundation for Public Health". This study was conducted in December 2021–January 2022 as part of a project to develop recommendations for the creation of a policy document "National Policy on Medical Care for ATO/JFO Veterans", implemented by the International Charitable Foundation "Ukrainian Foundation for Public Health" with the support of the United Nations Development Programme in Ukraine (Medical care availability and level of satisfaction with services for ex-combatants, 2024, n.d.).

The analysis of the data from the national surveys "Image of Veterans in Ukrainian Society" by the Ukrainian Veterans Fund of the Ministry of Veterans and the Rating Sociological Group allows us to record some points that affect the state and prospects of social work with war veterans.

The attitude of the Ukrainian society towards war veterans is reflected in the following indicators:

1. A high level of trust in veterans, namely, 94% of the polled citizens fully trust and rather trust veterans who will return from the ongoing war (the same number of citizens trust the military of the Armed Forces of Ukraine); ATO veterans who fought in the East of Ukraine in 2014–2021.
2. The vast majority of respondents (91%) believe that Ukrainian society respects war veterans.
3. Citizens show goodwill towards war veterans: 96% are ready to work in the same team with them.
4. Analysts note that there is a growing feeling in society that veterans may be forgotten, that they will face problems that are not being solved, and that this is the reason for the lack of representation of the topic of war veterans in the media and communication space.
5. Lack of awareness of veterans' issues, especially among residents of the centre and south of Ukraine, middle-aged people, and relatives of those who fought from 2014 to 2021 (two-thirds of respondents said they were completely or rather uninformed). Public awareness of veterans' issues has slightly increased from 37% in January 2023 to 53% in September 2023. Respondents learn information about veterans and their problems mainly from television (37%), relatives and friends (37%), and social media: Facebook, Instagram, YouTube, etc. (34%), channels in Telegram or Viber messengers (22%), and online media (19%). Conclusion: Since Ukrainian consumers of information about war veterans are quite active in using digitized data as an information resource, this allows us to use AI to analyse the interest of citizens in the problems of this large socially vulnerable group.

Attitudes of the state towards war veterans: Residents of the western regions, representatives of the younger generation, and those with higher income are more likely to have a negative opinion about the fulfilment of the state's obligations to veterans. In general, the number of people who believe that the state fulfils its obligations to war veterans continues to decline: in August 2022, the number was 69%, in January 2023—53%, and in September 2023—33% (Загальнонаціональне опитування #24 Образ ветеранів в українському суспільстві, n.d).

The situation of war veterans in the transition from military to civilian life. Respondents from among the veterans who did not serve in the Defence Forces of Ukraine noted that they need primarily financial assistance (50.2%), medical care (18.2%), information (8.9%), and psychological assistance (7.1%). A significant number of them said they wanted to start their own business (63.6%). Given that veterans have been changed by the war, the surveyed citizens named the following perspectives: a desire to change their country for the better (64.7%), a value of life and relationships (63%), a change in life priorities (61.5%), a change in life philosophy (54.8%), a circle of friends and like-minded people (46.6%), and a desire to live and develop (45.6%). The most common problems, according to the respondents, are physical health problems (68.4%), psychological disorders (66.4%), lack of understanding of society (63.8%), problems with social benefits (53.6%), difficulties in obtaining medical care (53.2%), difficulties in finding a job (50.4%), and

alcohol or drug addiction (47.9%). “At the same time, unlike the indicators of the survey of the population of Ukraine by the Rating Group, for the surveyed servicemen and veterans, the top 3 most common problems faced by veterans, in addition to physical health problems and psychological disorders, also include lack of understanding of society” (Друге анонімне онлайн-опитування серед ветеранів та діючих військовослужбовців «Портрет ветерана. Блок “Потреби ветеранів”, 2023, n.d.). This indicates that the topic of war veterans is not sufficiently represented in the public communication space (real and virtual) to form an adequate public opinion about this socially vulnerable group with special needs.

Studies also show “an increase in expectations of possible risks, the list of which has changed, in the lives of veterans, including: unemployment (over the year, the percentage increased from 34% to 42%); conflicts in veterans’ families (from 31% to 46%); alcohol and drug abuse (from 31% to 43%); suicide (from 18% to 28%); and violation of laws (from 13% to 23%). Among the most likely risks facing veterans are psycho-emotional instability, physical health problems, difficulties in obtaining medical care, lack of inclusive space and adapted workplaces for people with disabilities” (Образ ветеранів в українському суспільстві: які результати загальнонаціонального опитування, 2024, n.d.).

Medical Services for Ex-combatants

As a result of the protracted armed conflict in eastern Ukraine, a large number of military personnel have suffered physically and mentally. The study of the state of medical services for veterans shows that “The list of services for war veterans is quite wide, and the majority of respondents to the survey are satisfied with the services of their family doctors (3.8 points out of a maximum of 5) and specialised medical care (3.4 points). For the most part, respondents reported no significant obstacles to signing declarations with their family doctor: 80% of the ex-combatants surveyed have them. Some problems were encountered by ex-combatants who, for various reasons, ended up in the combat zone, do not have housing and, accordingly, no registration. More than half of the ex-combatants (62%) visited their family doctor without any obstacles. The rest are more often concerned about ‘long queues to see a doctor’ and ‘lack of free appointments for the next few days’, which is primarily due to the rapid increase in COVID-19 cases (the survey was conducted in December 2021–January 2022). At the same time, according to ex-combatants, they rarely seek advice from family doctors (more often they seek referrals for examinations and consultations with other specialists). And this is despite doctors’ assessments that most ATO participants have neurological, cardiovascular, musculoskeletal and psychological problems. In addition, 40% of the surveyed ex-combatants have disabilities. The inability to get to a family doctor promptly and the lack of specialists significantly slows down the time for ex-combatants to pass medical commissions necessary for receiving sanatorium treatment, confirmation of disability, etc.” (Демченко et al., 2022, p. 77).

The problem of territorial inaccessibility is also relevant: access to medical services for ex-combatants living in rural areas or small towns is significantly worse. As for prosthetic services, only a quarter of the doctors surveyed (22 out of 101) consider them accessible to ATO/JFO ex-combatants, and they are not always of high quality (9 out of 22). These services are considered inaccessible due to complicated registration procedures, long queues, etc. (Демченко et al., 2022, pp. 78–79).

Among war veterans, some need special assistance with rehabilitation and prosthetics. The results of the survey on the quality and accessibility of state programmes for rehabilitation and prosthetics for people with disabilities are less revealing, as the vast majority (55%) chose the option "difficult to answer". Such programmes are considered to be of high quality but difficult to access by 22% of respondents, and not of high quality but accessible by 6.5%. At the same time, according to almost 15% of respondents, state programmes for rehabilitation and prosthetics for persons with disabilities are of poor quality and difficult to access, and only 1.8% of respondents consider them to be of good quality and accessible (Друге анонімне онлайн-опитування серед ветеранів та діючих військовослужбовців «Портрет ветерана. Блок "Потреби ветеранів" 6–12 лютого 2023, n.d.).

Social protection of veterans and disabled veterans of the Russian-Ukrainian war of 2014–2023 will be most effective when they do not lose a single day or hryvnia (the currency of Ukraine) to realize and receive the full range of medical care and rehabilitation promised by the state, both in Ukraine and abroad, when they and their relatives do not partially or fully spend money (depending on the specific circumstances, the salary of a defender can be from 80,000 to 150–200,00 UAH per month).

With the continuation of war the number of wounded and disabled is growing. According to the State Statistics Service and the Ministry of Veterans Affairs, as of March 2023, 2.7 million civilians and almost 500,000 combatants in Ukraine had disabilities. The number of people with disabilities is growing because of the war.

After diagnostic, therapeutic, and rehabilitation measures, if there is information confirming a persistent impairment of the body's functions that cause disability, medical advisory commissions give a referral to the Medical and Social Expert Commission (MSEC) to establish disability (Коли дають групу інвалідності довічно, 2023).

When a Disability Group Is Granted for Life

Many medical documents are required to establish the cause of the deterioration in health, the nature of the injury, and its causal relationship to a particular event (coincidence of place, date, type of injury—shrapnel, mine, gunshot). Such medical documents are drawn up by different levels of military personnel—combat medic, stabilization centre, waiting for evacuation, specialized surgical unit, highly specialized hospital, and further treatment and rehabilitation. In addition, you need a report from the commander that the injury was sustained in combat with the enemy. A very

important decision is the decision of the Military Qualification Commission on fitness/restricted fitness or complete unfitness for military service.

All these stages, especially treatment and rehabilitation, take place in different medical institutions. Therefore, medicine is accompanied by bureaucratic red tape, as not every stage has an electronic version of medical records and a local database.

Conclusion. Artificial intelligence will be useful for searching and combining veterans' medical information, both in databases (state or local) and searching for medical information on social media. In addition, AI can receive and process medical information from the veteran himself. In addition, AI can receive and process medical information from publicly available medical records and from the veteran himself if it is posted on social media.

Veteran's Assistant as an Innovation in Social Support for War Veterans

The research records the following most common questions from war veterans:

- Appealing against the actions of social protection authorities regarding the refusal to transfer a one-time financial assistance, including as a family member of a deceased combatant.
- Appealing against actions of the Pension Fund of Ukraine regarding the refusal to recalculate a pension and the obligation to take certain actions.
- Appealing against a decision to remove a serviceman from the housing register and restore his housing rights.
- Appealing against the decision of the military medical commission.
- Divorce proceedings.
- Recovery of the unpaid amount of the salary indexation.
- Appealing against an order to bring to disciplinary responsibility.
- Recovery of monetary compensation for unused days of additional leave for a combatant.
- Resolution of disputes in the field of civil law, family, administrative, housing law, social security (З початком широкомасштабної війни з'явилися нові суб'єкти права на безоплатну вторинну правову допомогу, 2023). Of course, it is difficult for many veterans to solve the issues of their livelihood and defend their legal rights without assistance.

That is why Ukraine is implementing a project to train a war veteran assistant to provide professional social services as part of the mechanism for implementing the state policy on veterans and other relevant categories of citizens announced by the Government in early 2023. Four key components have been identified: economic independence through successful professional fulfilment; the possibility of further military career; provision of housing for veterans; and medical (rehabilitation) services. Additionally was underlinned the necessety for improvement of external and internal communications that are provided by the service office for veterans' affairs

as a legal entity under public law in accordance with the Law on Cooperation of Territorial Communities.

On 1 July 2023, the pilot project "Institute of Veteran's Assistant" was launched in four regions of Ukraine, and on 10 October 2023, the project "Institute of Veteran's Assistant" was expanded to six more regions. The main task of the veteran's assistant is to provide comprehensive assistance to Ukrainian defenders during their transition from military service to civilian life, after returning from the front and serving at home. The veteran's assistant will work in the community. It is envisaged that one assistant will be responsible for helping 100 veteran families. As early as 2024, 15,000 veterans' assistants will start their work across Ukraine (Адаптація ветеранів війни до мирного життя: помічник ветерана, 2023).

Currently, the Ministry of Veterans Affairs of Ukraine is planning to introduce a new area of activity—the establishment of a service office for veterans in territorial communities based on the results of the pilot project and the introduction of a new professional group—a veteran's assistant. The institute of a veteran's assistant is being introduced in territorial communities as part of the system of transition from military service to civilian life. The pilot project was launched on 1 July 2023 in Lviv, Vinnytsia, Dnipro, and Mykolaiv regions. On 10 October 2023, these regions were joined by Zakarpattia, Kyiv, Sumy, Poltava, and Kharkiv regions and the city of Kyiv.

Based on the results of the pilot project, summarized by the Ministry of Veterans, it is planned to scale up the institute of a veteran's assistant to the entire territory of Ukraine in 2024/25 (Помічник ветерана: дорожня карта для громад, кандидатів та родин Захисників і Захисниць, n.d.). The training of a war veteran's assistant involves the following tasks:

- To ensure the development of general and professional competences of a veteran's assistant to provide consultative and advisory assistance to veterans
- To equip veterans' assistants with techniques, methods, and tools to assist in the communication of war veterans with the relevant authorities (state and public sectors) to comprehensively assist this category of socially vulnerable citizens in full adaptation to peaceful life: from establishing everyday life and family understanding to employment, implementation of own business projects, learning new skills, rehabilitation and treatment, legal protection
- To facilitate a dialogue between a particular veteran and his/her community of residence, labour collective, the state, and NGOs in general
- To organize communication support for professional care for family members of veterans, families of deceased war veterans, and families of deceased defenders of Ukraine.

Conclusion. Artificial intelligence will be useful both for selecting candidates for training as an assistant to a war veteran and for selecting services to help a particular veteran and a particular stage of his/her life or challenging life situation.

Digital Application Diia in Providing Social Services as the Example of Using AI Tools in Providing Social Services for Ex-combatants

For almost 4 years now, Diia has been proving that the State in a Smartphone is not a distant future prospect, but a new reality. Diia has become pop culture and has taken an important place in the lives of Ukrainians: the app is used by more than 20 million citizens. This is every second resident of our country (State in a Smartphone: More than 20 million Ukrainians use Diia, 2024, n.d.). Diia has fundamentally altered the dynamics of interaction between the government and its citizens. Through this application, the state has transformed into a user-friendly service accessible to all with just a few clicks from the comfort of their homes. Presently, the app boasts 14 digital documents and over 30 electronic services, catering to the daily needs of millions of Ukrainians across a myriad of life scenarios.

Ex-combatants as a specific target group have two main forms to receive e-services via Diia (State enterprise DIIA, 2024, n.d.): (1) as citizens they have access to more than 70 services and (2) ex-combatant for the specific target-oriented services—Veteran Business. The last mentioned service oriented on veterans, their families, and the families of fallen defenders, providing comprehensive information and support for veteran entrepreneurship endeavours. Option Veteran Business is the opportunity for ex-combatants to avoid unemployment. Key opportunities and services include:

- Free education for veterans and their families. Returning from war poses unique challenges for veterans, both mentally and physically. The Ukrainian Veterans' Fund, in collaboration with UCU, offers an online course designed to address common obstacles faced by military personnel transitioning to civilian life. Topics covered include psychological adaptation (such as managing emotional states and PTSD) and physical rehabilitation, including prosthetics. Additionally, participants will gain insights into legal aspects related to obtaining the status of a combatant and explore economic development opportunities, including planning and launching their own businesses.
- Finding employment after military service: Combatants, veterans, and women veterans can benefit from an online educational series available on the Diia. Education portal. This series equips individuals with essential tools for effective career guidance, resume writing, job search strategies, and interview preparation, ensuring a smooth transition into the civilian workforce.
- Free online entrepreneurship school at Diia.Business: Entrepreneurs can access a wealth of resources through the Diia.Business platform. From setting up a company and maintaining proper accounting records to developing a financial plan, securing funding, marketing products, and managing teams, all aspects of entrepreneurship are covered in the format of educational series, provided free of charge.
- Ideas for launching your own business. For ATO/JFO participants, veterans, and women veterans seeking to start a business with minimal startup costs, this

section offers a range of business ideas tailored to their unique experiences and circumstances.

Risks of Using Artificial Intelligence in Social Work with War Veterans

Ukrainian developers from the Mantis Analytics team have created a unique AI-based platform that detects hostile fakes on the web and warns of information and psychological operations (IPO). Mantis Analytics, as an AI platform for monitoring the information field, collects information from the media, social networks, and other sources and processes thousands of messages and gigabytes of data from the information space—media, social networks, information platforms—in real time to identify false information. The platform then plots everything on an interactive map.

AI analyzes the information and makes it clear who is reporting what and how. The platform uses NLP and LLM models to analyse huge amounts of unstructured data quickly and efficiently. According to Mantis Analytics developers, AI can:

- Recognize propaganda, fakes, and disinformation
- Predict the future based on open data
- Collect statistics and generate reports
- Analyse trends

However, the use of AI in the medical field has a flexible conflict (Винен штучний інтелект. Австралійські лікарі потрапили у гучний скандал, 2023). In social work with veterans it can be useful. For example, there is evidence that AI can help clean up the veterans' community by identifying those who receive social assistance illegally or claim to do so through various registers and databases ("Штучний інтелект" допоможе виявити тих, хто отримує соціальну підтримку незаконно, 2019).

Empirical Social Research in AI FORA

Concepts and Methodologies Used to Investigate Research Question(s)

Due to the current war crisis in Ukraine, it is practically impossible to conduct workshops and participatory design sessions with war veterans. This is the reason why we decided to hold a workshop with Ukrainian refugees to understand their perception of DIIA as a pathway to access public unemployment services. This workshop also allowed us to identify values and morals regarding decision-making based on a digital system for the distribution of public services.

Our primary methods included Participatory Systems Modelling (PSM) and Gamification. Participatory Systems Mapping (PSM) (Barbrook-Johnson & Penn, 2022) is a participatory modelling method to create networked maps focused on a concrete issue, detailing the elements affecting it, the causal relationships between them, and the strength of these causal relationships. The causal map provides information about the interaction of stakeholders and the system, as well as about the system and the stakeholders creating the map. This tool focuses on exploring complex issues.

Gamification is a method to frame topics and problems of interest in the format of a game. Games usually represent situations, environments, and systems in which stakeholders must make decisions. Gamification has the advantage of setting up an engaging environment where everyone can play. Since playing is not demanding regarding abilities, this method offers a context where individual and power unbalances are left behind (Szczepanska et al., 2022). Gamification is often used to obtain information about people's behaviour or to understand the system of interest better and can be applied together with other methodologies, such as Participatory Systems Mapping or Agent-Based Models (Barreteau et al., 2003; Szczepanska et al., 2022).

Workshop

A workshop was held in November 2022, bringing together 24 Ukrainian refugee citizens between 15 and 61 years old. Due to the ongoing Russian invasion of Ukraine, participants were mostly women with children. The participants signed an informed consent document to collect audio and photos so they could refuse any practice that could make them uncomfortable.

A 3-day workshop was held to collect information about the public services provided through the Diia mobile app (https://diia.gov.ua). The research team included a native Ukrainian speaker, one of this chapter's authors. The partner helped to translate whenever needed and was familiar with the Diia app and the current situation in Ukraine.

The Participatory Systems Mapping (PSM) session began with an ice-breaking activity: the participants were asked to write down the Diia services they were familiar with. This activity also helped the participants to refer to their own experience and knowledge regarding Ukrainian public services, offline and online. The PSM session aimed to identify and analyse the conditions affecting fairness in public services provision via Diia. The research team proposed the concept of fairness as a working concept that participants had to elaborate on with their own meaning. Then, the participants were asked about the factors determining the fairness of public services provision through Diia performance and the relative impact of these factors in determining fairness. Participants clustered factors into personal attributes, government resources, geographical location, and technical issues. Participants gave weight to each factor to represent how much they affect fairness. Then, they placed the factors with a perceived higher impact closer to the centre of

a whiteboard. The lower the impact perceived, the farther away from the centre they would be.

The last task to be carried out during the session was selecting one of the Diia services mentioned at the beginning to work with the next day in the gamification session. Participants voted among the services listed in the ice-breaking activity. They chose to work with the employment service since they had the most experience in this kind of social service provision. At the end of the PSM session, the research team collected a list of experienced social services deployed by Diia, the PSM about the fairness of the Diia, the clustering of factors, and the employment service for the gamification session.

The game's design is the step where the PSM and gamification come together. The research team conducted this step based on the inputs collected in the PSM session.

The elements of this game are agents with specific fixed attributes and abilities that can be improved, the stations where the agents can go and do actions in each round, and the tokens provided to the agents, which can be wealth, money, or particular abilities' units. The attributes and abilities are featured according to each case study. The game includes six stations. (1) Home, where players can go or stay in each round and where discussions are allowed among the participants; (2) the City Council, where proposals can be handled to be voted by the participants; (3) the Public Service Institution, where a particular public benefit (always in limited quantity) is provided; (4) the Training Centre to improve agents' abilities; (5) the Holidays Resort where the agents can improve their happiness and; and (6) the Station where the public benefit of the game can be used. Also, an algorithm is part of the structure and has to be designed to set the rule by which the agents get the game benefits.

Based on this structure, a game for this case study was designed and focused on employment benefits. In the first place, the station "Employment Agency" was settled as delivering a low number of good jobs, unlimited bad jobs, and no jobs at all. The allocation of jobs was decided based on an algorithm. The station "Workplace" was settled as the one where the public benefits could be used (in this case, working with the job achieved). The agents' attributes were settled based on the PSM factors clustered: gender (man or woman), ethnicity (Ukrainian or not), household composition (single or not), having children or not, previous residence in Ukraine (in occupied, military operation or Western territory), and former history of the agent (with a good track record or not). Education level and networking abilities were included as abilities which could be improved. Each attribute gets one or no points and their sum gives the player a score. An "algorithm" was devised based on this table of attributes. The algorithm is a chart with limit values to assess the score of the player and decide whether the player gets access to a good job, a bad job, or cannot access any job. The values of the algorithm are set at the beginning of the game, and players do not have access to it (for a description, see Table 10.1).

The agent's attributes and abilities and their level of happiness were included in the "algorithm" as features of the agents to determine the kind of job they could achieve at the employment office.

Table 10.1 Algorithm: initial *vs* final weighting during gamification

Features	Initial weighting	Final weighting
Gender		
Male	1	1
Female	0	0
Household composition		
Single	1	1
Not single	0	0
Ethnicity		
Ukraine	*1*	*0*
Else	*0*	*1*
Children		
No kids	*1*	*0*
Kids	*0*	*1*
Previous residence		
Western Ukraine Terr.	*2*	*1*
Military Operation Terr.	*1*	*1*
Occupied Terr.	*0*	*1*
Former history		
Good track record	1	1
No track record	0	0
Education units	***(each 10 un)***	***(each 5 un)***
Yes	1	1
No	0	0
Networking units	***(each 3 un)***	***(each 1 un)***
Yes	1	1
No	0	0
Happiness	*Required to have a job*	–
Good job	> 7	> 7
Bad job	*5 – 7*	*3 – 7*
No job	*0 – 4*	*0 – 2*

In the last session, the game was run with the participants. Three different rounds were played, in which the participants could implement changes in the algorithm.

The dynamics of the game were as follows. At the beginning of the game, each agent had the same units of wealth, happiness, education, and networking. Some bad and good jobs were randomly distributed among the participants and there were also unemployed participants. The goal of each player is to maximize their wealth and happiness units. In each round of the game, the participants decided in which station they would spend the round. Only employed participants could attend the Workplace, where they earned wealth and networking units. Those with a bad job or no job could try to improve their situation at the Employment Agency, where they could become a new job depending on their scoring. Those who chose the Holidays Resort spent money but won happiness. Staying at home provided happiness only

for one round. Those who chose to go to the City Council could propose new game rules or changes in the algorithm and earned networking units. Those who went to the training centre earned education units. At the end of each round, each participant could donate wealth units for the community, which could be used if, after deliberation at the end of each round, they decided to change the game rules, the table of attributes, or the algorithm, which had a cost in terms of wealth.

The results obtained include the decisions made by each individual in the rounds and the collective behaviour captured (the discussions and the modifications agreed upon in the algorithm). Regarding the definition of the problem assessed, the modifications agreed upon in the algorithm are the most relevant result since they capture the participants' perspectives regarding values.

Co-design and participatory methods are pivotal for mitigating AI-related injustices. In the AI FORA project, a workshop convened 24 Ukrainian refugee participants, focusing on public services accessed through the Diia app. Facilitated by a native Ukrainian speaker, the workshop employed Participatory Systems Mapping (PSM) to explore fairness factors in service delivery. Participants prioritized the employment service for further examination during a subsequent gamification session. This session integrated insights from the PSM exercise. The game structure encompassed agents, stations, and tokens, tailored to specific case studies and governed by an allocation algorithm. The structure featured six key stations: Home, City Council, Public Service Institution, Training Centre, Holidays Resort, and a Station for utilizing public benefits. An independent Ethics Committee approved the project, ensuring adherence to ethical standards. Over 3 days, participants engaged in discussions, activities, and decision-making processes, fostering a collaborative approach to addressing AI risks. Through their involvement, the workshop aimed to amplify the voices and perspectives of marginalized communities, contributing to more equitable AI development and deployment.

Presentation of the Results Research

After running the workshop, we ended up with three main results: the PSM about the fairness of the Diia app and the clustering of the factors involved in this fairness, a game based on the employment service offered by the Diia app and framed through the factors of the PSM, and the results of the gamification run with the participants.

First, using a PSM, participants define their perception of the system to be analysed. The PSM contains the factors in which participants cluster in groups. This reflects what participants considered to be affecting the performance of the Diia. According to the PSM results, the participants consider that the fairness of the Diia app performance is affected by the weight the government assigns to the personal attributes of the applicants, the government resources, the geographical location of the beneficiaries, and other technical issues. Then, the interrelation between the PSM and game, inserting these factors as the main features of the gamification run

with the participants, we can somehow "put in practice" this definition and, through the possibility of changing the algorithm, the participants give information not only about how they perceive the system but also about how their desired system to be.

The original version of the algorithm and its final version allow us to track these changes to understand the participants' problem definition deeply. The original algorithm privileged Ukrainian single men with studies and networks. The game allowed participants to experiment with the system and discuss the results after each round. Participants then accessed the algorithm and profile chart, debated in the City Council station, and agreed to modify both their weights and content so that the expected outcome aligns with their view on what is fair.

In contrast to the original algorithm, participants changed the algorithm favoured single parents and non-Ukrainians and equated the weights for the participants' origin (in other words, in the selection process, it becomes irrelevant). Participants also modified the weights of the algorithm to lower the minimum for getting a "bad job", therefore widening the area for a "bad job" and narrowing the space for "no job". Participants changed the algorithm through a democratic majority voting as they were very much interested in the best situation for all. Moreover, they split the amount to be collected for the public fund to change the algorithm, and everyone had to donate the same.

Overall, the resulting modifications to the profile chart and algorithm suggest favouring people's families broken in a way by war (single parents) and also suggest that, as refugees, they may come from different parts of Ukraine, so their interest was in eliminating the relevance of their place of origin. Furthermore, perhaps giving a higher value to not being Ukrainian might be related to acknowledging ethnic diversity. Finally, lowering the minimum to get a "bad job" responds to the logic of "better to have a bad job than having no job at all" when conditions to improve changes to get a job are currently so hostile. All this is speculation and must be confirmed after a proper qualitative data analysis. Moreover, the participants' opinions and views depend on their current context and experience as refugees coming from the eastern and southern areas of the country, which are occupied territories and particularly unsafe. Therefore, the insights obtained during the workshop do not represent those of the Ukrainian people.

The game designed for this case study focuses on employment benefits, utilizing stations like the "Employment Agency" and "Workplace", with job allocation determined by an algorithm. Agents' attributes, including gender, ethnicity, and household composition, are incorporated, alongside education and networking abilities. An algorithm charts limit values to assess player scores, deciding job access. During gameplay, participants make algorithm changes across three rounds, aiming to maximize wealth and happiness units. Stations like the "City Council" and "Training Centre" allow for proposal and education unit earning. After the workshop, three main outcomes emerged: a fairness assessment of the Diia app via Participatory Systems Mapping (PSM), a game reflecting employment service, and gamification results. PSM clusters factors affecting Diia's performance, highlighting participant perceptions on fairness. The game integrates these factors, allowing participants to experiment and modify the algorithm. Algorithm changes favour single parents and

non-Ukrainians, reflecting participants' fairness views. Modifications suggest favouring families affected by war and acknowledging ethnic diversity. The analysis process remains ongoing, emphasizing the importance of understanding participants' motivations and experiences. The methods applied allow vulnerable groups to express values crucial for software design.

Discussion of Results

According to the results presented, we can state that participatory modelling, namely the combination between PSM and gamification, allows us to obtain information about the values and conceptions of the participants involved. Participatory modelling is a powerful tool to include vulnerable groups as epistemic agents in AI development and design. Participation can improve public services provision by defining the problems to handle with AI and providing information about the desired state of the situation. Having the participants in the process of problem formulation allows us to conclude that the inclusion of vulnerable groups through participatory modelling can foster the alignment of AI technologies applied to public services provision. Since vulnerable groups are often the target of public services, a good alignment of AI will only be possible if the needs of vulnerable groups are considered and not only the needs of bureaucratic administration or the computer scientists who design an algorithm with only good technical performance in mind. Additionally, the inclusion of the other stakeholders' perspectives should be included to foster a better alignment between performance and social values.

As was said above, these are only preliminary results as qualitative data analysis is undergoing. A deeper analysis of the participants' rationales for change factors and the algorithm could bring relevant insights into how their perspectives can improve AI alignment through the participatory modelling methodology. Finally, we implement agent-based modelling (ABM) to our methodology pipeline. By implementing games in an ABM simulation, it is possible to run several hundred rounds, which is impossible to do with real players/participants. Moreover, we can use data from real players/participants to seed it into the simulation, and we could test the long-term effects of the decisions made by the players in the distribution of benefits and later discuss the results with stakeholders and validate the results or introduce further changes.

Society's Responsibility for War Veterans

The responsibility of society for war veterans has been implemented since the beginning of the Russian-Ukrainian war in civil society projects, such as the ATO Veterans Society (NGO registered on 24.09.2014).

The Ukrainian Veterans' Fund of the Ministry of Veterans is already providing financial support to veteran businesses through micro- and macro-financing programmes. Experts believe that the experience of the USA, where almost every second veteran became a business owner or director after World War II, is a promising scenario, as more than 60% of veterans would like to have their own business, which is important to consider when reintegrating the military (Калмикова, 2023).

The After Service Foundation has positive experience of working with veterans. After Service was founded by Edward Marshall, an American of Ukrainian descent, at the end of 2022. The foundation adopts the American experience, adapts it to Ukrainian realities, provides treatment and rehabilitation (combat injuries), psychological support, legal support and access to benefits, assistance in finding employment and creating new career opportunities, and payment for educational courses, and cooperates with volunteers and veterans' organizations across the country. The Women's Veterans Movement is gaining popularity in Ukraine (Каразуб, 2023).

Conclusions

According to all-Ukrainian research, the factors that actualize changes in social work with war veterans include the presence, along with the public's respect for veterans, of growing dissatisfaction with the state's activities in this area. The trend in the digital era is to provide social assistance services based on the latest information and communication technologies and digitalization of services to socially vulnerable groups, including war veterans. There is a widespread opinion among experts that Ukraine is becoming a country of veterans due to the ongoing hostilities. The state is obliged to fulfil its social function, but citizens also have a duty to veterans. This leads to the realization that veterans' issues should not be perceived as a problem, but as a task not only for the state but for the entire indestructible Ukrainian society.

Social work with war veterans has specific features: first, its task is to ensure the transition from military service to civilian life, which requires, first, the development and implementation of various social services for those who plan to return to the security forces (national guard, police, border guards, armed forces) and those who seek employment in civilian life; second, taking into account the health status (especially disability) of veterans; third, the identification of target groups of veterans in each territorial community according to their socio-demographic composition (primarily by age, gender, level of education, marital status, place of settlement, housing, and livelihood); fourth, the consideration that some information about veterans should be classified or for official use, as it has signs of military secrecy. This once again requires an awareness that the digitalization of veteran services and the use of AI in social work with veterans should be a controlled process, not a chaotic one, with a clear algorithm and control system.

In Ukraine, artificial intelligence, based on the Veterans Register and numerous databases with unified access to them, not only changes the course of the war, but also provides an opportunity to take care of war veterans on an innovative basis through the digitalization of veteran services and practical innovations (e.g. the introduction of a veteran's assistant, the existence of independent forensic diagnostic centres). For example, a significant contribution to improving social assistance to war veterans is the inclusion of a veteran's assistant in the national ecosystem of veteran policy in Ukraine, who is trained to accompany a Ukrainian defender after returning to civilian life. There is also a need for special training and professional development of social workers to provide services to war veterans. It is important to instil in these workers the ability to develop community-specific projects in the field of social work with veterans and provide their communicative support and implementation. Therefore, with the emergence of war veterans as a new socially vulnerable group, further professionalization of social work is required! And all these efforts must now be made with consideration for the peculiarities of the digital age and the possibilities, advantages, and disadvantages of using AI.

We propose to measure the innovative potential of AI in the social integration of war veterans and meeting their needs using a system of indicators: (1) the ability of the system of social work with war veterans to get rid of outdated approaches, forms, and types that cannot be digitized; (2) the ability to perceive new information and communication technologies of social work; (3) the ability to implement these technologies in the practice of service delivery; and (4) the ability to retain the constructive in social work due to the manageability of the process of AI application by professional staff of the relevant services (primarily state as well as municipal at the level of territorial communities where a particular veteran recipient of social services lives). These indicators can be used to measure the effectiveness of the use of artificial intelligence to solve certain problems in social work with war veterans.

The system of social integration of veterans of the Russian-Ukrainian war is being actively developed. Reintegration into civilian life is one of the stages of every soldier's return from war. Equally important is the social support of veterans (medical, educational services, employment, ensuring a decent life, well-being, public respect, and relevance). Such social support in Ukraine is provided by the state, civil society, veterans' self-organization, volunteers (Ukrainian and international), international organizations, governments, and local branches of a number of countries.

There are two promising areas for further research: (1) studying the self-organization of war veterans and (2) the topic of war veterans in public communications (in real and virtual communication space).

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

References

"Образ ветеранів в українському суспільстві": які результати загальнонаціонального опитування. (n.d.). Armyinform.com.ua. Retrieved March 20, 2024, from https://army-inform.com.ua/2023/09/21/obraz-veteraniv-v-ukrayinskomu-suspilstvi-yaki-rezultaty-zagalnonaczionalnogo-opytuvannya/

"Штучний інтелект" допоможе виявити тих, хто отримує соціальну підтримку незаконно. (2019, July 18). *ГЛАВКОМ*. https://glavcom.ua/news/shtuchniy-intelekt-dopomozhe-viyaviti-tih-hto-otrimuje-socialnu-pidtrimku-nezakonno-611117.html

«Це категорія потужних людей». Міністр у справах ветеранів розповіла про допомогу ветеранам і їхнім родинам. (n.d.). *Міністерство У Справах Ветеранів*. https://mva.gov.ua/ua/news/ce-kategoriya-potuzhnih-lyudej-ministr-u-spravah-veteraniv-rozpovila-pro-dopomogu-veteranam-i-yihnim-rodinam

18% українців вважають, що суспільство не поважає ветеранів. (2023, September 21). Www.ukrinform.ua. https://www.ukrinform.ua/rubric-society/3764270-18-ukrainciv-vvazaut-so-suspilstvo-ne-povazae-veteraniv.html

Barbrook-Johnson, P., & Penn, A. S. (2022). *Systems mapping: How to build and use causal models of systems* (p. 186). Springer Nature.

Barreteau, O., Antona, M., D'Aquino, P., Aubert, S., Boissau, S., Bousquet, F., et al. (2003). Our companion modelling approach. *Journal of Artificial Societies and Social Simulation, 6*(1).

Digital 2023: Ukraine. (n.d.). *DataReportal – Global digital insights*. https://datareportal.com/reports/digital-2023-ukraine

Medical care availability and level of satisfaction with services for ex-combatants. (n.d.). UNDP. Retrieved March 20, 2024, from https://www.undp.org/ukraine/publications/medical-care-availability-and-level-satisfaction-services-ex-combatants

State Enterprise DIIA. (n.d.). Se.diia.gov.ua. Retrieved March 20, 2024, from https://se.diia.gov.ua/en/

State in a Smartphone: More than 20 Million Ukrainians Use Diia. (n.d.). *Cabinet of Ministers of Ukraine* (March 20, 2024). https://www.kmu.gov.ua/en/news/derzhava-u-smartfoni-diieiu-korystuietsia-ponad-20-milioniv-ukraintsiv

Szczepanska, T., Antosz, P., Berndt, J. O., Borit, M., Chattoe-Brown, E., Mehryar, S., et al. (2022). GAM on! Six ways to explore social complexity by combining games and agent-based models. *International Journal of Social Research Methodology, 25*(4), 541–555. https://doi.org/10.1080/13645579.2022.2050119

Адаптація ветеранів війни до мирного життя: помічник ветерана. (2023, November 6). *Національний інститут стратегічних досліджень*. https://niss.gov.ua/news/komentari-ekspertiv/adaptatsiya-veteraniv-viyny-do-myrnoho-zhyttya-pomichnyk-veterana

Аналіз системи соціального захисту ветеранів та військовослужбовців. (2022). Retrieved March 20, 2024, from https://legal100.org.ua/wp-content/uploads/2022/08/2022-Bila-kniga.pdf

Винен штучний інтелект. Австралійські лікарі потрапили у гучний скандал. (2023, July 28). *ГЛАВКОМ*. https://glavcom.ua/techno/hitech/vinen-shtuchnij-intelekt-avstralijski-likari-potrapili-u-huchnij-skandal%2D%2D945341.html

Демченко, Артюх, & Булига. (2022). *Оцінка доступності та задоволеності медичною допомогою в Україні екскомбатантами/екскомбатантками АТО/ООС*. UNDP. https://www.undp.org/sites/g/files/zskgke326/files/2022-08/Звітзсоц.дослідження-ветерани 02.06_КТ_Чупа_ред_web.pdf

Друге анонімне онлайн-опитування серед ветеранів та діючих військовослужбовців «Портрет ветерана. Блок "Потреби ветеранів" 6-12 лютого 2023. (n.d.). Retrieved March 20, 2024, from https://veteranfund.com.ua/doc/6-12-02-23.pdf

З початком широкомасштабної війни з'явилися нові суб'єкти права на безоплатну вторинну правову допомогу. (2023, May 3). *Безоплатна правова допомога*. https://legalaid.gov.

ua/novyny/z-pochatkom-shyrokomasshtabnoyi-vijny-zyavylysya-novi-subyekty-prava-na-bezoplatnu-vtorynnu-pravovu-dopomogu/

Загальнонаціональне опитування #24 Образ ветеранів в українському суспільстві. (n.d.). Retrieved March 20, 2024, from https://ratinggroup.ua/files/ratinggroup/reg_files/rg_ua_veterans_1000_ua_092023.pdf

Калмикова, Н. (June 25, 2023). *Як держава може підтримати ветеранів зараз і після війни*. УКРАЇНСЬКА ПРАВДА. https://www.pravda.com.ua/columns/2023/05/31/700703/

Каразуб, І. (15.09.2023). Історії ветеранок. Що робить держава та суспільство для соціалізації бійчинь. https://suspilne.media/570311-istorii-veteranok-so-robit-derzava-ta-suspilstvo-dla-socializacii-bijcin/

Коли дають групу інвалідності довічно. (2023, July 18). *ТСН.ua*. https://tsn.ua/ukrayina/koli-dayut-grupu-invalidnosti-dovichno-2371087.html

Помічник ветерана: дорожня карта для громад, кандидатів та родин Захисників і Захисниць. (n.d.). Mva.gov.ua. Retrieved March 20, 2024, from https://mva.gov.ua/ua/pomichnik-veterana/opis-proyektu

Chapter 11
Assessment for AI in Social Services: Community Virtual Nursing Home in Shanghai, China

Wu Qi and Li Hui

Abstract This chapter assesses the impact of introducing an AI system on the efficiency of service distribution and the quality of life (QOL) of the elderly in a community virtual nursing home in Shanghai, China. Using data from the Xinhong Street Virtual Nursing Home in Minhang District, we employ a difference-in-differences approach to analyze the changes in operational efficiency and quality of life indicators before and after the AI system was implemented. The results show that the AI system significantly improved operational efficiency and the physical health of the elderly. However, there was no significant improvement in social acceptance. The study also found that the AI system shifted the distribution of care services from proactive to responsive, leading to a change in care patterns. Overall, the study highlights the benefits of AI systems in improving service efficiency and health outcomes in elderly care while also raising questions about equitable service distribution and the need for human touch in care services.

Introduction

According to data released by the National Health Commission, China's elderly care has formed a "9073" pattern, in which about 90% of the elderly choose domestic self-care, about 7% rely on public nursing homes, and 3% move into private nursing homes. This pattern suggests that only those who need the community nursing service the most (i.e., severely disabled elderly) may be able to receive admission into public nursing homes, while the rest of residents stay home and receive remote services. Therefore, the improvement of digital infrastructure and intelligent services for nursing homes and domestic care has become the focus of the development of the elderly care industry.

W. Qi (✉) · L. Hui
Shanghai Institute for Science of Science (SISS), Shanghai, PR China
e-mail: wuqi@siss.sh.cn

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*,
Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_11

For smart elderly care, China has gone through a bottom-to-top process in policy making and has launched a series of policies from the perspective of civil pension and the integration of AI and civil service. The regional policies were released earliest as a primitive phase for national plan. In December 2022, the Shanghai Municipal Civil Affairs Bureau issued the "Three-Year Action Program of Shanghai Municipality to Promote the Construction of Smart Nursing Homes (2023–2025)," which clarifies the overall framework of Shanghai's smart nursing home construction and the main construction content. In January 2023, the construction of smart nursing homes was included in the Shanghai Municipal Government 2023 Project, and there are currently 36 of them throughout Shanghai, which have basically completed the construction task. The national policy comes as the following step, facilitating experiences of early regional policies. In March 2023, the "Overall Layout Plan for the Construction of Digital China" was released, proposing the overall framework layout of "2522" to realize that our country's digital development level will be among the world's top in 2035. "2522" means that China plans to consolidate the "2 foundations" of the digital infrastructure and the data resource system, promote the deep integration of digital technology with the construction of the "5-in-1" economic, political, cultural, social, and ecological civilization, enhance the "2 capabilities" of the digital technology innovation system and the digital security barrier, and optimize the "2 environments" for digital development at home and abroad.

Though the national policy has been released and is now under implementation, China's community-based elderly care policy still faces a number of challenges. First, the supply of elderly care facilities is insufficient. Due to the rapid growth of the elderly population, the demand for senior care facilities far exceeds the supply. Especially in some economically underdeveloped areas, the construction of senior care facilities is insufficient. Secondly, the quality and professionalism of elderly services need to be improved. There are problems with the poor quality of services and the poor quality of caregivers in some elderly care facilities, and training and supervision need to be strengthened. In addition, the coverage of community-based elderly care services is not yet broad enough, and older persons in some remote and rural areas still face the problem of inadequate elderly care services.

As a result, virtual nursing homes are introduced to urban communities as a method to tackle the problem of elderly care resource shortage. The virtual nursing home in Shanghai's Xinhong Street is an innovative model of elderly care services designed to provide comprehensive care and support for the elderly living at home. This project achieves 24-h health monitoring and safety assurance for the elderly by installing smart devices and sensors in their homes, such as smartwatches and intelligent monitoring mattresses. These devices can monitor key health indicators such as heart rate, blood pressure, sleep status, and time spent out of bed and automatically alert in case of abnormalities.

The operational model of the virtual nursing home combines online real-time monitoring with offline personalized services. In the event of an emergency or service need at home, the system responds immediately and can quickly dispatch community managers or service personnel to provide assistance. Additionally, the

elderly or their family members can also check the health and alarm status of the elderly through WeChat and other means.

In this chapter, we investigate the adoption of AI systems in nursing homes for the elderly using health-related data of the seniors under care on a daily basis. Combining quantitative and qualitative research methods, we evaluated the effectiveness of the virtual nursing home from five aspects:

- The efficiency of distribution of AI services
- The impact of AI services on the physical health of the elderly
- The impact of AI services on the social acceptance of the elderly
- Changes in the content of staff services after the addition of AI
- The feelings of the elderly after the addition of AI

These data are for elderly residents in Xinhong community who have experienced AI-assisted healthcare services, as well as the performance data of the implemented AI system.

AI-powered Elderly Care

As a revolutionary and disruptive emerging technology, artificial intelligence and human economic and social services are constantly integrating, aiming at the optimization of digital social services (He, 2019; Wirtz & Müller, 2019; Chen, 2019; Han, 2019; Dong, 2019; Hu, 2018). The introduction of AI technologies to public services has also brought a series of challenges to social security that cannot be ignored (Hu, 2018; Song, 2018; Scherer, 2017).

Regarding the implementation of intelligent home care services in Chinese urban communities, in the study of Shanghai, Jing and others, based on the examination of the policy and practice of home care services in Shanghai, pointed out that there is a contradiction between the efficiency of service and the adaptive capacity of the system, as well as a contradiction between collaboration and self-governance in the service integration of home care services. In this regard, it is necessary to establish a cooperative service supply system based on a collaborative perspective to realize the innovation of public management of urban home care services (Jing & Chen, 2009). Zhu et al. showed that the highest degree of demand is for housekeeping, nursing care, senior dining table, emergency rescue, elevator installation, free medical checkups, etc.; among them, the list of services has local characteristics, such as "elevator for old houses" and "indoor ventilation and sunshine" (Zhu et al., 2013). Li Bin and other surveys show that Shanghai elderly people have a high degree of demand for medical, health, spiritual, meal assistance, and housekeeping services, and there is a strong demand for senior dining table, dispensing medicine, maintenance, and emergency call; at the same time, there is an inverse relation between the cost and demand for meal assistance services—i.e., the lower the cost is, the higher the demand is; and the degree of specialization of medical services is inversely correlated with the demand—i.e., the higher the specialization is, the lower the demand

is (Li et al., 2016). Peng suggests that the low acceptance of smart devices by the Chinese elderly will bring difficulties to smart aging (Peng et al., 2016). Luo et al. examined the typical cases of home care service informationization construction, and found that due to the failure of government–society–enterprise interaction, it is necessary to build a macro-institutional environment of benign interaction between multiple subjects and guide and encourage more resources to be pooled into home care service (Luo & Tong, 2022).

Case Introduction

China Case Study Research Focus

The research focus of this study is on the deployment and impact of AI systems within the context of social welfare and public services distribution in China, a field that is currently in its nascent stages and predominantly focused on public safety initiatives. Despite the limited availability of cases for research, the present investigation has selected a nursing home in the Shanghai community that recently implemented an AI management system as the subject of inquiry. The selection of this case is predicated on its alignment with the research framework of the AI FORA project, as well as the availability to data and the possibility of conducting interviews. Furthermore, the chosen application scenario is deeply rooted in Chinese cultural norms, particularly the enduring values of filial piety and respect for the elderly.

Virtual Nursing Home in Minhang District

By the end of 2023, Minhang District had a household population of 1.26 million, of which 428,300 were aged over 60, accounting for 34.01% of the household population. Xinhong Street Comprehensive Service Center for the Elderly is one of the 90 comprehensive service centers for the elderly in Minhang District, with services covering about 11,000 household residents over the age of 60 in the street community. Since 2019, Minhang District has been piloting the "Virtual Nursing Home" service in Xinhong Street. At the "2022 Global Smart City Cooperation and Development Conference," the results of the "2022 City Digital Transformation Excellent Case Selection" were announced, and the Xinhong Street "Home Virtual Nursing Home" service platform project was selected as one of the 52 excellent cases nationwide.

Home Care Needs of Disabled Elderly Face Service Gap

According to the National Working Committee on Aging, there will be more than 42 million people over the age of 60 suffering from disability and dementia (hereinafter referred to as "disabled") in China in 2020, accounting for about 16.6% of the

total population of elderly people over the age of 60. This means that for every six elderly people over the age of 60, there will be about one who cannot take care of themselves and needs long-term care services. The virtual nursing home in Xinhong Street, Shanghai, is mainly for "disabled elderly"; i.e., it includes "community service centers for the elderly + home care" for two parts of the elderly people, totaling more than 300 people. According to the survey data, by the end of 2023, this virtual nursing home had provided services to a total of 719 individuals, including 154 elderly residents in the institution and 565 receiving home-based care. Based on the proportion of disabled elderly in the XinHong Street population, it can be roughly estimated that out of the total approximately 1800 disabled elderly in the area, the 719 individuals currently served by the virtual nursing home represent about 39.3% of the disabled elderly population. Additionally, nearly 60% of the disabled elderly are in need of seeking other forms of services.

Smart Services in Virtual Nursing Homes

Through smart mattresses, smart bracelets, and other devices, the sleep quality and health status of elderly people can be monitored in real time, and appropriate health advice and services can be provided to the elderly based on data about their living habits such as sleeping duration and exercise frequency. Elderly people can also call the service center to express their needs. The Elderly Service Center is mainly staffed by government employees and outsourced personnel. Volunteers supplement the daily work as "community housekeepers."

The AI in virtual nursing home utilizes intelligent agents and smart management platforms, where data and computers drive the management process. This involves using intelligent interactive devices, IoT sensory devices, and other perception terminals, combined with technologies such as facial recognition for data collection and the generation of management decisions. For example, face recognition is used to prevent seniors with Alzheimer's disease from leaving the management campus, and data on the health status of seniors at home is simultaneously collected by interactive devices (smart bracelets) and IOT sensing devices (smart mattresses), and trends in the seniors' health are analyzed through the accumulation of data, and so on.

The Virtual Nursing Home provides two smart devices; one is a smart bracelet that can monitor heart rate, blood pressure, and other indicators; the other is a smart mattress that can monitor the elderly's heart rate, breathing, time out of bed, and so on. They can also choose the "6 Choose 2" home service (choose 2 items from haircut, pedicure, massage, housekeeping, fruit delivery, and medication dispensing), and the monthly community volunteers' community housekeepers can also come to the home to provide nursing care. If an elderly person has an emergency at home, he or she can connect to the "24-h service" through the smart wristband or smart mattress, and the back-end service will respond immediately.

Shanghai Virtual Nursing Home Admission Process

The targets of the virtual nursing home are selected by the sub-district office. The welfare information of the elderly was collected via national census to firstly choose those who meet the criteria, and then the community committee will offer door-to-door reminder to the eligible families and see if they need the nursing home service.

- Age requirement: Over 60 years old, to ensure that the elderly can receive better care and attention in the nursing home.
- Financial condition: Difficult financial conditions, for example: receive basic living allowances, or with family income less than 20% of the local minimum living standard. This ensures that elderly people with financial difficulties can receive social assistance and be admitted to nursing homes.
- Living conditions: Priority is given to elderly people living alone, who often lack the companionship and care of a family and need the care and attention of a community nursing home. At the same time, priority will also be given to families where multiple members suffer severe disabilities, as well as elderly veterans who receive preferential benefits.

Methods

Quantitative Research Design Overview

The research design for this case study is quantitative in nature, employing a pre-post experimental design with a control group to evaluate the impact of AI technology implementation on the service efficiency of a nursing home's AI system. The study spans a 22-month period, divided into two distinct phases: the pre-implementation phase (first 10 months) and the post-implementation phase (last 12 months).

During the pre-implementation phase, data on the nursing home's service efficiency, as well as the health index and social acceptance index of the elderly residents, were collected. This phase serves as the baseline for measuring the effects of the AI technology implementation. In the post-implementation phase, the same data were collected to assess any changes that occurred after the introduction of AI technology.

To analyze the data, a difference-in-differences (DID) approach is utilized. The DID method compares the changes in the outcome variable (service efficiency index) over time (pre- and post-implementation) between the treatment and control groups. This method helps to control for any time-invariant factors that could affect service efficiency, providing a more accurate estimate of the causal effect of AI technology implementation.

In addition to the primary outcome measure (service efficiency index), the health index and social acceptance index are included as control variables in the analysis.

These variables are considered to account for any potential confounding factors that could influence the dependent variable.

The quantitative research design overview outlines the study's framework, including the research questions, the data collection process, the sample selection, and the statistical methods employed. This design allows for a rigorous evaluation of the impact of AI technology on the service efficiency of the nursing home's nursing home, providing valuable insights into the potential benefits and challenges of integrating AI technology in healthcare settings.

Applicability of the DID Model in This Case Study

The difference-in-differences (DID) method is particularly well-suited for this case study examining the quality of care improvements in a nursing home over a 22-month period for several reasons. Firstly, the DID approach allows us to mitigate the effects of confounding factors by comparing the changes in outcomes over time between a treatment group (the nursing home in question) and a control group (a similar nursing home not experiencing the same interventions or quality improvement measures). This is crucial in isolating the causal impact of the specific interventions or changes implemented in the treatment nursing home.

Secondly, the DID method is effective in situations where random assignment is not feasible, as is often the case in real-world settings such as healthcare facilities. By relying on a natural experiment framework, where the treatment and control groups are determined by the presence or absence of the treatment, we can still obtain estimates of the treatment effect that are less biased than those from simple before-and-after comparisons.

Lastly, the monthly data collected over the 22-month period provides a rich source of information that allows for the estimation of both the immediate and longer-term effects of the quality improvement measures. The DID model will help us to analyze whether any observed changes in the quality of care are indeed attributable to the interventions by comparing the changes, thus providing a more robust evaluation of the nursing home's quality improvement efforts.

Data Sources

The data used in this study come from the Virtual Nursing Home of Xinhong Street in Minhang District, Shanghai, which provides data on residents' life information, and from the daily records of the artificial intelligence system of Hengyan & Lebang Elder Care, a partner company in the community nursing program. The company is responsible for the virtual nursing home project in Xinhong Street, which includes smart devices, the construction of the smart system, and the provision of nurses. The data content is monthly data, extracted from a database of health-related data of all

community residents who agreed to receive virtual nursing home services, starting the data collection 10 months before they were enrolled in the program and continuing for 10 months after enrolment (22 observations in all). In terms of the QOL-related data, we collect the health and social acceptance indicators from the nursing home in the same time span.

Main Variables

In the context of this case study, the implementation of AI technology in the nursing home's nursing home serves as the primary explanatory variable. This variable is dichotomous, with a value of 0 indicating the absence of AI technology during the first 10 months of the study period, and a value of 1 indicating the presence of AI technology during the subsequent 12 months. The introduction of AI technology is expected to have a significant impact on the efficiency of the nursing home, which is the dependent variable in this study. The nursing home's service efficiency index is a composite measure that captures the effectiveness and productivity of the nursing home operations.

In addition to the service efficiency, two other variables are considered as explained variables in the analysis. The first explained variable is the health index of the elderly residents, which is a standardized measure that reflects the overall health and well-being of the residents in the nursing home. This variable is included in the analysis to account for the potential influence of residents' health conditions on the nursing home's service efficiency. The second explained variable is the social acceptance index, which gauges the level of acceptance and satisfaction among the residents and their families regarding the services provided by the nursing home. This variable is incorporated into the analysis to discover the potential impact of residents' and families' perceptions and satisfaction on the nursing home's service efficiency. The detailed variable settings are shown in Table 11.1.

e Service distribution efficiency, the efficiency of the service distribution of the elderly institutions, first of all, according to the number of calls received by the elderly service center and the number of problems handled each month, using entropy value method to measure the efficiency of the elderly institutions.

ph Physical health, physical health index of the elderly in the community. This health index is provided by the elderly center for each elderly person through measurement. The questionnaires are averaged out for all the elderly.

sa Social acceptance, social acceptance index of the elderly in the community. This social acceptance index is provided by the elderly center for each elderly person through the questionnaire. All the questionnaires are averaged here.

Table 11.2 is the monthly value of each variable. It is calculated based on the number of service orders received by the elderly service center and the number of problems handled each month, using the entropy value method. The data covers a

Table 11.1 Main variables and statistical method

Variable name	Variable meaning	Outcome method
e	Service distribution efficiency of the nursing home	Entropy value method
ph	Physical health indicator of elderly people	Mean value from questionnaire
sa	Social acceptance indicator of elderly people	Mean value from questionnaire

Table 11.2 Service distribution efficiency and QOL of elderly people

		Service distribution efficiency		Elderly people quality of life		
Year	Month	Service order volume	Service order finished	Health index	Social acceptance index	Treat (AI)
2021	3	31		4.50	3.76	0
	4	37	36	4.12	4.01	0
	5	55	42	4.30	3.52	0
	6	79	65	4.65	4.33	0
	7	1	60	4.4	4.54	0
	8	128	128	4	4.27	0
	9	116	116	3.8	4.16	0
	10	352	178	3.76	3.90	0
	11	97	177	3.78	3.64	0
	12	173	179	3.65	3.52	0
2022	1	241	236	4.01	3.43	1
	2	157	250	4.25	3.65	1
	3	413	97	4.45	3.72	1
	4	256	365	4.57	4.21	1
	5	346	329	4.60	4.20	1
	6	135	243	4.65	4.13	1
	7	254	211	4.63	4.65	1
	8	476	348	4.70	4.42	1
	9	378	401	4.76	4.23	1
	10	395	433	4.72	4.16	1
	11	421	329	4.78	4.03	1
	12	438	445	4.77	4.10	1

22-month period from January 2021 to December 2022, with a total of 22 observations. The health index and social acceptance index are offered by the nursing home.

To better understand the characteristics of the research sample, we conducted descriptive statistical analyses on the collected data. Table 11.3 presents the basic statistics for each variable, including the mean, standard deviation, minimum, maximum, and sample size.

Table 11.3 Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
did	22	0.545	0.51	0	1
e	22	0.457	0.306	0.029	0.964
ph	22	4.357	0.376	03.65	4.78
sa	22	4.026	0.342	3.43	4.65

The DID Model

The double difference model (DID) is widely used and can well circumvent the endogeneity problems. This chapter constructs PSM-DID model for empirical analysis. The time span of the study is 10 months before and 10 months after the commissioning of the virtual nursing home services. The elderly who receive virtual nursing home services in the community are set as the treatment group, and the elderly who do not receive services in the same community are set as the control group. In this chapter, the initial samples are first processed by propensity score matching (PSM), whereby the experimental group in each month is paired with the matched control group through year-by-year matching to ensure that the long-term trend of the elderly in the experimental group and the control group is the same prior to obtaining virtual nursing home care, so as to exclude the influence of other factors. The multi-period DID model is constructed as follows:

$$Y_{it} = \propto_0 + \alpha_1 did_{it} + \mathrm{X}_{it} + \theta_t + \lambda_i + \varepsilon_{it}$$

The dependent variable Y_{it} reflects the senior residents' QOL, and the subscripts i and t indicate the AI system and the time, respectively.

did_{it} is treat*after, which refers to the introduction of the AI system in daily operation of the nursing home, where treated is an individual dummy variable with a value of 1 for the experimental group and 0 for the control group; after is a dummy variable for the time of policy implementation, with the year of policy implementation and the year after taking the value of 1, and the rest taking the value of 0. X is each control variable. θt is a time fixed effect. λi is an individual fixed effect, and εit is the random error term.

Empirical Analysis

Robustness Check

In this section, we conduct a multiple covariance test on the explanatory variables within the model so as to check the relations between variables. The test revealed that there were significant differences among the variables e, ph, and sa, as indicated by their variance inflation factor (VIF) values of 1. The VIF is a measure that quantifies how much the variance of an estimated regression coefficient increases when

the predictors are correlated. A VIF value of 1 indicates that there is no correlation between the explanatory variable and the other variables in the model. Given that there is no highly correlated relationship observed among the variables, it is appropriate to proceed with the double difference model testing. The absence of multicollinearity ensures that the model will not be affected by issues related to unstable coefficients or difficulties in interpretation due to overlapping information provided by correlated variables. Thus, the results of the double difference model testing can be considered reliable and robust, providing valuable insights into the relationships and effects of the variables under study.

Statistical Analysis Results

Correlation Analysis

The introduction of an artificial intelligence (AI) system has been found to significantly influence the governance efficiency and objectives within virtual nursing homes. Prior to conducting the difference-in-differences (DID) test, a correlation analysis was performed on four key variables, revealing a substantial impact of the DID value on both the efficiency of service distribution (E) and the physical health (PH) of the elderly. According to Table 11.4, results indicate that the integration of AI systems has positively affected these outcomes. However, the analysis did not detect any significant influence on the social acceptance status of the elderly, suggesting that while AI systems can enhance operational and health-related aspects, their impact on social dynamics within these communities may be limited.

The Impact of the AI System on Virtual Nursing Home: Basic Results

As can be seen in Table 11.5, the DID values in the e- and ph-terms show good significance over the 22-month period of observation. That is, after the introduction of the AI system, there is a significant increase in both the efficiency of management and the physical health of the elderly. The effect on the sa term is not significant, and the improvement in the social acceptance of the elderly is not strong. Meanwhile, by

Table 11.4 Correlation analysis result

	Did	e	ph	sa
did	1			
e	**0.801**	1		
	0			
ph	**0.649**	0.490	1	
	0.001	0.021		
mh	0.168	0.200	0.519	1
	0.456	0.373	0.013	

Table 11.5 Time-varying DID* result

	(1)	(2)	(3)
Variables	E	ph	sa
Did	0.480***	0.478	0.113
	(5.978)	(3.815)	(0.760)
Constant	0.195	4.096	3.965
	(3.283)	(44.246)	(36.287)
Observations	22	22	22
R-squared	0.641	0.421	0.028

Note: t-statistics in parentheses, ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

observing the original data, it is concluded that the high value of social acceptance of the elderly in 22 months is distributed in June–August every year, perhaps related to the fact that children and grandchildren have more leisure time for visiting in summer; these arguments need to be further verified.

Advantages and Disadvantages of AI-Powered Elderly Care Services

Interviews with Elderly Care Center Staff

After collecting the data, we conducted interviews with the staff of the elderly care centers based on the first question of the AI FORA research framework, which is the specific operation in the implementation of social welfare allocation, to explore how they use intelligent devices to provide smart elderly care services for the disabled elderly, as well as the challenges encountered and strategies adopted in the implementation of this service.

First, the interviews focused on the intelligent management of the disabled elderly. Officials introduced the smart management system developed since 2013, which monitors the elderly's movements and provides in-home services through electronic fencing technology. Initially, the system generated about 300 alarms per month, mainly because the elderly triggered the electronic fence during activities such as getting out of bed at night. However, through continuous optimization of the system and education on geriatric science, the number of alarms has been reduced to 10 per month. This improvement was achieved by reducing the false alarm rate and increasing users' acceptance of the device. In addition, the product was improved to address issues such as inconvenient charging, product quality, and WI-FI connectivity to increase the convenience of the elderly in using the device.

Second, the interviews focused on the impact of intelligent management on staff when providing services. Prior to the introduction of the AI system, the nursing home staff adopted a proactive management approach, i.e., face-to-face

communication with the elderly in the form of home visits to gather information about their needs. This approach reflects the staff's attention to the individual needs of the elderly and their high level of responsibility for service quality. Through home visits, staff are able to directly observe the elderly's living environment and understand their habits and preferences to better understand their real needs. After completing the community interviews, the staff will organize and analyze the information collected and use this demand information as an important basis for formulating the next stage of work strategies. This demand-oriented work strategy helps ensure that nursing home services are closer to the actual needs of the elderly, thus improving service quality.

However, with the introduction of AI systems, operation routine of the nursing home has changed significantly. There has been a shift from proactive to reactive management, with a reduction in the number of visits made by staff. Instead, they maintain communication with the elderly mainly through phone, Internet, or face-to-face conversations. This change has resulted in raised monthly service efficiency implied by elevated number of and also enabled staff to make better use of modern communication technology to maintain close contact with the elderly. However, the number of face-to-face meetings between staff and the elderly during home visits is still relatively limited. This may be due to the fact that venues such as activity centers for the elderly provide more opportunities for communication, making it easier for staff to communicate with the elderly face-to-face. In addition, the introduction of the AI system has made it possible for staff to organize their working hours more flexibly, allowing them to better balance the need for visits and communication.

According to the results of the data analysis, the efficiency of the nursing home's services has increased and the health of the elderly has improved, which are supposed to be the benefits of the AI management system. However, the statistics also show that the social acceptance of the elderly has not improved, indicating that the introduction of the AI management system has not positively affected the mood of the cared-for group. Whether this is due to the way the management is set up or whether it reflects the boundaries of the AI's own capabilities is an issue that needs to be further explored and discussed.

Group Interviews with Three Seniors in a Virtual Nursing Home

In addition, we learned about the content of seniors' needs over the phone to study how the AI system recognizes and responds to seniors' needs. With the consent of the nursing home staff, we reviewed the nursing home's call logs and found that the main needs were health and housekeeping, but about 10% of the calls were made by seniors who wanted to test the usefulness of the AI system or wanted to talk to customer service agents because they were lonely. After speaking with staff, we conducted group interviews with three seniors in the virtual nursing home.

Interviewee inclusion criteria:

- Aged 60 years and above
- Have lived in the community for ≥1 year
- Be conscious and have no obstacle to communicate with the investigator
- Voluntary participation in the study

The community nursing home interviews provided insight into the challenges of implementing a virtual nursing home in the community, particularly the elders' perceptions of the new model of care. In the interviews, the elders raised the following four main concerns:

First, elders were generally concerned about the security of their personal information. They were worried that their personal information might be misused or leaked in the process of adopting virtual nursing home services, which increased their sense of insecurity about virtual nursing home services. They want to protect their privacy rights and interests and avoid improper use of their personal information.

Second, many elderly people have a negative attitude towards intelligent devices. Some of the elderly do not have smartphones or are not very good at using smart devices, and they are unfamiliar with and confused by these new technologies. They are worried about the problems and inconveniences that may arise during the operation process, so they have a conservative attitude towards the services of virtual nursing homes.

In addition, the elderly expressed concerns about the daily maintenance of smart devices. They mentioned that the difficulty of charging smart devices, the mismatch between product functions and actual use scenarios, and the lack of WI-FI at home caused them problems. They hoped that the AI service could be more convenient and adaptable to meet their actual needs.

On the point of communicating and chatting with the customer service staff, the elderly also confessed that because their children are not around, they sometimes like to "talk more" with the customer service staff in the nursing home. An old man said, especially in the New Year holidays, the first day of the year when the body is not feeling well you cannot help but talk about the heart; she said the old man almost wanted to go to the supermarket to buy his own medicine; she said that the couple also quite miss their son studying abroad.

Finally, the old people expressed doubts about the radiation that may be produced by smart devices. They are worried that wearing smart devices for a long time may have a negative impact on their health, and in particular, they are concerned about the potential harmful effects of radiation on their bodies.

In summary, virtual nursing homes face the challenge of the difficulty of changing the attitudes of the elderly in the process of implementation in the community. Worries about the security of personal information, rejection of intelligent devices, daily maintenance problems, and doubts about radiation are common concerns among the elderly. Therefore, those elderly people who do not reject new things and are able to adapt to new technologies are more likely to accept the services of virtual nursing homes. Judging from the current implementation, families above the middle

class have a higher degree of acceptance of virtual nursing homes, which may be related to the notion that they are more willing to try and accept new technologies.

A Comparison of Face-to-Face and AI Management Method in Nursing Home

Based on the content of the two interviews, it can be concluded that there are advantages and disadvantages to both home-visit management style in nursing homes and the management style conducted by AI, as follows:

The advantages of home-visit method include the following:

- The home-visit method allows for a profound emotional connection between the elderly care center staff and elderly people. Through face-to-face interactions, staff can engage in a holistic understanding of their clients' experiences and emotions. The nuances of non-verbal communication, including body language and facial expressions, provide insights that may be overlooked in remote or digital interactions, enhancing the elderly care center staff's capacity for empathy and allowing for a more nuanced and responsive approach to client needs.
- The home-visit method is instrumental in building and maintaining trust between the elderly care center staff and the client. The consistent, personal contact that characterizes home visits fosters an environment of trust and confidentiality, which is essential for clients to feel secure in sharing sensitive information and for the subsequent development of an effective working relationship.

The disadvantage of the home-visit management approach in nursing homes primarily lies in its low efficiency under the circumstances of insufficient staff and a high number of elderly residents in Xinhong community. As a result, the elderly have to wait for an extended period before they can receive a home visit, which can lead to feelings of neglect and frustration among the residents. This prolonged waiting time can also result in delayed identification and addressing of the elderly's needs, potentially compromising their well-being. Moreover, the distribution of services among the elderly is unequal. Those who have more connections in the community, are more integrated into the community, or are more extroverted and willing to communicate tend to receive more attention and care from the staff. This preferential treatment based on personal characteristics or social relationships can lead to feelings of injustice and exclusion among those who do not fall into these categories. It can also create a divide within the nursing home community, as residents may perceive favoritism and bias in the staff's actions.

The deployment of AI in elderly care has resulted in a substantial improvement in the ability to monitor the safety and health of the elderly. Advanced sensors and monitoring technologies facilitate real-time tracking of vital signs and environmental conditions, ensuring that any deviations from the norm are promptly detected and addressed. This proactive approach to health and safety monitoring has led to a reduction in the incidence of accidents and health complications, thereby

contributing to a safer living environment for the elderly. In addition, the AI system's capabilities in data analysis and pattern recognition have enabled healthcare providers and caregivers to gain deeper insights into the health trends and behaviors of the elderly. This data-driven approach to elderly care allows for more informed decision-making, personalized care plans, and timely interventions, thereby optimizing the allocation of resources and improving the overall efficiency of care delivery. Last but not least, the advent of AI systems has introduced a new dynamic to the provision of elderly care services. In terms of service distribution, with the implementation of responsive service, the distribution of care is increasingly contingent upon the direct expression of needs to the care center. This means that elderly people who actively voice their requirements to the nursing center are more likely to receive additional care and services relative to their peers who do not articulate their needs as assertively.

This transition to a responsive management system has the potential to democratize access to care, as it relies less on subjective judgments and social dynamics and more on the objective articulation of care needs by the elderly themselves. However, it also raises concerns about the digital divide, as older adults with lower levels of technological literacy may be disadvantaged in navigating and utilizing the AI systems to express their needs.

Conclusion

The results of the empirical study and the semi-structured interviews show the current situation of adding AI systems to the community care of the elderly in Shanghai, China, The previous home-visit method in elderly care offers several advantages, including the development of a strong emotional connection between staff and elderly individuals, the building of trust, and the provision of personalized care. These benefits arise from face-to-face interactions that allow for a deeper understanding of the elderly's experiences and emotions, as well as the fostering of a trusting relationship through consistent personal contact.

However, the home-visit approach also has its drawbacks, particularly in terms of efficiency when there is a shortage of staff and a high number of elderly residents. This can lead to extended waiting times for home visits, neglect, and frustration among the elderly, as well as delayed addressing of their needs. Additionally, the distribution of services can be unequal, with those who are more socially connected or extroverted receiving more attention, potentially causing feelings of injustice and exclusion among others.

The introduction of AI in elderly care has significantly improved safety and health monitoring, with real-time tracking of vital signs and environmental conditions. AI systems also enhance decision-making through data analysis and pattern recognition, leading to more personalized care plans and timely interventions. This proactive approach has resulted in a safer living environment for the elderly and optimized resource allocation.

The shift to responsive governance in elderly care, facilitated by AI, has the potential to democratize access to care by prioritizing the direct expression of needs. However, it also raises concerns about the digital divide, as older adults with lower technological literacy may face difficulties in using AI systems to articulate their needs.

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

References

Chen, P. (2019). Government governance in the age of artificial intelligence: Adaptation and transformation. *E-Government, 3*, 27–34.

Dong, L. R. (2019). Research on artificial intelligence development and government governance innovation. *Journal of Tianjin Administrative College, 3*, 3–10.

Han, X. (2019). Let information flow: Artificial intelligence and government governance change. *The Social Studies, 4*, 79–86.

He, J. H. (2019). The technical logic and value reshaping of AI-empowered government governance. *Journal of Shanghai Normal University (Philosophy and Social Science Edition), 5*, 112–121.

Hu, H. B. (2018). Change and innovation of government governance mode in the era of artificial intelligence. *Academia, 4*, 75–87.

Jing, Y. J., & Chen, R. J. (2009). The development and management innovation of China's home care service system from the perspective of collaboration. *Fudan Journal (Social Science Edition), 5*, 133–140.

Li, B., Wang, Y., Li, X., et al. (2016). Demand for elderly care services in urban communities and its influencing factors. *Journal of Architecture, S1*, 90–94.

Luo, Y., & Tong, Y. L. (2022). Service levitation: The practical dilemma of informatization of home-based elderly services under the logic of subject action. *Journal of Zhengzhou University (Philosophy and Social Science Edition), 3*, 25–30.

Peng, G., Garcia, L. M. S., Nunes, M., & Zhang, N. (2016). Identifying user requirements of wearable healthcare technologies for Chinese ageing population. In *Proceedings of the 2016 IEEE international smart cities conference (ISC2), Trento, Italy. 12–15 September 2016* (pp. 1–6).

Scherer, M. (2017). Regulating artificial intelligence systems: Risks, challenges, competencies and strategies. *Harvard Journal of Law and Technology, 2*, 353–400.

Song, J. L. (2018). AI and governance reform. *Academic Exploration, 12*, 55–61.

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review, 21*, 1–25.

Zhu, R., Hao, Y., & Wang, F. F. (2013). Survey on the demand for elderly services in Shanghai. *Social Security Research, 6*, 21–26.

Chapter 12
AI Integration in Mental Health Services: Examining Trends in the USA and Peoria, Illinois

Margaret Hinrichs, Jieshu Wang, Caity Roe, and Erik W. Johnston

Abstract In the USA and globally, public provisioning systems are evolving in two fundamental ways. The first is to reorganize from decentralized services to coordination around systems of care. The second is the widespread integration of AI into multiple social service areas including mental health diagnosis, needs assessment, and service delivery. While AI has displayed tremendous potential across various dimensions of mental health, including prediction, monitoring, diagnosis, treatment, and assessment, the use of AI also introduces new challenges to performance and accountabilities. This chapter explores the use of systems of care in Peoria, Illinois, for coordinating public service provisioning across multiple organizations serving vulnerable populations. Practitioners identified barriers for the public including logistical, social, cultural, and internal organizational challenges. Lessons from the case motivate a broader exploration of the use of AI in public service provisioning in the USA with a deeper dive into the use of AI in the mental health social service area. Concerns and challenges are included to promote a balanced conversation on the opportunities and accountabilities for using AI in public service provisioning. As the use of AI becomes more widespread, continuous interrogation and reflection are necessary to realize the potential of AI consistent with the values of the public service organizations, to be in service of the publics that benefit from these programs, and to minimize unintended consequences.

Case Study Introduction and Focus

There is a broad trend in governance and public administration to reorganize social service provisioning systems from being coordinated around specific provisions (education, safety, health) to being organized around the individual (NTAEC, 2009).

M. Hinrichs (✉) · J. Wang · C. Roe · E. W. Johnston
Global Futures Laboratory, School of Complex Adaptive Systems, College of Global Futures, Arizona State University, Tempe, AZ, USA
e-mail: mhinrich@asu.edu; jwang490@asu.edu; cgroe@asu.edu; erik.johnston@asu.edu

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_12

These new arrangements are commonly referred to as "systems of care." The distributed and integrated nature of these systems results in novel threats and opportunities to attend to the accountabilities of individual parts of the system and the system in its entirety. Additionally, public service provisioning that is responsible for essential public services, such as public health, public safety, and mental health services, is facing increasing internal and external pressure to embrace data-driven tools to augment their decision-making processes. This push is driven by limited resources and the availability of vast amounts of digital health information, with the aim of improving the quality and efficiency of their operations. However, many of these public service organizations lack the necessary human and technical infrastructures to effectively understand, manage, analyze, and extract actionable insights from these datasets.

In response to the gap between expectations and capacity, private companies and universities are developing computational systems capable of integrating, analyzing, and visualizing data obtained from the public sector. Despite being externally owned and sometimes externally operated, these technological systems are becoming deeply integrated into public services, forming a fundamental part of the public infrastructure for managing and delivering these critical services.

While these data tools address the immediate need for data management, automated decision-making, and insight generation, they also raise significant questions regarding accountability. Other sectors operate under different accountability frameworks and oversight mechanisms compared to the public sector, primarily due to their market-driven approach to designing and developing software solutions or their academic approach that incorporates research and experimentation sensibilities. This discrepancy highlights the importance of addressing conflicting or even absent accountability issues in the utilization of private data systems within public service delivery to ensure equitable provision of public services. In particular we explore the gap between who designs a system and who uses the system.

For the rest of this chapter, we follow the AI Fora framework. First, we provide context on the need for public health provisioning in mental health, the social service area that is the focus of our study, in the state of Illinois. We then present our specific case which explores system of care in Peoria, Illinois. Although our case was halted because of significant turnover in key personnel, we present early findings from a series of meetings and conversations with key stakeholders. These insights motivated two broader conversations on the use of AI in public service provisioning in the USA and then focused more on the use of AI specifically in the provisioning of mental health services in the USA. For both, we also discuss concerns and challenges moving forward.

Social Service Area: Mental Health

The USA grapples with a significant burden of mental health issues, with more than 20% of adults (over 50 million Americans) experiencing at least one mental illness, and nearly 94% of adults with substance use disorders failing to receive any treatment. Disturbingly, almost 5% of US adults (over 12 million) report serious thoughts of suicide, with youth facing a similarly distressing scenario. Over 10% of US youth contend with severe major depression, which detrimentally affects their lives and studies, and an alarming 60% of them do not receive the necessary treatment.[1] The situation has exacerbated, notably due to a sharp increase in suicide rates among children and youth (Curtin, 2020), making suicide the second leading cause of death among youth aged 10–14 in the USA.[2]

Illinois, ranked 28th among states in the prevalence of mental illness (Masterson, 2023), grapples with its share of mental health challenges. Mental health issues among adolescents have long been identified as a top health concern in the state (Illinois State Board of Education, 1991). Before the onset of the COVID-19 pandemic, a SAMHSA report revealed that, on average, 15% of youth experienced major depressive episodes in a year, with only 43% of them receiving depression care. At the same time, 11% of young adults grappled with serious thoughts of suicide, and 8.7% faced serious mental illness (SAMHSA, 2020). Studies have also illuminated significant racial disparities among Illinois youth, with African American youth demonstrating the highest need for mental health services (Rawal et al., 2004). Hispanic youth and those on Medicaid were also less likely to access hospital care for mental health issues (Brewer et al., 2020).

The onset of the COVID-19 pandemic exacerbated these issues further. Mental health-related emergency department (ED) visits among youth aged 5 to 19 years in Illinois spiked in 2019, and subsequent hospitalizations surged during the pandemic (Brewer et al., 2022). Addressing these challenges is imperative, particularly considering the disparities in access and care.

Efforts to address the mental health challenges among youth and adolescents in Illinois have been initiated. For instance, the Children's Mental Health Act, passed in 2003, established the Illinois Children's Mental Health Partnership (ICMHP) to formulate a comprehensive and coordinated mental health system for Illinois children, adolescents, and their families.[3] ICMHP has devised two statewide Illinois Children's Mental Health Plans, with the most recent spanning 2022–2027 (ICMHP, 2020). Additionally, the Screening, Assessment, and Support Services (SASS) program, an offshoot of the Children's Mental Health Act, intervenes and aids children and adolescents experiencing mental health crises.[4]

[1] https://mhanational.org/issues/state-mental-health-america

[2] https://wisqars.cdc.gov/data/lcd/home

[3] https://www.icmhp.org/about-us/our-mission/

[4] https://hfs.illinois.gov/medicalproviders/behavioral/sass/sasshome.html

To address the mental health needs of youths arrested and detained in the juvenile justice system, the Mental Health Juvenile Justice (MHJJ) initiative was introduced in Illinois (Burnett-Ziegler et al., 2009; Lyons et al., 2003). This initiative identifies youths with psychotic disorders within the juvenile justice system, connects them with community services, and tracks their progress. Follow-up studies have underscored the initiative's significant role in reducing emotional problems among youths and enhancing various dimensions of their lives (Lyons et al., 2003). Another noteworthy initiative is the Evidence-Based Training Initiative in Illinois (EBTI), launched in 2005 by the Illinois State Mental Health Authority (Starin et al., 2014), with the aim of implementing evidence-informed practices in the children's mental health system.

In addition to these statewide initiatives, Illinois offers school-based mental health resources. School-wide systems in Illinois have been instrumental in fostering positive behaviors and improving the mental health functioning of students (Atkins, 2002). A hotline is accessible to children, parents, and educators in Illinois, providing information and guidance on cyberbullying and Internet safety.[5] The Crisis and Referral Entry Services (CARES Line), part of the SASS program, operates round the clock, offering assistance during mental health crises. Furthermore, Safe2Help Illinois,[6] a free multi-platform application, is available to students, enabling them to confidentially report school safety concerns in the absence of a trusted adult.

The Illinois government has also allocated substantial efforts and resources to enhance mental health services. In 2022, Illinois launched a federally funded youth mental health program with $2.5 million to ensure that children's mental health needs are adequately met by healthcare providers.[7] Additionally, in August 2023, Illinois announced an additional $10 million federal grant to strengthen mental health services in schools across the state. This grant will support initiatives such as trauma assistance, training for medical and school staff, and workforce enhancement.[8] Furthermore, Illinois youth under the age of 21 are eligible to enroll in YouthCare Illinois,[9] a specialized healthcare program selected by the Illinois Department of Healthcare and Family Services (HFS). This program extends coverage to include behavioral and mental health services, in addition to standard medical, dental, vision, and pharmacy services.

[5] https://ag.state.il.us/cyberbullying/index.html

[6] https://www.safe2helpil.com/make-a-difference/

[7] https://www.wgem.com/2022/11/01/illinois-launching-new-youth-mental-health-program/

[8] https://www.wjol.com/illinois-announced-10m-grant-to-strengthen-mental-health-services-in-schools-2/

[9] https://www.ilyouthcare.com/

US CASE: Systems of Care in Peoria, Illinois

Peoria, Illinois, is home to zip code 61605, one of the poorest zip codes in the country and the most distressed in the state of Illinois according to the Distressed Communities Index (Economic Innovation Group, 2020). 61605 suffers from decades of disinvestment, concentrated poverty, and racial segregation resulting in Peoria being named the worst city for African Americans to live in 2016 (Comen, 2019). The cause for this dubious distinction is the stark disparity between Black and White residents in the metropolitan region when it comes to unemployment (15.3% for Black residents compared to 5.4% for White), arrest rates (nine times higher for Blacks than Whites), and poverty level (four times higher for Black residents than White residents).

In October 2020, Peoria, IL, was selected as one of five communities to be funded by the Illinois Children's Healthcare Foundation (ILCHF) to design and implement a children's mental health system of care (SOC) over the course of 4 years. ILCHF defines systems of care using the definition developed by Stroul et al. (2010): "a spectrum of effective, community-based services and supports for children and youth with or at risk for mental health or other challenges and their families, that is organized into a coordinated network, builds meaningful partnerships with families and youth, and addresses their cultural and linguistic needs, in order to help them to function better at home, in school, in the community, and throughout life." Children and youth with or at risk of mental health disorders and their families need support and services from many different child- and family-serving agencies. Often, these services are provided in a fragmented fashion. By creating partnerships and integration among agencies and organizations, systems of care are able to coordinate services and support to meet the ever-changing needs of children and families, which leads to improved outcomes. Whereas previous approaches to addressing at-risk children and youth focus on the efforts of disparate sectors, the system of care approach provides individualized services in accordance with the unique potential and needs of each individual child and family, providing individualized service plans which ensure cross-system collaboration with linkages between child-serving agencies and programs which cross administrative boundaries. Such systems of care aspire to incorporate continuous accountability mechanisms to track, monitor, and manage the achievement of a system of care goals, providing services and supports without regard to race, religion, national origin, gender, sexual orientation, disability, socioeconomic status, immigrant status, or other characteristics. However, despite best efforts the civic technologies which support systems of care are susceptible to bias and often require intentional mechanisms to increase accountability, particularly to the publics which are subjected to the increased surveillance and results of automated decision-making by software systems.

In the state of Illinois, the use of the Integrated Referral and Intake System (IRIS) developed by the University of Kansas significantly enhances the referral process across various sectors. IRIS serves as a comprehensive platform facilitating efficient referrals by integrating and organizing information about social services, healthcare

providers, educational institutions, and other essential resources. Through IRIS, agencies, professionals, and individuals can access a centralized database, enabling streamlined referrals for individuals in need of specific services or assistance. This system not only ensures quicker and more accurate referrals but also promotes collaboration among different entities, fostering a more interconnected network of support for Illinois residents seeking assistance and guidance. Its user-friendly interface and robust database play a pivotal role in improving the effectiveness and accessibility of referral services across the state.

A comprehensive case study analyzing the Integrated Referral and Intake System (IRIS) developed by the University of Kansas presents an intriguing opportunity to explore the integration of AI within social service systems, particularly in the realm of mental health services. Understanding local stakeholder needs and knowledge is pivotal in shaping the potential uses of AI within IRIS. AI integration intends to improve the system by offering predictive analytics to identify patterns in service utilization, optimize resource allocation based on demand, and provide personalized recommendations for mental health services. Additionally, AI-powered chatbots or virtual assistants could enhance user interactions, providing immediate support, information, and guidance to individuals seeking mental health resources through IRIS. This case study could serve as a blueprint for leveraging AI to augment the efficiency, accessibility, and effectiveness of social service systems, thereby improving mental health outcomes for individuals within the community.

However, the integration of AI into a social service system like IRIS also presents several challenges, especially concerning mental health services. One significant challenge is ensuring the ethical use of AI algorithms, particularly in sensitive areas like mental health, to avoid reinforcing biases or perpetuating stigmas. Maintaining data privacy and confidentiality is crucial, especially when dealing with sensitive patient information. Moreover, AI-driven decision-making in mental health services raises concerns about the potential loss of human touch and empathy, as well as the risk of misinterpretation or misdiagnosis due to the complexity of mental health conditions. Striking a balance between AI-driven insights and the expertise of mental health professionals is essential to harness the benefits of AI while ensuring that it complements, rather than replaces, human care and empathy in providing mental health services within systems like IRIS.

Although the research was interrupted because of turnover of key personnel before the participatory mapping and modeling phases of the research, we gained insights from a series of conversations across multiple planning meetings with various stakeholders. Specifically, in Peoria, although the system of care intended to incorporate many different agencies and stakeholders, the social service ecosystem can still seem very siloed with disparities in resources and services. Additionally, like many communities in the USA, the demographic dynamics include a growing minority population that leads to language gaps, trust gaps, and a gap between who the system is designed by and who it intends to serve.

The practitioners we talked with highlighted a number of barriers to the use of public service provisioning by community members as well as their own challenges in using the system and their own challenges in providing service. Regarding the

barriers for the public include logistic, social, and internal challenges. Logistically, many that needed public services lacked the means of transportation or time to work with public agencies to enroll and access services. Socially, the public servants highlighted that many of their clients had a language barrier, had a lack of trust in public services, or noticed a lack of cultural mirroring between those that needed service and those that provided the services. Regarding their own internal challenges in providing services, the public servants highlighted a lack of accountability, operating in a politically fraught environment with different groups protecting their turf or programs, employee burnout, and a high level of turnover in their agencies.

The high levels of turnover, organizational dynamics, and technological integration are not just a common challenge with the stability of any public service provisioning system, but they also motivated us to look more systematically about the broader environment of which Peoria is one example. The early insights from this case study prompt us to take a step back and explore the larger context in which this research is situated. Specifically exploring, what is the role of AI in the USA around social service provisioning and what is the role of AI when applied to mental health services in the USA and what are the associated concerns and challenges of both.

The Broader Environment for Using AI in the US Social Service Provisioning

The landscape of governmental operations in the USA has been undergoing a marked transformation with the increasing integration of AI and algorithmic systems across various functions. A noteworthy surge in the exploration and use of AI and related technologies has been witnessed not only within broader government functions but also in the intricacies of decision-making processes associated with social services (Crawford et al., 2019; Schiff et al., 2022).

In recognition of this technological shift, the US government has taken concrete steps to endorse the adoption of trustworthy AI within the federal government through an executive order (Executive Office of the President of the United States, 2020), directing each Federal agency to compile an "AI inventory" and enhance its expertise in AI, emphasizing principles including accuracy, reliability, safety, security, responsibility, resiliency, transparency, and accountability. In 2023, the executive order on AI safety and security was released, which provides guidance for Federal agencies to acquire AI products and talents (The White House, 2023).

Remarkably, nearly half (45%) of US federal agencies with more than 400 employees either are in the planning stages of implementing AI and other algorithmic techniques or have already put them into practice. Some departments, such as the Office of Justice Programs, have even gone a step further by deploying over 10 AI/ML programs (Engstrom et al., 2020). Even the US Patent Office (USPTO) has

embraced AI tools to augment the work of patent examiners, particularly in tasks related to searching and classifying patent applications (USPTO, 2021).

The deployment of these AI systems is strategically aimed at harnessing the vast data resources accumulated by the government. Their primary objectives encompass enhancing the quality and efficiency of public services, fostering greater responsiveness to policy changes (Margetts & Dorobantu, 2019), and nurturing public trust in government processes (Dwivedi et al., 2021). These AI solutions are being employed across a diverse spectrum of government functions, spanning healthcare, child welfare, social benefits, housing, education, immigration management, and poverty alleviation (AI Now Institute, 2018; Crawford et al., 2019; Nalbandian, 2022; Wirtz et al., 2019; Zuiderwijk et al., 2021).

AI in Identifying Individuals in Need of Social Services

The role of the government in providing essential social services in the USA is crucial. These services are designed to support individuals in various aspects of their lives, including healthcare, education, housing, income assistance, and more. At the federal, state, and local levels, government agencies craft and execute programs to ensure that vulnerable populations have access to essential resources and opportunities. Social service initiatives are instrumental in promoting equity and equality, improving overall well-being, and addressing societal challenges.

The process of delivering social services by governments typically traverses several phases, including the identification of needs, program development, implementation, monitoring, and continuous improvement (Chambers & Bonk, 2012). Within the implementation phase, each service journey typically consists of pre-service engagement and assessment, the actual service delivery, and post-service evaluation. Throughout each of these steps, a multitude of decision-making processes come into play, from the perspectives of both service providers and recipients.

Prior to the provision of a social service, an engagement process unfolds, initiated either by the government or the individuals, wherein automated decision-making tools have been developed and deployed. These tools serve as instrumental aids for government agencies in identifying individuals and communities potentially in need of specific services. These automated mechanisms often rely on risk scores computed by software programs, leveraging the extensive data repositories collected by various government agencies.

A notable example is the Allegheny Family Screening Tool (AFST), an algorithmic tool developed by Allegheny County, Pennsylvania, to determine which children are at risk of abuse or neglect, and whether intervention is necessary (Gerchick et al., 2023). The risk score of a child, ranging between 1 and 20, is computed using data such as medical records and interactions with juvenile probation systems. These scores help call screening workers make decisions regarding the necessity of investigating potential child neglect cases. Similarly, Los Angeles County, California, has adopted an AI-based program that mines data from county agencies,

aiding county case workers in identifying residents at risk of homelessness (Pimentel, 2023). A poignant example of its success lies in the case of a mother of nine who received vital assistance due to improved coordination resulting from the AI program, uniting two case workers who had previously operated independently on her case (Pimentel, 2023). Facing significant understaffed situation, Charleston County, South Carolina, has deployed AI service in 911 call center to answer calls and decide whether the callers' needs are emergent enough to dispatch law enforcement or just have the callers fill out online requests for non-emergency service (Grzeszczak, 2023). On the federal level, the Social Security Administration is experimenting with AI tools to help make decisions about who is eligible for disability benefits (Glaze et al., 2022).

In this critical engagement and assessment phase, in addition to identifying individuals in need of service, AI has also been instrumental in detecting fraudulent claims and guiding decisions on withholding service provision. Several federal agencies have led the way in pioneering AI tools for fraud detection, including the Securities and Exchange Commission's deployment of an insider-trading algorithm and the IRS's efforts to scrutinize tax documents and financial records for indications of tax evasion and other financial misconduct (Engstrom et al., 2020; Heckman, 2020). The Centers for Medicare & Medicaid Services (CMS) have committed resources and efforts to engage AI-driven solutions to "reshape the way [they] use data to make decisions."[10] Among their primary focuses is the utilization of AI to locate potential fraud, such as medical identity theft, which has witnessed a notable uptick during the COVID-19 pandemic (Heckman, 2020; Patterson, 2022; Whitfield, 2023). Michigan's Unemployment Insurance Agency employed the Michigan Integrated Data Automated System (MiDAS) to detect fraudulent unemployment claims (Giest & Klievink, 2022). However, it inadvertently led to false accusations of fraud against over 34,000 unemployed individuals (Charette, 2018).

AI in Public Service Delivery

Once engagement is established, decisions must be made regarding the nature, quality, timing, delivery methods, and other aspects of the service. In this process, algorithmic tools prove invaluable across a multitude of domains, enhancing service efficiency, accuracy, and accessibility. The applications of AI in public service delivery span a wide spectrum, from AI-powered chatbots providing immediate assistance to predictive analytics optimizing resource allocation.

Chatbots, a type of AI programs equipped with natural language processing (NLP) capabilities that enable human-like conversations through text or voice (Makasi et al., 2022), stand out as one of the most common AI-powered tools employed by governments in the USA and worldwide to facilitate public services.

[10] https://ai.cms.gov/

The ability of these tools to interpret questions and provide responses has been widely acknowledged by public service delivery agencies. Chatbots are primarily deployed to handle information requests from the public or, if necessary, to direct them to human assistance. Notable examples include New Orleans' "Jazz" 311 chatbot, San Jose's chatbot addressing COVID-related inquiries, Knoxville's Census chatbot (Pattison-Gordon, 2021), USCIS virtual assistant Emma, assisting with immigration services (USCIS, 2018), and Mississippi State's MISSI chatbot, offering support to residents, businesses, and visitors (Glasscock, 2020). Table 12.1 provides some examples of chatbots used by the US government.

Furthermore, automated decision-making tools are developed to determine how, when, and in what amount benefits are disbursed to individuals. For instance, the Idaho Department of Health and Welfare employs a formula to automatically decide the amount of Medicaid assistance provided to residents with developmental disabilities (AI Now Institute, 2018). Similarly, the Medicaid waiver program at the Arkansas Department of Human Services incorporates an algorithm to determine the allocation of home healthcare hours for individuals with disabilities (AI Now Institute, 2018; Lecher, 2018; McCormick, 2021).

Moreover, automated tools have been instrumental in connecting individuals with the essential services they require or improving the efficiency and quality of service. Los Angeles County's Housing for Health Program (HFH) utilizes a digital solution called the Coordinated Entry System to comprehensively assess the needs of homeless individuals and families, effectively linking them with appropriate housing resources (Gupta et al., 2020). Furthermore, through the implementation of the Content Classification Predictive Service (CCPS), an AI-powered tool incorporating NLP and ML technologies, the Department of Veteran Affairs (VA) has achieved a remarkable reduction in veterans' waiting times before receiving their entitled benefits (Barnett, 2020; DigitalVA, 2019; Makridis et al., 2021). In addition, the VA has also documented 41 distinct use cases where AI is deployed to support veterans,[11] spanning crucial areas such as suicide prevention, physical therapy, cardiac surgery, and diabetes prediction.

Concerns and Challenges of AI in Social Service Assessment and Delivery

Despite the promising potential of AI to enhance efficiency and cost-effectiveness within the public sector and social services, there are notable apprehensions surrounding the possible negative and unintended consequences associated with these applications. Notably among these concerns are issues related to potential biases and privacy issues. For example, the AFST program has faced criticism for its alleged practice of assessing a child's risk "by association" (Gerchick et al., 2023),

[11] https://www.research.va.gov/naii/ai-inventory.cfm

Table 12.1 Examples of chatbots deployed by the US government (Makasi et al., 2022)

Chatbot	Description	URL
Arkansas.gov AI assistant (Arkansas bot)	Arkansas State government (US) chatbot provides information on various services across the state government	www.portal.arkansas.gov/pages/chat-bot/
Dave	Arizona Department of Economic Security (US) chatbot for general information concerning public financial benefits	www.des.az.gov/services/employment/unemployment-individual
Miles	California Department of Motor Vehicles (US) chatbot provides assistance with locals' vehicles related services	https://www.dmv.ca.gov/portal/
CHIP (City Hall Internet Personality)	Los Angeles (US) chatbot provides information on local business opportunities	https://govlaunch.com/collections/chat/projects
EMMA	Department of Homeland Security (US) chatbot for citizenship and immigration service support	www.uscis.gov/tools/meet-emma-our-virtual-assistant
Dayne	Maryland Department of Labor (US) chatbot provides general information about unemployment benefits	www.dllr.state.md.us/employment/unemployment.shtml
MISSI	Government of Mississippi State (US) chatbot for supporting users with local online services	https://www.ms.gov/technology
Missouri Department of Labor virtual assistant (MDLva)	Missouri Department of Labor (US) chatbot provides information and resources for employment-related questions	https://info.mo.gov/labor/chatbot/
REAL ID bot	Montana Motor Vehicle Division (US) chatbot provides requirements information for obtaining an official ID card	https://mtrealid.gov/
DOLi	New Jersey Department of Labor (US) chatbot provides links to specific unemployment insurance services	https://nj.gov/labor/myunemployment/
IDES Assistant	Illinois Department of Employment Services (US) chatbot provides general information about unemployment insurance	www2.illinois.gov/ides/Pages/default.aspx

resulting in the disproportionate labeling of economically disadvantaged children as higher risk, while neglecting to screen more affluent families (Engstrom et al., 2020).

AI applications in this domain are often characterized by complexity and opacity, frequently functioning as black-box systems. These systems are sometimes implemented without necessary assessments, training, and oversight (AI Now Institute, 2018). For instance, automatic decision-making processes, particularly those tied to social benefits, are designed with a primary objective of reducing costs (AI Now Institute, 2018). Consequently, they inherently target populations

perceived as receiving more benefits, or those viewed as more "expensive" from the government's fiscal perspective. These targeted populations frequently belong to some of the most marginalized communities.

A pertinent example involves the previously mentioned algorithm employed by the Arkansas Department of Human Services to determine home care hours, which failed to understand the unique needs of patients with conditions such as cerebral palsy or diabetes. Furthermore, instances of human errors, such as categorizing an individual with double amputations as "not having a mobility problem" solely because they utilized a wheelchair, further exacerbated the situation. These adjustments, made ostensibly to enhance efficiency, resulted in the arbitrary reduction of care hours allocated to specific patients (AI Now Institute, 2018; McCormick, 2021). This reduction in care disproportionately affected low-income populations with disabilities across multiple states, including Pennsylvania, Iowa, New York, Maryland, New Jersey, and Arkansas (McCormick, 2021). As a result of legal battles, Arkansas ultimately terminated the usage of such an algorithmic system in 2018.

Central to the discussion of AI governance in this context is the crucial matter of public trust (Robles & Mallinson, 2023). The role of public attitudes toward AI and understanding of AI in government use is a topic that warrants comprehensive examination.

The Application of AI to Mental Health Services

The USA has witnessed the beneficial impact of computational and technological advancements on healthcare, including the domain of mental health. These advancements have led to the development and experimentation of new treatments, a deeper understanding of existing therapeutic approaches, the creation of novel medications, and the introduction of innovative predictive, analytical, and diagnostic methods (Frank & Glied, 2006).

The mental healthcare system within the US public sector stands as a unique and complex system, characterized by multi-level financing and management involving state and local governments. This system functions as a crucial "safety net" for the public, particularly to individuals without insurance or those with inadequate commercial healthcare plans (Hogan, 1999). Mental health services in the USA encompass a wide spectrum of programs, agencies, and focal points, with provision stemming from a combination of federal, state, and local initiatives. The integration of AI into the field of mental health has garnered significant research attention, both within the USA and on a global scale. AI has displayed tremendous potential across various dimensions of mental health, including prediction, monitoring, diagnosis, treatment, and assessment. Its widespread adoption is evident in mental health service applications such as personal sensing (commonly referred to as digital phenotyping), clinical text and social media content analysis, and the deployment of AI-driven chatbots (D'Alfonso, 2020).

At a service provider level, there is a breadth of research that has explored the implementation of AI into mental health services, diagnostics, and treatments. More specifically, there is evidence to suggest that AI can be beneficial for assisting with a number of mental health services needs, including accurate diagnosis of mental health disorders (Tutun et al., 2023). Specifically, machine learning has shown to be accurate in diagnosing ADHD by using electrophysiological data (Mueller et al., 2011), computer-assisted models show accuracy in diagnosing schizophrenia and related disorders (Razzouk et al., 2006), and they have also been successful at predicting depressive symptoms, suicidal ideation, and suicide attempts (Graham et al., 2019; Stewart et al., 2020). As research in this area has continued to expand, there seems to be some evidence to suggest that AI can possibly be a tool in the hands of mental health professionals. There is even some suggestion that we go beyond these applications of AI by utilizing phenotyping, natural language processing of clinical texts and social media content, and chatbots to assist with mental health services (D'Alfonso, 2020).

For instance, machine learning-based approaches have demonstrated notable proficiency in the real-time monitoring of psychological and behavioral changes (Ćosić et al., 2020, 2021) and the accurate prediction of mental health conditions, including conditions like Alzheimer's disease, schizophrenia, and anxiety (Baune, 2019; Grassi et al., 2019; Kalmady et al., 2019). AI's capacity to identify and diagnose psychotic disorders has also garnered recognition (Lejeune et al., 2022). Researchers have undertaken investigations employing diverse data sources, such as social media data, electroencephalogram (EEG) data, and facial expression data, to track individuals' mental states and facilitate the diagnosis of conditions such as depression and autism (Fei et al., 2020; Mumtaz et al., 2018; Wall et al., 2012).

Furthermore, AI-driven programs, with their inherent ability to interact with individuals, detect patterns, and respond adaptively, have spurred the development of innovative applications and experimental interventions in the field of mental health. This is particularly exemplified by the proliferation of AI-driven chatbots (Creed et al., 2022; Meheli et al., 2022; Omarov et al., 2023; Rathnayaka et al., 2022), as well as their utilization in specific therapeutic domains like ADHD intervention (Sibley et al., 2023) and music therapy (Lu, 2022).

Nonetheless, despite extensive research efforts, substantial barriers persist in the integration of AI into mental health practice. These barriers primarily stem from issues related to trust and confidence in the system, end-user acceptance, and system transparency. Further research is imperative to gain a comprehensive understanding of the potential role of AI in aiding decision-making within the realm of mental health practices. This necessitates an exploration of the attitudes and beliefs surrounding AI and its influence on end-users (Higgins et al., 2023).

While the public sector has recognized the potential benefits of AI in the realm of mental health, its current role primarily revolves around facilitating research, fostering collaboration, enacting regulatory measures, and overseeing approval processes, rather than direct implementation. For instance, in 2022, the National Institutes of Health (NIH) made a substantial investment of $130 million to launch

the Bridge2AI program,[12] aimed at expediting the adoption of AI in biomedical and behavioral research (NIH, 2022). Notably, this program has been a funding source for mental health-related research initiatives (Mudd, 2022).

Documentation of specific use cases involving the US government's application of AI within mental health services is rare, with only a handful of instances documented. Among these notable explorations, the Department of Veterans Affairs (VA) has emerged as an active participant, proactively utilizing technology, particularly AI, to enhance the mental health support provided to veterans. For instance, the Behavidence model app employs non-intrusive smartphone data and machine learning-based data mining to monitor veterans' mental health, with a specific focus on tracking their daily anxiety levels (Choudhary et al., 2022). Another noteworthy application is the SoKat Suicidal Ideation Detection Engine, which leverages NLP and large language models to identify instances of suicidal ideation among veterans (SoKat, 2023).

Concerns and Potential Consequences of Utilizing AI in Mental Health Services

While implementation of AI and computer models seems to show promise in application to clinical and systems in mental health, it is important to consider how this technology might have unintended consequences. There is some evidence to suggest that social and systemic bias exists in the current mental health system and that those biases introduce harm to patients. For example, Black men are more likely to be diagnosed with schizophrenia than White individuals, with some evidence suggesting that this is not due to genetic predisposition of the disorder in this population, but rather the introduction of clinician racial bias and client experiences of racism compounding symptoms and influencing diagnosis (Faber et al., 2023). Another example of this is the possible gender bias in the diagnosis of autism spectrum disorder, attention deficit hyperactivity disorder, and certain types of personality disorders (Garb, 2021). If we are modeling these technological systems on present diagnostic practices and training it on biased data, we risk reinforcing and exacerbating the consequences of bias already present in the current mental health system. Another consideration for the implementation of AI into mental health is how the technologies themselves might impact patients. For example, the use of technology to aid the treatment of individuals with psychosis can increase symptoms, such as paranoia (Bradstreet et al., 2019). Paying attention to these consequences will be important as we continue to research, design, and implement technologies such as AI into mental health services.

Carr (2020) suggests that one way to address these concerns is by including patients, service providers, and caregivers in the research and design of AI as a tool

[12] https://commonfund.nih.gov/bridge2ai

for mental health. By including those who will be most impacted, there may be an opportunity to address some of the biases before they are built into these systems and mitigate harm done by the already existing systemic issues in mental health. Additionally, it will be important to consider what datasets are used in training and assisting AI (Cirillo et al., 2020), how clinicians and AI will work together to make serious clinical decisions (Koutsouleris et al., 2022), and how we can best monitor these systems and make adaptations to them when needed.

Conclusion and Outlook

As public service provisioning continues to reorganize from decentralized services to coordination around systems of care and AI is integrated into more elements of mental health service delivery, ongoing reflections of who designs, uses, and benefits from these systems are necessary to promote improvements in performance and accountabilities. This chapter explores the use of systems of care in Peoria, Illinois, for coordinating public service provisioning across multiple organizations serving vulnerable populations. Practitioners identified barriers for the public including the insight that the public is more likely to trust and subsequently use a system if it reflects them. How to engender trust in the use of AI systems that regularly feel like a black-box created by distant experts is an ongoing challenge that deserves more attention. Additionally, with the inclusion of AI elements at different parts of the overall system of care, it is important to recognize and study how errors or biases introduced at one part of the system might propagate, exacerbate, or be contained across the system. Bias present anywhere in the system is a threat of bias everywhere throughout the system.

Acknowledgments Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

The ASU Knowledge Exchange for Resilience is supported by Virginia G. Piper Charitable Trust. Piper Trust supports organizations that enrich health, well-being, and opportunity for the people of Maricopa County, Arizona. The conclusions, views, and opinions expressed in this chapter are those of the authors and do not necessarily reflect the official policy or position of the Virginia G. Piper Charitable Trust. This material is based upon work supported by the National Science Foundation under Grant No. 2217706.

References

AI Now Institute. (2018). *Litigating algorithms: Challenging government use of algorithmic decision systems*. AI Now Institute. http://www.law.nyu.edu/sites/default/files/litigatingalgorithms_0.pdf.

Atkins, M. (2002). *School-wide systems of positive behavioral support promoting the mental health of all students, including those with SED*. Symposium, Distributed by ERIC Clearinghouse.

Barnett, J. (2020). *By using AI, the VA dramatically decreased claims processing intake times, official says*. FedScoop. https://fedscoop.com/veterans-benefits-ai-mail-processing/.

Baune, B. (2019). *Personalized mental health: Artificial intelligence technologies for treatment response prediction in anxiety. In Personalized psychiatry*. Elsevier Science & Technology.

Bradstreet, S., Allan, S., & Gumley, A. (2019). Adverse event monitoring in mHealth for psychosis interventions provides an important opportunity for learning. *Journal of Mental Health, 28*(5), 461–466. https://doi.org/10.1080/09638237.2019.1630727

Brewer, A. G., Davis, M. M., Sheehan, K., & Feinglass, J. (2020). Sociodemographic characteristics associated with hospitalizations for anxiety and depression among youth in Illinois. *Academic Pediatrics, 20*(8), 1133–1139. https://doi.org/10.1016/j.acap.2020.01.009

Brewer, A. G., Doss, W., Sheehan, K. M., Davis, M. M., & Feinglass, J. M. (2022). Trends in suicidal ideation-related emergency department visits for youth in Illinois: 2016–2021. *Pediatrics, 150*(6), e2022056793. https://doi.org/10.1542/peds.2022-056793

Burnett-Ziegler, I., Brennen, J., & Jackson, C. (2009). Illinois's Mental Health Juvenile Justice Initiative: Use of standardized assessments for eligibility and outcomes. In *Behavioral health care: Assessment, service planning, and total clinical outcomes management* (pp. 1–20). Civic Research Institute.

Carr, S. (2020). 'AI gone mental': Engagement and ethics in data-driven technology for mental health. *Journal of Mental Health, 29*(2), 125–130. https://doi.org/10.1080/09638237.2020.1714011

Chambers, D., & Bonk, J. (2012). *Social policy and social programs: A method for the practical public policy analyst* (6th ed.). Pearson.

Charette, R. N. (2018, January 24). Michigan's MiDAS unemployment system: Algorithm alchemy created lead, not gold: A case study into how to automate false accusations of fraud for more than 34,000 unemployed people. *IEEE Spectrum*. https://spectrum.ieee.org/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold

Choudhary, S., Thomas, N., Alshamrani, S., Srinivasan, G., Ellenberger, J., Nawaz, U., & Cohen, R. (2022). A machine learning approach for continuous mining of nonidentifiable smartphone data to create a novel digital biomarker detecting generalized anxiety disorder: Prospective cohort study. *JMIR Medical Informatics, 10*(8), e38943. https://doi.org/10.2196/38943

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine, 3*(1), Article 1. https://doi.org/10.1038/s41746-020-0288-5

Comen, E. (2019, November 5). The worst cities for Black Americans. *24/7 Wallstreet*. https://247wallst.com/special-report/2019/11/05/the-worst-cities-for-black-americans-5/3/

Ćosić, K., Popović, S., Šarlija, M., Kesedžić, I., & Jovanovic, T. (2020). Artificial intelligence in prediction of mental health disorders induced by the COVID-19 pandemic among health care workers. *Croatian Medical Journal, 61*(3), 279–288. https://doi.org/10.3325/cmj.2020.61.279

Ćosić, K., Popović, S., Šarlija, M., Kesedžić, I., Gambiraža, M., Dropuljić, B., Mijić, I., Henigsberg, N., & Jovanovic, T. (2021). AI-based prediction and prevention of psychological and behavioral changes in ex-COVID-19 patients. *Frontiers in Psychology, 12*, 782866–782866. https://doi.org/10.3389/fpsyg.2021.782866

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., Sánchez, A. N., Raji, D., Rankin, J. L., Richardson, R., Schultz, J., West, S. M., & Whittaker, M. (2019). *AI now 2019 report*. AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.html.

Creed, T. A., Kuo, P. B., Oziel, R., Reich, D., Thomas, M., O'Connor, S., Imel, Z. E., Hirsch, T., Narayanan, S., & Atkins, D. C. (2022). Knowledge and attitudes toward an artificial intelligence-based fidelity measurement in community cognitive behavioral therapy supervi-

sion. *Administration and Policy in Mental Health and Mental Health Services Research, 49*(3), 343–356. https://doi.org/10.1007/s10488-021-01167-x

Curtin, S. C. (2020). *State suicide rates among adolescents and young adults aged 10–24: United States, 2000–2018* (69, 11; National Vital Statistics Reports). https://stacks.cdc.gov/view/cdc/93667

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology, 36*, 112–117. https://doi.org/10.1016/j.copsyc.2020.04.005

DigitalVA. (2019, December 27). *VA Launches smart tool to reduce veteran wait times for disability claims [DigitalVA]*. https://digital.va.gov/general/va-launches-smart-tool-to-reduce-veteran-wait-times-for-disability-claims/

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., et al. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Economic Innovation Group. (2020). *The space between us: The evolution of American communities in the new century*. Economic Innovation Group https://eig.org/wp-content/uploads/2020/10/EIG-2020-DCI-Report.pdf

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies* (SSRN Scholarly Paper 3551505). doi:https://doi.org/10.2139/ssrn.3551505.

Executive Office of the President of the United States. (2020). *Promoting the use of trustworthy artificial intelligence in the Federal Government*. https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

Faber, S. C., Khanna Roy, A., Michaels, T. I., & Williams, M. T. (2023). The weaponization of medicine: Early psychosis in the Black community and the need for racially informed mental healthcare. *Frontiers in Psychiatry, 14*, 1098292. https://doi.org/10.3389/fpsyt.2023.1098292

Fei, Z., Yang, E., Li, D. D.-U., Butler, S., Ijomah, W., Li, X., & Zhou, H. (2020). Deep convolution network based emotion analysis towards mental health care. *Neurocomputing (Amsterdam), 388*, 212–227. https://doi.org/10.1016/j.neucom.2020.01.034

Frank, R. G., & Glied, S. A. (2006). *Better but not well: Mental health policy in the United States since 1950* (pp. xv–xv). Johns Hopkins University Press.

Garb, H. N. (2021). Race bias and gender bias in the diagnosis of psychological disorders. *Clinical Psychology Review, 90*, 102087. https://doi.org/10.1016/j.cpr.2021.102087

Gerchick, M., Jegede, T., Shah, T., Gutierrez, A., Beiers, S., Shemtov, N., Xu, K., Samant, A., & Horowitz, A. (2023). The devil is in the details: Interrogating values embedded in the Allegheny family screening tool. *2023 ACM Conference on fairness, accountability, and transparency* (pp. 1292–1310). doi:https://doi.org/10.1145/3593013.3594081.

Giest, S. N., & Klievink, B. (2022). More than a digital system: How AI is changing the role of bureaucrats in different organizational contexts. *Public Management Review, 0*(0), 1–20. https://doi.org/10.1080/14719037.2022.2095001

Glasscock, A. (2020). *Chat with us: How states are using chatbots to respond to the demands of COVID-19 (United States of America) [Report]*. National Association of State Chief Information Officers. https://apo.org.au/node/319464.

Glaze, K., Ho, D. E., Ray, G. K., & Tsang, C. (2022). Artificial intelligence for adjudication: The social security administration and AI governance. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford handbook of AI governance* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.46

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports, 21*(11), 116. https://doi.org/10.1007/s11920-019-1094-0

Grassi, M., Loewenstein, D. A., Caldirola, D., Schruers, K., Duara, R., & Perna, G. (2019). A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: Further evidence of its accuracy via a transfer learning approach. *International Psychogeriatrics, 31*(7), 937–945. https://doi.org/10.1017/S1041610218001618

Grzeszczak, J. (2023, April 27). *Charleston County's 911 center using AI, other tactics to solve staffing shortage*. Post and Courier. https://www.postandcourier.com/news/charleston-countys-911-center-using-ai-other-tactics-to-solve-staffing-shortage/article_db800620-e436-11ed-b8a8-e3c3b009df40.html

Gupta, R., Ghaly, M., Todoroff, C., & Wali, S. (2020). Creating value for communities: Los Angeles County's investment in Housing for Health. *Healthcare, 8*(1), 100387. https://doi.org/10.1016/j.hjdsi.2019.100387

Heckman, J. (2020, October 27). CMS untangles its data infrastructure to enable AI-powered fraud detection |Federal News Network. *Federal News Network* https://federalnewsnetwork.com/automation/2020/10/cms-untangles-its-data-infrastructure-to-enable-ai-powered-fraud-detection/

Higgins, O., Short, B. L., Chalup, S. K., & Wilson, R. L. (2023). Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review. *International Journal of Mental Health Nursing, 32*(4), 966–978. https://doi.org/10.1111/inm.13114

Hogan, M. F. (1999). Perspective: Public-sector mental health care: New challenges. *Health Affairs, 18*(5), 106–111. https://doi.org/10.1377/hlthaff.18.5.106

ICMHP. (2020). *Illinois chidren's mental health plan (2022–2027). Illinois created the Illinois Children's Mental Health Partnership*. https://www.icmhp.org/wp-content/uploads/2022/06/FINAL_ICMHP-2022-Childrens-Mental-Health-Plan_rev1-1.pdf

Illinois State Board of Education, S. (1991). *The Illinois 9th grade adolescent health survey. Full Report*. Distributed by ERIC Clearinghouse.

Kalmady, S. V., Greiner, R., Agrawal, R., Shivakumar, V., Narayanaswamy, J. C., Brown, M. R. G., Greenshaw, A. J., Dursun, S. M., & Venkatasubramanian, G. (2019). Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *NPJ Schizophrenia, 5*(1), 2–2. https://doi.org/10.1038/s41537-018-0070-8

Koutsouleris, N., Hauser, T. U., Skvortsova, V., & De Choudhury, M. (2022). From promise to practice: Towards the realisation of AI-informed mental health care. *The Lancet. Digital Health, 4*(11), e829–e840. https://doi.org/10.1016/S2589-7500(22)00153-4

Lecher, C. (2018, March 21). A healthcare algorithm started cutting care, and no one knew why—The Verge. *The Verge*. https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy

Lejeune, A., Robaglia, B.-M., Walter, M., Berrouiguet, S., & Lemey, C. (2022). Use of social media data to diagnose and monitor psychotic disorders: Systematic review. *Journal of Medical Internet Research, 24*(9), e36986–e36986. https://doi.org/10.2196/36986

Lu, D. (2022). Evaluation model of music therapy's auxiliary effect on mental health based on artificial intelligence technology. *Journal of Environmental and Public Health, 2022*, 1–10. https://doi.org/10.1155/2022/9960589

Lyons, J. S., Griffin, G., Quintenz, S., Jenuwine, M., & Shasha, M. (2003). Clinical and forensic outcomes from the Illinois mental health juvenile justice initiative. *Psychiatric Services, 54*(12), 1629–1634. https://doi.org/10.1176/appi.ps.54.12.1629

Makasi, T., Nili, A., Desouza, K. C., & Tate, M. (2022). A typology of chatbots in public service delivery. *IEEE Software, 39*(3), 58–66. https://doi.org/10.1109/MS.2021.3073674

Makridis, C., Hurley, S., Klote, M., & Alterovitz, G. (2021). Ethical applications of artificial intelligence: Evidence from health research on veterans. *JMIR Medical Informatics, 9*(6), e28921. https://doi.org/10.2196/28921

Margetts, H., & Dorobantu, C. (2019). Rethink government with AI. *Nature (London), 568*(7751), 163–165. https://doi.org/10.1038/d41586-019-01099-5

Masterson, L. (2023, May 1). The worst states for mental health care, ranked – Forbes Advisor. *Forbes Advisor*. https://www.forbes.com/advisor/health-insurance/worst-states-for-mental-health-care/

McCormick, E. (2021, July 2). What happened when a 'wildly irrational' algorithm made crucial healthcare decisions. *The Guardian*. https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions

Meheli, S., Sinha, C., & Kadaba, M. (2022). Understanding people with chronic pain who use a cognitive behavioral therapy-based artificial intelligence mental health app (Wysa): Mixed methods retrospective observational study. *JMIR Human Factors, 9*(2), e35671–e35671. https://doi.org/10.2196/35671

Mudd, L. M. (2022). *Bridge to artificial intelligence (Bridge2AI).* NIH. https://dpcpsi.nih.gov/sites/default/files/Bridge2AI.pdf.

Mueller, A., Candrian, G., Grane, V. A., Kropotov, J. D., Ponomarev, V. A., & Baschera, G.-M. (2011). Discriminating between ADHD adults and controls using independent ERP components and a support vector machine: A validation study. *Nonlinear Biomedical Physics, 5*, 5. https://doi.org/10.1186/1753-4631-5-5

Mumtaz, W., Ali, S. S. A., Yasin, M. A. M., & Malik, A. S. (2018). A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD). *Medical & Biological Engineering and Computing, 56*(2), 233–246. https://doi.org/10.1007/s11517-017-1685-z

Nalbandian, L. (2022). An eye for an 'I:' A critical assessment of artificial intelligence tools in migration and asylum management. *Comparative Migration Studies, 10*(1), 32–32. https://doi.org/10.1186/s40878-022-00305-0

NIH. (2022, September 13). *NIH launches Bridge2AI program to expand the use of artificial intelligence in biomedical and behavioral research*. National Institutes of Health (NIH). https://www.nih.gov/news-events/news-releases/nih-launches-bridge2ai-program-expand-use-artificial-intelligence-biomedical-behavioral-research

NTAEC. (2009). *A closer look: An overview of systems of care in child welfare*. The National Technical Assistance and Evaluation Center. https://www.childwelfare.gov/pubPDFs/overview.pdf

Omarov, B., Zhumanov, Z., Gumar, A., & Kuntunova, L. (2023). Artificial intelligence enabled mobile chatbot psychologist using AIML and cognitive behavioral therapy. *International Journal of Advanced Computer Science and Applications, 14*(6). https://doi.org/10.14569/IJACSA.2023.0140616

Patterson, A. (2022, September 21). *Data analytics, data sharing help combat fraud at health agencies*. GovCio. https://governmentciomedia.com/data-analytics-data-sharing-help-combat-fraud-health-agencies.

Pattison-Gordon, J. (2021, June 25). *New Orleans launches its AI-Powered, Textable 311 Chatbot*. GovTech. https://www.govtech.com/computing/new-orleans-launches-its-ai-powered-textable-311-chatbot

Pimentel, B. (2023, February 9). *How AI can help fix its homelessness in San Francisco*. San Francisco Examiner. https://www.sfexaminer.com/news/how-san-francisco-could-use-ai-to-help-fix-its-homelessness/article_2f67b464-a836-11ed-b90a-4f1ccef82528.html.

Rathnayaka, P., Mills, N., Burnett, D., De Silva, D., Alahakoon, D., & Gray, R. (2022). A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. *Sensors (Basel, Switzerland), 22*(10), 3653. https://doi.org/10.3390/s22103653

Rawal, P., Romansky, J., Jenuwine, M., & Lyons, J. S. (2004). Racial differences in the mental health needs and service utilization of youth in the juvenile justice system. *The Journal of Behavioral Health Services & Research, 31*(3), 242–254. https://doi.org/10.1007/bf02287288

Razzouk, D., Mari, J. J., Shirakawa, I., Wainer, J., & Sigulem, D. (2006). Decision support system for the diagnosis of schizophrenia disorders. *Brazilian Journal of Medical and Biological*

Research = Revista Brasileira De Pesquisas Medicas E Biologicas, 39(1), 119–128. https://doi.org/10.1590/s0100-879x2006000100014

Robles, P., & Mallinson, D. J. (2023). Artificial intelligence technology, public trust, and effective governance. *The Review of Policy Research*. https://doi.org/10.1111/ropr.12555

SAMHSA. (2020). *Behavioral Health Barometer, Illinois, Volumn 6 (HHS Publication No. SMA–20–Baro–19–IL; Behavioral Health Barometer: Illinois, Volume 6: Indicators as Measured through the 2019 National Survey on Drug Use and Health and the National Survey of Substance Abuse Treatment Services)*. Substance Abuse and Mental Health Services Administration. https://www.samhsa.gov/data/sites/default/files/reports/rpt32830/Illinois-BH-Barometer_Volume6.pdf

Schiff, D. S., Schiff, K. J., & Pierson, P. (2022). Assessing public value failure in government adoption of artificial intelligence. *Public Administration (London), 100*(3), 653–673. https://doi.org/10.1111/padm.12742

Sibley, M. H., Bickman, L., Atkins, D., Tanana, M., Coxe, S., Ortiz, M., Martin, P., King, J., Monroy, J. M., Ponce, T., Cheng, J., Pace, B., Zhao, X., Chawla, V., & Page, T. F. (2023). Developing an implementation model for ADHD intervention in community clinics: Leveraging artificial intelligence and digital technology. *Cognitive and Behavioral Practice*. https://doi.org/10.1016/j.cbpra.2023.02.001

SoKat. (2023). *SoKat Suicide Ideation Engine*. SoKat. https://www.sokat.com/copy-of-va-mission-daybreak

Starin, A. C., Atkins, M. S., Wehrmann, K. C., Mehta, T., Hesson-McInnis, M. S., Marinez-Lora, A., & Mehlinger, R. (2014). Moving science into state child and adolescent mental health systems: Illinois' evidence-informed practice initiative. *Journal of Clinical Child and Adolescent Psychology, 43*(2), 169–178. https://doi.org/10.1080/15374416.2013.848772

Stewart, S. L., Celebre, A., Hirdes, J. P., & Poss, J. W. (2020). Risk of suicide and self-harm in kids: The development of an algorithm to identify high-risk individuals within the children's mental health system. *Child Psychiatry and Human Development, 51*(6), 913–924. https://doi.org/10.1007/s10578-020-00968-9

Stroul, B. A., Blau, G. M., & Friedman, R. M. (2010). *Updating the system of care concept and philosophy*. National Technical Assistance Center for Children's Mental Health. https://portal.ct.gov/-/media/DCF/Mental_Health/pdf/UpdatingTheSOCConcept2010pdf.pdf

The White House. (2023, October 30). *Executive order on the safe, secure, and trustworthy development and use of artificial intelligence*. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Tutun, S., Johnson, M. E., Ahmed, A., Albizri, A., Irgil, S., Yesilkaya, I., Ucar, E. N., Sengun, T., & Harfouche, A. (2023). An AI-based decision support system for predicting mental health disorders. *Information Systems Frontiers, 25*(3), 1261–1276. https://doi.org/10.1007/s10796-022-10282-5

USCIS. (2018, April 13). *Meet Emma, our virtual assistant* | USCIS. https://www.uscis.gov/tools/meet-emma-our-virtual-assistant

USPTO. (2021, March 18). *Artificial intelligence tools at the USPTO*. USPTO. https://www.uspto.gov/blog/director/.

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., & Deluca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS One, 7*(8), e43855–e43855. https://doi.org/10.1371/journal.pone.0043855

Whitfield, J. (2023, May 22). *How health tech leaders use AI to combat fraud*. GovCio. https://governmentciomedia.com/how-health-tech-leaders-use-ai-combat-fraud.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration, 42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly, 38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577

Chapter 13
Towards Culture-Sensitive, Responsive, and Participatory AI

Petra Ahrweiler

Abstract This chapter evaluates the contribution of research in this volume towards culture-sensitive, responsive, and participatory AI for social assessment in public service provision. Research offers some opportunities for a cross-cultural comparison among case studies concerning AI-based social assessment: The chapter starts by discussing some more general commonalities and differences in terms of value cultures from a nation state perspective. After this, the comparison of case studies and their results will evaluate the actual findings regarding the hypotheses and questions of AI FORA. Insights concerning the utility of including vulnerable groups in participatory research will be compared across case studies, as well as findings about challenges for existing or future AI use in social assessment for public service distribution. Then, the chapter assess research experiences with the Safe Spaces concept for inclusive technology co-design from the point of view of Safe Space coordinators, of participants in Safe Space workshops, and of AI FORA researchers. After this, the chapter discusses impact paths of research for its two target audiences, i.e. policy and the general public. Finally, the last evaluation exercise of this chapter consists in an ex ante assessment of the future, i.e. the upcoming modelling and simulation work in the case studies for improved AI-based social service provision in national welfare systems.

AI Assessing People: People Assessing AI

As demonstrated by the previous chapters, AI is already in place in many countries to support social assessment, deciding on provision or refusal of public social services. The case studies presented in this volume analysed social assessment practices in nine countries with different value reference frameworks, investigating the

P. Ahrweiler (✉)
Department of Technology- and Innovation Sociology, Social Simulation (TISSS Lab), Institute of Sociology, Johannes Gutenberg-University Mainz, Mainz, Germany
e-mail: petra.ahrweiler@uni-mainz.de

P. Ahrweiler (ed.), *Participatory Artificial Intelligence in Public Social Services*, Artificial Intelligence, Simulation and Society,
https://doi.org/10.1007/978-3-031-71678-2_13

context-dependent service provision selectivity of their national welfare systems in order to identify the requirements, constraints, and consequences of the use of AI to support or replace conventional practices. In empirical research, Participatory Mapping was used to identify and analyse actors, values, resources, inputs, outputs, processes, AI in use, policies in use, performance, stakeholders, networks, etc. Context, especially in terms of the cultural and social embeddedness of values and attitudes, proved everywhere to be key to understanding the current rulesets for social assessment used for distributing social services in existing systems. Some case studies already started to explore requirements to system changes from existing towards more desired systems in participatory deliberation processes with stakeholders.

AI assessing people—people assessing AI: Where are we now in the sequence of AI FORA's assessment exercise? What are the limitations of our work, what are the lessons learnt so far, and where to go from here? This concluding chapter is set out to address these questions.

However, it can only provide a mid-term "assessment of the assessor" (the latter, in this case, is the AI FORA project) because the part on modelling and simulation, using all these collected data for scenario analysis and policy recommendations, is still outstanding (see Volume II). Furthermore, a comprehensive output, outcome, and impact assessment of the whole project needs to give time for knowledge transfer, dissemination, uptake of results, and potential implementation of recommendations. Effects might only be seen as a long-term legacy of the project. Thus, currently only work covered in this volume can be evaluated for its immediate scientific contribution, for its experiences with the research process, and for its utility to connect to following work. A continuous feature of this mid-term assessment will focus on limitations of research produced so far.

Valuation consists of a "triad of valuator, valuee, and audience" (Waibel et al., 2021: 35), translated for this case as the assessor, the assessed, and the project's target audiences. The assessor in this case is the coordinator of the AI FORA project, which makes this review not only subjective but also a case of project-internal self-assessment. The audiences will be discussed at the end of the chapter. Assessed will be here, on the one hand, the project's contribution to the conceptual work outlined at the beginning of this volume, and, on the other hand, the process experiences and scientific results of empirical case studies.

The chapter is structured as follows: After assessing the chances for a general cross-cultural comparative approach to discuss results of country cases concerning AI-based social assessment and coming back to the cross-cultural approaches of Hofstede (2003) and Inglehart-Welzel (2005), the most interesting commonalities and differences of our case studies will be discussed in more detail. This will reveal what we have learnt from applying a more local lens about the relationship between universality and particularity, i.e. observations and concepts that cross all chapters and are easily transferable to other cases for universality (Kitcher, 1993), and knowledge reported as highly contextualised bound to locations and situations as particularity (Knorr-Cetina, 1981). After this, the chapter assesses research experiences with participatory methods, especially the Safe Spaces, for inclusive

technology co-design. The chapter turns then to knowledge transfer, i.e. dissemination of results to the target audiences of research evaluating their points of access for benefitting from project results. The section starts with an assessment of policy learning with the policy community as primary target audience; this is followed by an assessment of communication requirements with regard to other expert communities, and, particularly, to what is called "the general public". Finally, the last assessment exercise of this chapter consists in an ex ante assessment of the future, i.e. the upcoming modelling and simulation work: The outlook provided will discuss whether and how the picture can be completed to produce knowledge on the role of culture and context for AI-based social service provision in national welfare systems.

Assessing the Comparative Approach

Case study countries were originally chosen to capture as much cultural heterogeneity as possible from a nation state perspective (see Chap. 1) following the Ingelhart/Welzel cultural map supported by a factor analysis from the Hofstede dimensions. They differed in the Inglehart/Welzel coordinate system of Survival vs. Self-Expression Values and Traditional vs. Secular-Rational Values, in the six Hofstede dimensions as explained above, and in three more general categories (level of digitisation, government form, and country size).

These cross-cultural comparison frameworks with their focus on the nation state were mainly used for selecting case study countries and motivating this choice. However, is it generally possible to check for the assumed cultural differences on this high granular level and enable a generic comparison between case studies using the specific methodologies of AI FORA? During research, some experiments were set up to address this question.[1]

The objective of these experiments was not to generally and systematically test cross-cultural measurement frameworks; their limitations for AI FORA's research had led to the development of a more local and behaviour-centred approach (cf. Chap. 1). Rather, experiments were set up to check for basic assumptions of cultural heterogeneity in a small sample of cases.

Hofstede's six-dimensional framework was implemented as an agent-based model (ABM) of cultural-dependent decision-making, and the ABM was translated into a serious game for human players (Ahrweiler et al., 2024; Szczepanska et al., 2022) to be played in different national contexts. Games were used not only to see whether the same game played in one national context produced different outcomes in another context, but also whether the outcome was the one suggested by the Hofstede dimensions when the decision-making situations in the game were

[1] The role of experiments in social research, and especially the status of computer simulations as experiments, is extensively discussed in Ahrweiler (2017).

value-coded according to the cultural comparison framework. People are drawing in their decision-making on their prior knowledge, beliefs, and systems of ethics; therefore, a gamification approach is useful for examining the relationship between AI service provision and Hofstede's six cultural dimensions. Thus, integrating insights from gamification can allow for a more complete understanding of national cultural differences and how they relate to social assessment routines for distributing social services organised and institutionalised in different societies.

Of course, this approach also reveals much heterogeneity. Social service provision is a problem context everybody is exposed to, with each of us deciding how to behave faced with specific situations. Much of the decision-making is related to cultural background, but there is a huge variety of all kinds of behaviours within a culture due to individual or group characteristics, for example, very fundamentally, being committed to societally shared norms and values, or defecting them. However, it is not about liking or disliking cultural norms: The task is to connect an individual's behaviour to the wider cultural context of its society, i.e. the social meanings and understandings that are deeply rooted in its members. An illustrative and popular example for this type of approach in interpretative sociology (David, 2010), which does not need samples or representativity, e.g. in who and how many play a serious game, is the 1992 novel by Peter Hoeg *Miss Smilla's Feeling for Snow*:[2] The title centres around the phenomenon that Scandinavians have many more different notions for "snow" than non-Scandinavians have due to their shared experience. The society around is responsible for object formation: It let people see things; people from other societal backgrounds would not even notice because they are meaningless for them. Members of a certain society might differ in their approaches, perceptions, worldviews, opinions, attitudes, and approaches, but they share meaningfulness. The societal "footprint", however, can be found in every member regardless of personal preferences.

Gamification, i.e., applying game elements in non-game contexts (Strahringer & Leyh, 2017), is a low-threshold entry point for non-scientists to contribute to research.

- A serious game will implement a problem context as a game with anonymous players sharing a certain cultural background.
- There will be "stations" where players make decisions that are recorded in a quantitative database.
- The stations will be "value-coded" capturing the conflicting value propositions behind the decisions; station chairs (members of the scientific team) will record the conversations of players and other qualitative data to check quantitative outcome.
- Value coding of stations and value decisions of players will be checked against the results of a value survey following the example of the Ingle-Welzel questionnaire that is run among the players afterwards.
- Further checks do follow (interviews with players, qualitative data of chairs, etc.).

[2] Also set as a mystery thriller film in 1997 directed by Bille August; starring Julia Ormond.

- The game can be played in all cultural contexts, and it will make a difference where it is played.
- It can be played by everybody, regardless of group membership, background, or personal attributes (a medium group size of 20 is advisable, less would not display interesting commonalities, more would rather serve statistical purposes).
- Though each player will have individual decision choice, the game will display the context-bound societal value sets through the corridor choices checked by the qualitative interpretations of coding.
- Games create a controlled setting with observability, measurability, and comparability.

Before developing a game about social service provision, the general feasibility to value-code Hofstede dimensions into a serious game was tested using a decision context, which was most prominent in the minds of players at that time, and where everybody had immediate and same exposure: the everyday decision context in the acute COVID-19 crisis.

Stations of the game to simulate everyday life during the pandemic were Home, Office, Supermarket, Bank, School, Lounge, and Townhall. Players could get infected at public places but needed money and happiness to survive. Different stations offered different payoffs and costs/risks. Players could stay at Home, go to work in the Office to earn money, go to the Supermarket to buy supplies, go to School to get a good office job, go to the Lounge and get happiness, go to the Bank and invest, or go to the Townhall to change policies, i.e. rules of the game. Stations, or more precisely decisions players could make at the stations, were value-coded according to Hofstede dimensions. The Individualism-Collectivism and Motivation towards Achievement and Success (formerly called Masculinity-Feminity) dimension, for example, was value-coded at the Bank and the Townhall. At the Bank, players could choose between various investment decisions: to invest in stock funds, to invest in vaccination research, to invest in personal virus protection, or to donate money for infected people. Data collection at this station revealed whether a group of players systematically preferred to invest into their individual benefit or into group benefit, i.e. whether their behaviour reflected more individualistic or collectivistic values. In the Townhall, as another example, players could discuss submit petitions, deliberate, and vote on them. If successful, they could alter the game dynamics (e.g. they could introduce a social welfare system with income tax and health care, etc.). With this, it was possible to measure whether a group of players had more egalitarian or elitist values.

The game was played by case study groups in Spain and Germany. Results revealed general differences and confirmed the Hofstede factor analysis as displayed in Fig. 13.1. While Spanish players focused on inclusion and communication played a central role during the game, German players focused on efficient behaviour, group discipline, and coordination.

Insights of the “Corona Game” confirmed the general feasibility to value-code a serious game according to the Hofstede dimensions and to use the game for data collection on the role of culture and context for complex decision-making.

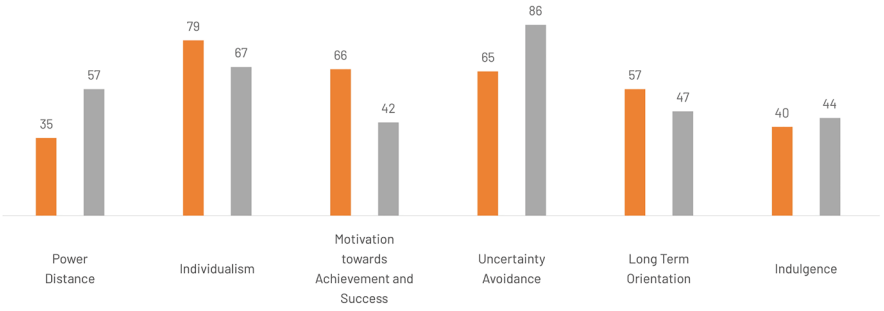

Fig. 13.1 Hofstede comparison of Germany (orange) and Spain (grey) according to the country comparison tool at https://www.hofstede-insights.com/ (last accessed 11.06.2024)

This led to the development of a game about AI-based social assessment for public service provision. The game built on the basic structure of the previous context but simulated a job market in non-COVID times. Familiar stations transferred from the old game were Home, Office, Bank, Lounge, School, and Townhall. The central new station was a Job Agency where AI-based or non-AI-based social assessment took place. Players had personal attributes such as age, gender, education status, family status, and area of residence according to assigned "identity cards" and had a good job, a bad job, or no job (for game dynamics, please see the section on China). They could find out about existing bias and discrimination, choose between AI-based or human assessment, and change the assessment algorithm in the Job Agency after discussion and deliberation. Again, most stations were value-coded according to the Hofstede dimensions. Besides with Chinese players, the "Unemployment Game" was also played with Spanish, Ukrainian, and German players from the AI FORA case study set. Again, some differences in comparing general cultural features to be expected according to Fig. 13.2 could be confirmed by data collected in the games.

However, results were ambiguous. For example, the hypotheses around power distance where China scored high suggested that Chinese people would accept inequalities among people concerning well-being as well as power hierarchies easier than Western countries, that there would be no strong incentive to change current distributions and algorithms, and that there would be no action focus on vulnerable groups. This could not be confirmed by the game results. Furthermore, in the individualism dimension where China scored low, the hypotheses that egoistic strategies (self-optimisation) would be not accepted and that system optimisation would be favoured at all times could not be supported by game results as well.

Outside of game experiments, the Hofstede dimensions were sometimes used to illustrate case study results. As demonstrated, for example, by the case study chapter of Spain, the Hofstede dimensions were referred to for explaining adoption behaviour of AI in Catalonian social services, especially concerning the Fear of Missing Out (FOMO): "At the national level, cultural dimensions may further influence both AI adoption and the manifestation of FOMO in social services. For

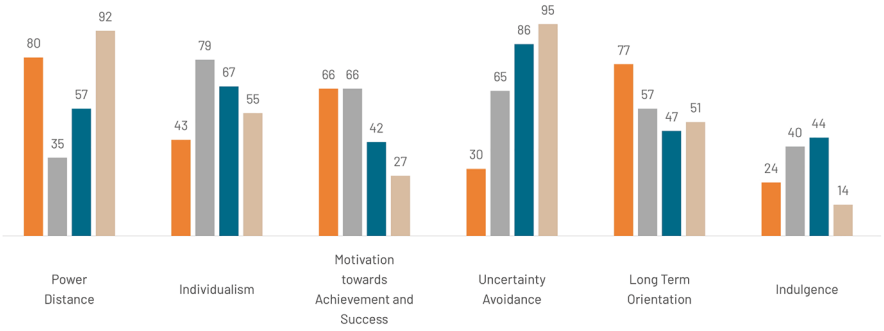

Fig. 13.2 Hofstede comparison of China (orange), Germany (grey), Spain (blue), and Ukraine (light-brown) according to the country comparison tool at https://www.hofstede-insights.com/ (last accessed 11.06.2024)

instance, on the one hand, the pronounced power distance is expected to make the AI adoption in social services predominantly flow from top-down directives, with FOMO potentially being fueled by leadership rather than grassroots demand. On the other hand, the collectivist cultures, which emphasise communal harmony and consensus, might experience a moderated form of FOMO, where the collective decision-making process naturally tempers the rush towards new technologies, thus fostering a more deliberate and inclusive approach to embracing AI in social services."

In summary, for a few illustrative examples it can be shown that it is meaningful to apply a cross-cultural country comparison approach such as Hofstede (2003) and Inglehart and Welzel (2005) to reveal general cultural imprints in behaviour. However, not much more than well-known country stereotypes were roughly confirmed, and sometimes results were ambiguous. For the purposes of this book's research, the lens was far too big anyway to allow deeper insights into the role of culture and context for AI-based social service provision.

Furthermore, limitations such as different funding situations and different access to stakeholders among country cases led to different opportunities for primary research and data collection. Thus, opportunities to answer research questions concerning the role of culture and context for AI-based social assessment were unequally distributed because this unique perspective was hardly available in secondary research materials. It also led to different chances to apply AI FORA participatory methods for inclusive technology co-design, which was in some cases additionally limited due to political and cultural structures less open to participatory approaches, questioning of fairness issues with regard to bias and discrimination, and inclusion of minority perspectives. Generally, this led to a limitation of comparability between case studies.

This is why we chose the case study design localised in a "culture cube" perspective as introduced in Chap. 1 and contextualised within participatory multi-stakeholder innovation networks as explained in Chap. 2. The next section will now

look at commonalities and differences between case studies where findings were enabled by this dedicated AI FORA approach.

Assessing Results of Empirical Research

Following the inclusive technology co-design approach with multiple stakeholders of the Quadruple Helix, in particular vulnerable groups, as described in Chap. 2, the comparative categories chosen for reviewing suggestions for "Better AI" in case studies centre around two central items: insights into the utility of including vulnerable groups in participatory research and challenges for existing or future AI use in social assessment for public service distribution. Of course, this focus needs to state first whether and to which degree vulnerable groups have been included, and whether and to which degree AI is in use. This thematic choice for case study comparison led to Table 13.1. Entries are introduced and discussed in more detail below.

The **chapter on Spain** examined the perceptions, attitudes, and acceptance of AI-based social assessment technologies by policymakers and administrative agencies locally in Catalonia, a frontrunner Spanish region in the adoption of digital technologies for the public sector. Given the overrepresentation of vulnerable groups who use social services in the cities of Barcelona, Girona, and Mataró, special emphasis was given to the impact of AI systems on such groups, particularly migrant populations. The tool investigated was the "Self-Sufficiency Matrix Catalunya (SSM-Cat)" used by municipalities in Catalonia to assign social services.

The chapter shows that the SSM-Cat, and accordingly any AI system that would be based on that, does not produce unambiguous and shared decision-making across municipalities. Complexities of individual cases are difficult to cover. The chapter highlights the necessity of balancing AI's analytical capabilities with human empathy and judgement. Adopting AI within social services is, according to the findings of the chapter, deeply intertwined with the cultural fabric, particularly regarding attitudes towards technology, social values, and, above all, digital literacy. Furthermore, the role of organisational culture, i.e. whether an agency is more or less championing innovation, plays an important role. The chapter points out that the continuous involvement of stakeholders is required, because the conditions for AI-based social assessment permanently change, be it due to the emergence of new technological solutions or due to social, political, economic, or regulatory developments. It is concluded that the inclusion of stakeholders is necessary for periodic reassessment and updates of ethical standards in emerging ethical challenges and responsibilities in AI-based social services. The chapter recommends rigorous oversight to mitigate unforeseen impacts and ensure alignment with societal norms and ethical standards through responsive governance structures.

The **chapter on Estonia** being a frontrunner in digital innovation for Europe and worldwide shows that utilising AI systems to assist human advisors, who act as intermediaries for all social services the state provides, is an application that can be expected soon to be adopted everywhere. For public administration and policy, it

Table 13.1 Country comparison of AI FORA case studies

Feature Country	Insights into utility of including vulnerable groups in participatory research	Challenges for existing or future AI use ("AI in use" refers to ML and neural networks; however, any pattern-recognising data analytics software running on large databases is positively considered here as well) in social assessment for public service distribution
Spain	Vulnerable groups are overrepresented in those who use social services in the cities of Barcelona, Girona, and Mataró: Therefore, their views are important; their input is necessary for periodic reassessment and updates on ethical standards	AI is currently in use for social service provision. Required: Finding solutions that can deal with complex individual cases; balancing AI's analytical capabilities with human empathy and judgement; aiming at unambiguous and shared decision-making across municipalities; making AI subject to periodic reassessment and updates on ethical standards in emerging ethical challenges and responsibilities in AI-based social services
Estonia	So far, research focused on service time—in the future work service quality and fairness need to be checked with clients, especially vulnerable groups; there are many multifaceted and complex cases of vulnerable people— cases that show legislative vagueness, case uniqueness, and sensitivity	AI is currently in use for social service provision. Required: Distinguishing complex from simple cases and customising support; tailoring technical interventions to direct different cases and clients towards appropriate channels; acknowledging the multifaceted landscape of social services requiring an alignment with the diverse needs of stakeholders
Germany	Working with refugees helped to identify and understand "blind spots" in assessment practices and which criteria as personal attributes of the applicants were perceived as relevant; identified high diversity in assessment decisions for applicants sharing criteria due to differences in interpreting and weighing criteria	AI is partially in use for services around granting asylum and follow-up services. Required: AI for assisting processes along the whole social service sequence for refugees along a "fairness value chain"; incorporating complexity and volatility of assessment process and criteria; finding a solution to the issue of criteria interpretation and weighing, because assessment is far more than what "is in the law"
Iran	Social acceptance of new technologies such as AI in social service provision is an issue; there are acceptance challenges from experiences with vulnerable groups in comparative cases; participatory approaches integrating heterogeneous stakeholders might raise acceptance and adoption levels in the population	AI is currently in use for social service provision. Required: Culture-sensitive and dynamic AI could increase social acceptance and satisfaction with services concerning the targeted subsidies plan and other state applications

(continued)

Table 13.1 (continued)

Feature Country	Insights into utility of including vulnerable groups in participatory research	Challenges for existing or future AI use ("AI in use" refers to ML and neural networks; however, any pattern-recognising data analytics software running on large databases is positively considered here as well) in social assessment for public service distribution
India	Vulnerable groups suffer from "cultural resistance" of Indian traditional values that partly counteract and outmanoeuvre the state distribution system and AI technologies supporting it; marginalised communities often face discrimination and exclusion when accessing public services; traditional social hierarchies and caste dynamics influence access and treatment; diverse cultural demands on characteristics of services; systemic/policy failure concerning vulnerable groups; corruption hits vulnerable groups particularly hard	AI is partially in use for social service provision. Required: AI needs to be able to address systemic inequalities based on cultural norms and power dynamics embedded within local contexts; AI needs to be sensitive to local realities and inclusive of diverse voices; dynamic AI is particularly important due to the context of the PDS, where factors such as market fluctuations, climate variability, and policy changes constantly impact service delivery; responsive AI can facilitate timely and informed decision-making; there needs to be (policy) awareness that AI is a tool that makes policy in the current struggle between constitutional values versus traditional values
Nigeria	Pervasive poverty and a high level of corruption in social protection lead to multiple vulnerabilities; current bias and discrimination in geographic locations and ethnicity; inclusion of all stakeholders necessary for social protection delivery	AI is partially in use for social service provision. Required: AI development that provides opportunities for addressing system and policy failures. However, it needs to be ensured that current discriminations and injustices are not simply re-instantiated. Combining AI research with social science is required
Ukraine	Highly vulnerable environment due to the war in the country; AI use for social services through the DIIA App; research centring on war veterans and Ukrainian refugees showed multiple vulnerabilities and complex cases	AI is currently in use for social service provision. Required: AI technologies offering a pathway to access public services in times of war; new technologies offer a way to stay in contact with the public administration of the home country for those seeking refuge in other countries
China	Elderly citizens in need of social care can be considered as vulnerable group; social care issues are deeply rooted in Chinese cultural norms, particularly the enduring values of filial piety and respect for the elderly; research shows that elderly with more social capital receive more care than others	AI is currently in use for social service provision. Social acceptance and satisfaction of new AI technologies in social care is low among the elderly. Required: Culture-sensitive AI to increase social acceptance and satisfaction with services concerning social care. Participatory AI might be useful to raise satisfaction with services and acceptance of AI use among the elderly

(continued)

Table 13.1 (continued)

Feature Country	Insights into utility of including vulnerable groups in participatory research	Challenges for existing or future AI use ("AI in use" refers to ML and neural networks; however, any pattern-recognising data analytics software running on large databases is positively considered here as well) in social assessment for public service distribution
USA	Significant, evidence-backed concern about bias in AI algorithms (especially race, ethnicity, gender); Peoria is one of the poorest zip codes in the country and the most distressed in the state of Illinois according to the Distressed Communities Index; demographics dynamics include a growing minority population that leads to language gaps, trust gaps, and a gap between who the system is designed by and whom it intends to serve; social and systemic bias exists in the current mental health system against minorities and marginalised groups	AI is currently in use for social service provision. AI is at risk of re-enforcing and exacerbating the consequences of bias already present in the current mental health system; errors or biases introduced in one part of the system might propagate, exacerbate, or be contained across the system. Required: Including patients, service providers, and caregivers in the design of AI as a tool for mental health; including those who will be most impacted, there may be an opportunity to address some of the biases before they are built into these systems and mitigate harm done by the already existing systemic issues in mental health; AI is perceived as a blackbox by stakeholders and needs more transparency; AI requires greater oversight

would be advantageous to have a tool for testing and prototyping these applications. The Estonian chapter focuses on services around job finding for the unemployed.

Results focused on service time, i.e. allocation of social workers' time to individual cases, which brought attention to vulnerable groups, because cases featuring legislative vagueness, case uniqueness, and sensitivity need more time to advise on than on simpler standard cases. Vulnerability can be described here as situations where several interrelated factors hinder job finding.

The effectiveness dimension—specifically service quality and fairness—fell beyond the scope of this specific case study due to encountered limitations. Definitive conclusions regarding the consistent impact of AI assistance on specialists' perspectives towards clients and their advisory remained elusive. Research relied on interviews rather than on-site observations, limiting the analysis of real-life behaviours. Standardised AI applications are sufficient for simpler cases increasing operational efficiency. However, for effectiveness to reduce unemployment they need to address the complex cases successfully. Furthermore, standard cases need to be distinguished from complex cases in the first instance, which already could be done by AI.

The **chapter on Germany** found that including vulnerable groups in research (in this case refugees seeking asylum in Germany and their supporters) helped to identify and understand "blind spots" in assessment practices. Importantly, this means that there were yet unknown "blind spots" in the first instance. Thus, it is useful and relevant to include vulnerable groups in research: Information is made available that cannot be gained by a different approach centred on experts. Including vulnerable

groups also helped in identifying and understanding which criteria at the level of personal attributes of people applying for asylum were perceived as relevant in the procedure and beyond. They were by far more, and they were different from the official administrative guidelines. A limitation of our research is here that these were the criteria *perceived* as relevant. They were not checked against the views of the decision-makers.[3] However, the applicants *experienced* to be judged against them (religious background, family size in home country as part of personal conditions, personal charisma, etc.), and the supporters *confirmed* most of these perceptions. Research could also confirm existing fairness-related issues in the whole process, such as different decision-making regarding protection status among federal states in Germany, and barriers to work according to personal attributes which became visible later in the integration process. Furthermore, there is a high volatility in assessment criteria. Research pointed to the political nature of an assessment in this particular context due to changing circumstances in countries of origin as well as host societies and the quick changes in geopolitical dynamics.

The idea often heard from an administrative perspective that "it is all in the law, and it does not matter what the individual case is about, the decision is clear" can be discarded according to research presented. The process is much more complex, and there is much leeway for biased decision-making on the way, especially where human decision-making is involved. Criteria are differently interpreted and weighed—not only by different social workers at agencies or across federal states, but also along the whole "fairness value chain" of the process.

It highly depends on the availability and quality of supporters who are familiar with all difficulties and intricacies to navigate the refugees through the system. Using AI for these processes, existing or projected/suggested, is not present in public discourse in Germany, and it is not very visible, or does not seem to play a role, at grassroots level. For AI systems assisting in the process, the task would be to address this complexity and volatility for fairness along the whole social service sequence for refugees, from arrival at German borders and applying for asylum to full integration in German society, applying for housing, education, jobs, kindergarten places, etc. So far, no comprehensive AI systems are in place to assist in longer or all sequences of the process though such systems have been suggested (Cserpes et al., 2019). So far, mostly identity management systems such as the language recognition system DIAS predicting country of origin are in use, which have a quite narrow concept of fairness. As stated in the case study chapter: "The tension between individual case deliberation and categorisation of countries/regions is not solved by the DIAS". Assisting AI systems would need to incorporate complexity and volatility of assessment process and criteria and find solutions to the issue of criteria interpretation and weighing, because assessment is far more than what "is in the law".

[3] The analysis of practices, norms, and values of decision-makers was based on document analysis only. Furthermore, a media analysis would have shed light on important value discourses, especially on the "welcome culture", solidarity discourses, and the crucial role of civil society, more particularly of civil society organisations.

The **chapter on Iran** showed that the integration of AI into the targeted subsidies plan had indeed been a transformative step in data-driven policy and decision-making in Iran streamlining the identification of deserving recipients and resulting in substantial financial savings. However, the chapter also states that identifying individuals with AI algorithms, as demonstrated in the "Woman, Life, Freedom" movement, poses significant challenges, raises questions about such AI applications in social services, and might impact on acceptance and trust concerning new technologies.

Though data-driven decision-making is generally seen as valuable, organisational data privacy concerns have hindered regular and up-to-date data provision. This highlights requirements for adequate legal mechanisms concerning access and use of data. The chapter emphasises that for achieving sustainable and effective outcomes, continued efforts are required to address societal acceptance, data access, and comprehensive governance.

The **chapter on India** showed that that the position of vulnerable groups is particularly precarious in the Indian Public Distribution System, the PDS. They mostly suffer from "cultural resistance" of Indian traditional values that partly counteract and outmanoeuvre the state distribution system. Especially, marginalised communities often face discrimination and exclusion when accessing public services. Traditional social hierarchies and caste dynamics deeply ingrained in Indian society influence access to and treatment within the PDS. Culture plays a pivotal role in influencing consumption patterns, social norms, and community dynamics, all of which directly impact the functioning of the PDS. The system operates within a rich tapestry of cultural and contextual intricacies, profoundly shaping its effectiveness and impact. Since the PDS is a system providing food and goods of daily demand, very specific cultural intricacies play a role in the Indian system that do not occur in the other country case studies. There is a high diversity of cultural demands on specific characteristics of services provided. Culture profoundly influences food preferences, dietary habits, and traditional practices across India's diverse communities. Regional cuisines, religious dietary restrictions, and cultural celebrations contribute to unique consumption patterns that must be considered in PDS implementation (e.g. in regions where rice holds cultural significance, the availability and quality of rice through the PDS greatly affects community satisfaction and well-being). Addressing these systemic inequalities requires a nuanced understanding of cultural norms and power dynamics embedded within local contexts. However, there is also systemic and policy failure concerning vulnerable groups (e.g. localised services but supra-local job market, disadvantages for the homeless, disadvantages for unmarried women, etc.); corruption in the distribution systems hits vulnerable groups particularly hard. Indian government tries to mitigate these disadvantages; however, some of them are still in place. Therefore, the chapter recommends enhancing gender sensitivity within the PDS, optimising ration allocations for smaller family units, implementing tailored interventions for vulnerable communities, and developing strategies to improve savings, particularly for low-income households. With regard to policy, the need for continuous improvement of the PDS

to effectively address the diverse needs of its beneficiaries and enhance its role as a critical safety net in India's social welfare landscape is obvious.

Concerning AI use, certainly biometrical data analytics is run on selecting beneficiaries; it is unclear whether and to which degree ML is already used. Case study researchers suggest developing a community-based vulnerability prediction model tailored to the Indian context, which would incorporate principles of fairness, accountability, and transparency to ensure equitable outcomes. The model would explore the pivotal role of deep learning in the development of community-specific vulnerability prediction models in the Indian scenario, which would be an innovation in approach to community-based interventions. By harnessing AI technologies sensitive to local realities and inclusive of diverse voices, India can enhance the effectiveness, efficiency, and equity of its public service delivery systems, ultimately advancing social welfare and inclusive development. Given the complexity of cultural and contextual factors shaping the PDS, the adoption of contextualised, participatory, responsive, and dynamic AI for social assessment becomes crucial. In the dynamic context of the PDS, where also factors such as market fluctuations, climate variability, and policy changes constantly impact service delivery, responsive AI can facilitate timely and informed decision-making.

The **chapter on Nigeria** stated that AI can be an enabler of effectively targeting beneficiaries of social protection (SP). Nigeria is a context where poverty is pervasive, and there is an opportunity to use AI for making meaningful progress in SP delivery which so far faced system failure in many parts of the service delivery chain. For example, there is massive current bias and discrimination in geographic locations and ethnicity to be overcome. The chapter sees a chance for technology to help with this and points to the necessity of policy discourse to ensure inclusion of all stakeholders in developing the delivery chain for SP intervention (design, monitoring, and evaluation).

The **chapter on Ukraine** presented a highly vulnerable environment due to the war in the country. Research of the chapter centred on war veterans showing multiple vulnerabilities and complex cases. Additionally, the coordinating team of AI FORA had organised a Safe Spaces workshop on "The role of AI for global social goods provision in times of crisis—How the world deals with the war in Ukraine"[4] with Ukrainian refugees having found refuge in Clifden, Ireland.

Starting point for this workshop in 2022, i.e. during the war time, was the observation that the Ukrainian welfare system was under severe pressure but was still working. People could apply for social services through the App called "Diia", which many Ukrainians had on their mobile phones, and which connected more than 80 national authorities. With the help of Diia, the Ukrainian government provided financial assistance to citizens, both in- and outside of the country, but details

[4] Since the Ukrainian case study was not funded as part of the six original case studies but was added later, funding for the workshop had to be found elsewhere. We gratefully acknowledge workshop funding of the Interdisciplinary Public Policy unit (IPP) of Johannes Gutenberg University Mainz, Germany (https://ipp-mainz.uni-mainz.de/files/2022/10/Flyer_Ukraine_Workshop_2022.pdf; last accessed 17.06.2024).

sensitive to location and other attributes of the mobile phone holder (e.g. some services did not work for people outside the Ukraine where host countries have taken responsibilities for refugees).

Furthermore, the war in the Ukraine had elicited an unprecedented shoulder-to-shoulder cooperation of Western national welfare systems: Many types of social goods provision—be it food, clothing, medical help, military devices, refugees and migrants' admission policies, or donations for monetary support of vulnerable groups confronted with barriers—had been provided. Data were freely exchanged to bring goods to the needy. Huge logistic efforts that had to be made were supported by new techniques of digitalisation such as data analytics and AI hoping for efficiency and objectivity gains in allocation. The App, the still existing welfare system in the Ukraine, and the relationship of that system to the supporter countries were the general topics of the Safe Spaces workshop. Its agenda included experimental and participatory sessions where refugees could explain whether and how they used the App, how they liked it, and what their suggestions were for improvement. Ukrainian policymakers and App developers from the DIIA provider company were participating online in the workshop. Results showed that AI technologies offer a pathway to access public services in times of war. These new technologies can offer a way to stay in contact with the public administration of the home country for those seeking refuge in other countries.

The **chapter on China** confirmed that adding AI systems to the community care of the elderly improved safety and health monitoring as well as quality of service. The chapter found that elderly with more social capital received more care than others: Those elderly people proactively making their needs known to nursing centres received more care and services than other seniors; furthermore, those who had more relatives in the community or knew the nursing home staff received more care than those who did not. This demonstrates, as the chapter stated, that social care issues are deeply rooted in Chinese cultural norms, particularly the enduring values of filial piety and respect for the elderly. This might also be an explanation for the fact found by the chapter's research that social acceptance and satisfaction of new AI technologies in this area is low: Face-to-face contact is seen as important.

A participatory approach might be useful to raise satisfaction with services and acceptance of AI use among the elderly, who can be considered as vulnerable group here. Satisfaction and acceptance proved to be very important during the COVID-19 pandemic when dissatisfaction with lockdown measures based on technological social assessment measures was significant among the overall Chinese population (Ren, 2020; Gan et al., 2022). In order to study how data analysis and AI influence China's public services and may best be used, it is necessary to consider China's special socio-cultural background and its specific approach to modern technology. To stay at the cutting edge of international research, policy instruments such as China's national AI development plan or the AI Governance Forum of the World Congress on Artificial Intelligence demonstrate awareness of such ethical issues (Ahrweiler & Neumann, 2021; Ahrweiler, 2020) for enhanced understanding and creating better AI usage on social service provision in the national welfare system.

As China was particularly hard hit by the COVID-19 pandemic with long lockdowns particularly in Shanghai, participatory research about the role of culture and values for AI-based social assessment in public service provision was restricted. Therefore, a Safe Spaces workshop with 12 Chinese students born and bred in China was organised by the AI FORA coordinator in Ireland to explore the relationship between Chinese culture, social welfare, and AI. Ireland hosts a large-sized Chinese community compared to other European countries, both in terms of university students who chose Ireland as their place of study and in terms of professionals working in the country.

The workshop used a gamification approach based on Hofstede's cultural dimensions theory (see the section below for comparative results): Stakeholders participated in a game on AI-based assessment for unemployment benefits designed to study how participants would create better systems from their cultural point of view, by letting them play a situation that informs this improvement of AI systems. Each player was given an identity, each with mixed backgrounds in personal attributes such as age, gender, education, employment status, area of residence, and household composition that they had to play the game with. They could, however, change their wealth and happiness levels by taking up/changing employment, training, or making policy changes to the (unknown to the players by default) algorithms that dictated their abilities to gain better employment, training, and/or happiness. At the end of each round, the participants went "home", i.e. a Safe Space where they discussed their experiences, the events of their "day", and how they felt. Each participant recorded events from the day in their diary each "evening". Proposals to alter the system could be submitted, e.g. to have the algorithm disclosed or to change the algorithm. They could be enacted when voluntary donations were made (e.g. donations of 50 credits could buy a change in the algorithm for one round only, donations of 100 credits could buy a change in the algorithm for all time) and when proposals were selected in a secret vote by the majority after discussion. After two rounds, test data was collected and presented to the participants for further review.

Results showed high sensitivity of players for bias, discrimination, and inequalities in the system, especially concerning gender, ethnicity, and area of residence. Remedies were sought and implemented collectively. The value assigned to "education and training" was very high—even as a value per se without immediate returns through the incentive structure of the game.[5] Discussions showed that bias would be most accepted here—participants claimed more internal differentiation within the education domain to distinguish between beneficiaries and non-beneficiaries.

The **chapter on the USA** reported findings from the case study on systems of mental care in Peoria, Illinois, which was carried out by the University of Arizona. Research was interrupted because of turnover of key personnel before the participatory mapping and modelling phases of the research, but key insights could be gained.

[5] This might be due to the fact that all players were Chinese students at Irish universities with high personal investments in this dimension.

The chapter reports that Peoria is one of the poorest zip codes in the country and the most distressed in the state of Illinois according to the Distressed Communities Index. It shows that demographics dynamics include a growing minority population that leads to language gaps, trust gaps, and a gap between who AI systems in mental health care are designed by and whom they intend to serve.

The chapter identifies social and systemic bias existing in the current mental health system against minorities and marginalised groups. AI is at risk of reinforcing and exacerbating the consequences of bias already present in the current mental health system. Errors or biases introduced in one part of the system might propagate, exacerbate, or be contained across the system. AI requires greater oversight. The chapter recommends that ongoing reflections of who designs, uses, and benefits from AI-based social systems are necessary to promote improvements in performance and accountabilities. Research suggests including patients, service providers, and caregivers in the design of AI as a tool for mental health. Including those who will be most impacted, there may be an opportunity to address some of the biases before they are built into these systems and mitigate harm done by the already existing systemic issues in mental health. Otherwise, the chapter explains, AI is perceived as a blackbox by stakeholders in lack of transparency.

How the public perceives AI-based social assessment is the dedicated subject of another US sub-study within the AI FORA project. Treyger et al. (2023) investigated in their study "Assessing and Suing an Algorithm: Perceptions of Algorithmic Decisionmaking" how public perceptions of algorithmic decision-making in employment and unemployment compare with perceptions of traditional human decision-making and how encounters with negative outcomes produced by algorithmic decision-making will shape people's assessments of these technologies.[6]

Treyger et al. (2023) used a survey experiment providing respondents with two scenarios in employment/unemployment where AI systems were in use. The experiment focused on perceptions of fairness, the potential shortcomings of algorithms, and people's willingness to resort to the law to challenge algorithmic decisions. Furthermore, it investigated how encountering negative algorithmic decisions shapes views of AI decision-making more broadly. "Because minority groups and women are at times disproportionately adversely affected when algorithms do not perform as designed, we further considered whether these more-affected groups assess AI decisionmaking differently" (Treyger et al., 2023: v).

The questionnaire did not only ask for standard demographic variables such as gender, age, education, household income, and employment status but also for ethnicity/race ("Non-Hispanic White" and "Other" as categorial options), political orientation, region of residence, confidence in the US courts and the legal system, belief that merit and hard work are rewarded, and computer literacy. It also asked for

[6] The two US case study teams in AI FORA (University of Arizona and RAND Corporation) were particularly impacted by consequences of the COVID-19 pandemic and key personnel turnover up to the point that the originally planned RAND case study did not materialise. However, the findings of this case study's preparatory phase led to a publication, which is introduced here outside of chapters, because it yielded interesting insights on public perception of AI-based social assessment.

the personal civil suit history (being sued and having sued), but, unfortunately, it did not ask for a history of having felt discriminated by an algorithm or human being in applying for state benefits.

Most respondents felt that "some decisions are too important for AI" and that greater exposure to algorithmic decision-making, especially if the decision-making is perceived to be unfair, corresponds to greater scepticism about the future and possibilities of algorithmic processes (cf. Treyger et al., 2023: vi).

Results from the survey were interesting with regard to AI FORA's focus on vulnerable groups, because there is significant, evidence-backed concern about bias in AI algorithms, especially race, ethnicity, and gender, in the USA (cf. RAND reports "An Intelligence in our Image" and "Algorithmic Equity"). "The concern that algorithmic tools will disproportionately affect disadvantaged groups—notably, non-White and women—is well founded. At the same time, human decision-making processes have also historically dis- advantaged the same groups. Whether injecting algorithmic technology into high-stakes decisions is a net benefit or a net detriment to these populations is an open question—and one that likely has different answers across different decisionmaking contexts." (Treyger et al., 2023: 28). However, results of the study revealed that there were no big differences in respondents' likelihood of pursuing legal remedies for perceived unfair or incorrect decisions made by algorithms compared with those made by humans.

Assessing the Safe Spaces

Most case studies worked with Benedictine monasteries as "Safe Spaces" (cf. Chap. 2) to enable eye-level interaction, agency, and empowerment of heterogeneous stakeholders including vulnerable groups. How did it work? And how was it perceived by participants of research? There was extensive Safe Spaces use in Germany (Monastery of Nuetschau), Spain (Montserrat Abbey), and India (Saccidananda Ashram Shantivanam). Nigeria and Iran as non-funded case studies could not work with the Safe Spaces concept due to financial reasons. The Chinese and the Ukrainian case study worked with a Safe Space outside their countries (Kylemore Abbey in Ireland), and the USA case study could not realise their planned Safe Space activities at New Camaldoli Hermitage due to issues explained above.

Assessment from the Point of View of Safe Space Leaders

What was the assessment of the Safe Spaces leaders, the superiors of the participating Benedictine monasteries, and their communities? For them, two aspects seemed particularly relevant to the Safe Spaces concept. First, the cloistered space of a monastery marks an inside-outside boundary in which those inside are "safe" from the demands of the outside world. In this sense, "enclosure" signifies exclusion

(independence and autonomy, especially from ecclesiastical hierarchies). On the other hand, the "small world" of a monastic Benedictine community living together in everyday life is a role and pilot model or training ground for the "big world" outside (cf. Tredget, 2002).

In Table 13.2, the subjective perceptions of the Safe Spaces leaders are summarised in the form of spontaneous expressions as a pro-contra table of quotes

Table 13.2 Self-assessment of the Benedictine monasteries as Safe Spaces

Pro	Contra
It is a positive challenge to engage with people and accept them as they are. A monastery does not mean that it has walls	Monasteries are too specific for non-Catholics, e.g. secular young people; they are exclusive
We can act as a "greenhouse" by moving and empowering people to change their thinking by communicating our way of life	Is this really in the interest of the monks/nuns? Isn't it just a big distraction from the inner core of our lives? Shouldn't we rather retreat?
It is a two-way service: The Safe Space experience is also meaningful to the monastic community and will have a positive impact on the community itself	Fear of openness. Communities may not want to engage in secular issues. "It's not Catholic". Identity gets in the way
It raises awareness in the community of the issues of their fellow people. It is a good way to combine contemplation and service	There is not enough information, and there is not enough acceptance—either for the community or for the guests
It's a "reality meets island" experience where the fear of the unknown stops. The AI topic is relevant for us. (The interest in the future of AI was considerable. Discussions centred around a message of Pope Francis sent to participants of a Vatican workshop titled "The 'Good' Algorithm? Artificial Intelligence: Ethics, Law, Health" in February 2020. Here, the Pope had said that in the newly emerging discipline of "the ethical development of algorithms", a critical contribution can be made by the principles of the Church's social teaching, namely the dignity of the person, justice, subsidiarity, and solidarity. However, any challenges, the Pope had continued, must not detract from the immense potential that new technologies offer (cf. Rome Call for AI Ethics, 2020; https://www.romecall.org/wp-content/uploads/2024/02/RomeCall_report-web.pdf, last accessed 13.06.2024))	It's risky. We are dealing with difficult issues that can put communities at risk

(continued)

Table 13.2 (continued)

Pro	Contra
This is part of the natural evolution of the Order, because God is in creation and therefore in society	It takes a lot of preparation to allow communities to participate
It is our duty to do this because it is part of our spiritual charism. [Here, discussion referred to a quote in lecture at the 2021 Benedictine Religious Conference titled "A time of empty churches—a prophetic warning?" by philosopher of religion Tomas Halik: "Can the church today offer enough places where the whole truth can be told? I dream of a church that is a **safe space**—a space of truth that heals and sets free. In this area I have great hopes for religious communities" (Tomas Halík, lecture on 23.07.2021: "A time of empty churches—a prophetic warning?", Benedictine Religious Conference 2021, p. 6)]	It is hubris to assume that we are the only ones who can do something here

worth considering on the question "Benedictine monasteries as Safe Spaces?" The contributions, which were collected on the kickoff workshop of Safe Spaces leaders, also reflect the different attitudes towards life and future issues that are currently affecting monastic communities.

The range of opinion was broad between "duty" and "hubris"; it anticipated community responses to the Safe Spaces challenge. However, one point stood out in particular: In the very realm of everyday life, there were striking similarities between what would take place in Safe Spaces and in the daily decision-making of a monastic community. It was also these commonalities that provided the background and experience to address the needs of Safe Spaces. For example, there are similar tensions between inclusion and exclusion, between participation and authority, and between the need for open deliberation and structured regulation. These tensions must be negotiated daily in order to follow the best argument in a complex situation and find the best solution to a problem. The Safe Spaces leaders defined monastic communities not as static structures, but as in motion: According to them, a monastic community is a fragile project, just like the group of participants at a Safe Space workshop. Here as there, a heterogeneous group of people must be brought into cooperation for creating a common product.

Assessment from the Point of View of Participants in Safe Space Workshops

For feedback on the venue, a questionnaire was run among workshop participants for most workshops.

Questions were:

- Did you like the venue? (yes/no/do not know)
- Did you know about the venue before? (yes/no)
- Did you visit the venue before?

- The venue is in a remote place. What was the advantage of not having the workshop at university? (open question; keywords)
- The venue is in a remote place. What was the disadvantage of not having the workshop at university? (open question; keywords)
- The place is run by a Benedictine community. Was it a "too religious" environment for you? (yes/no/do not know; keywords could be provided for reason)
- Did you experience the venue as a "Safe Space"? For example, could you speak openly about topics/issues that are relevant to you? (yes/no/do not know: keywords could be provided for reason)
- How do you rate the staff and infrastructure of the venue? (high/okay/low)
- The workshop was about a topic with high ethical and social implication for the future of our societies. Are there features of the venue that supported value discussions about such crucial issues? If so, please mention these features (open question; keywords).
- How do you rate the overall workshop experience? (high/okay/low; keywords)

Since the questionnaire was not run at all workshops due to language issues and results could not claim to be representative, only a summary of majority answers is provided here. Most participants liked the venue and had heard about it even if they had never visited it before. Participants saw locations as part of the shared local cultural heritage which fostered trust and common sense of place.

Some Safe Spaces had a background in distribution practices of social service provision themselves (Monastery of Nuetschau, Saccidananda Ashram Shantivanam) and were known to vulnerable participants as helping the "poor and needy", which further increased trust. The remote non-university location was appreciated by most (keywords: silence, serenity, tranquillity, peacefulness, nature, atmosphere, historical scene, time, etc.); some mentioned as disadvantages long travel time, missing technological infrastructure, and limited leisure facilities. Interestingly, none of participants said the venue was "too religious" though this had been one of the biggest concerns of the workshop organisers beforehand.

The venue was considered as "Safe Space" by nearly all the participants; most people mentioned hospitality as a key feature. Especially workshop participants from particularly vulnerable backgrounds (Ukrainian refugees; Indian casteless women from extreme poverty backgrounds) appreciated the Safe Space component and felt empowered by the environment.

Assessment from the Point of View of Research

Across the case study countries, nine dedicated Safe Spaces workshops at four different venues were held with the objective to work with vulnerable groups for AI innovation.

How research in case studies benefitted from integration of vulnerable groups enabled by the Safe Spaces was already reported and assessed in section "Assessing

Results of Empirical Research" (summarised in Table 13.1). Table 13.3 summarises the Safe Space activities of case studies. For the social researchers in case studies, it was easy to work with the Safe Spaces due to the Benedictines' familiarity with academia, sometimes even supported by formal institutional cooperations (e.g. Kylemore Abbey with Notre Dame University in the USA).

All participating monasteries were running guest houses and retreat centres with accommodation facilities and educational infrastructures such as seminar rooms for

Table 13.3 Safe Spaces in case study countries

Safe Space country	Available/used	Experience
Spain	Yes	Two workshops held at Montserrat Abbey; using the networks of the Abbey for recruiting workshop participants from NGO and for dissemination; including workshop participants from vulnerable groups such as migrant workers; using the hospitality functions of the Abbey such as retreat facilities, the guest house, and the seminar infrastructures
Estonia	No	
Germany	Yes	Three workshops held at the Monastery of Nuetschau; personnel of the monastery trained in working with heterogeneous groups helped to conduct the workshop; using the networks of the monastery for recruiting workshop participants; using knowledge and input of the monastery for researching the domain (monastery as church asylum provider); including vulnerable groups such as refugees (also refugees with unclear status, e.g. in church asylum) and migrants; using the hospitality functions of the Abbey such as retreat facilities, the guest house, and the seminar infrastructures.=
Iran	No	
India	Yes	Two workshops held at Saccidananda Ashram Shantivanam; using the networks of the monastery for recruiting workshop participants; including workshop participants from vulnerable groups such as women of low caste membership or casteless and participants from very poor rural areas; using knowledge and input of the ashram for researching the domain (ashram as food provider); using the hospitality functions of the ashram such as retreat facilities, the guest house, and the seminar infrastructures
Nigeria	No	
Ukraine	Yes	One workshop held at Kylemore Abbey; personnel of the abbey trained in working with heterogeneous groups helped to conduct the workshop; using the networks of the abbey for recruiting workshop participants; including workshop participants from vulnerable groups such as refugees from Ukrainian war fled to Ireland; using knowledge and input of the abbey for researching the domain (abbey as employer and supporter of refugees); using the hospitality functions of the abbey such as retreat facilities and guest house; using cooperation partners of the abbey for further workshop issues (Notre Dame Global Centre for seminar infrastructures on ground)

Table 13.3 (continued)

Safe Space country	Available/used	Experience
China	Yes	One workshop held at Kylemore Abbey; personnel of the abbey trained in working with heterogeneous groups helped to conduct the workshop; workshop participants included young people studying abroad who can be considered as vulnerable; using the hospitality functions of the abbey such as retreat facilities and guesthouse; using cooperation partners of the abbey for further workshop issues (Notre Dame Global Centre for seminar infrastructures on ground)
USA	Yes/No	Planned was a scenario workshop at New Camaldoli Hermitage on cultural diversity to check with heterogeneous stakeholders including vulnerable groups on biases which had not been covered by the RAND questionnaire

bigger groups, break-out rooms, presentation technologies, and active learning environments; the retreat components offered opportunities for social activities. Staff of monasteries and infrastructure provided approaches appealing to workshop participants on different levels of their existence and experience (intellectual, psychological, emotional, social, etc.). In some cases, personnel of monasteries trained in working with heterogeneous groups supported the research staff in conducting workshops. Networks of the monasteries helped to recruit workshop participants, because monasteries were in contact with stakeholders and applicants for social services directly, as well as with other organisations such as NGOs that helped to access or to back up for missing social services. As such, monasteries also proved to be valuable information sources regarding insider knowledge about the social service domain under research. The remote places of most monasteries, seen as benefit by participants, sometimes led to logistical challenges in workshop organisation, particularly transport of participants.

Assessing Impact

Including the full variety of multi-stakeholder perspectives is crucial for the legitimacy of AI design and implementation. The approach presented here tries to provide a quality space for participation and negotiation, where the diverse voices of all stakeholding communities can impact the shape of future AI systems in social welfare.

Birhane et al. (2022), however, had warned against "participation washing" (see discussion Chap. 2, section "The Need to Include Vulnerable Groups"), which means avoiding the participation of stakeholders, especially vulnerable groups, but has no consequences for decision-making. Stakeholders need to see that their agency is increased by participation and that their input makes a difference. How

can knowledge produced by inter- and transdisciplinary participation of heterogeneous stakeholders reach the level of decision-makers? The assessment exercise now turns to policy as one of the target audiences (Waibel et al. 2021) of research presented in this volume. Governmental decisions on AI use in public administrations of national welfare systems provide one important access point for "Better AI". Thus, the policy community in both domains, AI policy and public policy, is the first target audience of research. Results achieved by the participation of heterogeneous stakeholders including vulnerable groups need to be made known to this audience and lead to policy learning and impact for change. AI FORA research has shown that this can be mainly done in two ways.

Policy Learning by Information

The case study of Catalonia, **Spain**, worked with public administrators as the main client group and concluded that the inclusion of multiple stakeholders would be necessary for periodic policy *evaluation* and updates of ethical standards in emerging ethical challenges and responsibilities in AI-based social services; research recommends rigorous oversight to mitigate unforeseen impacts and ensure alignment with societal norms and ethical standards through responsive governance structures. In **Estonia**, outcomes have already been shared with the management of the Estonian Unemployment Insurance Fund (EUIF) as the main client group of research, enabling EUIF to enhance its *implementation* processes. In **Germany**, a similar policy workshop has taken place at the side of the European Workshop on Algorithmic Fairness "EWAF 2024" to *evaluate* current AI policies against results of the German case study with regard to the major policy goal to integrate refugees into German society and job market. The need for policy awareness that AI is a tool that *implements* policy is central in **India**: Here, the current competition between constitutional values versus traditional values will be decided. With regard to policy *evaluation*, the need for continuous improvement of the PDS to effectively address the diverse needs of its beneficiaries and enhance its role as a critical safety net in India's social welfare landscape was convened to decision-makers by a 2024 policy workshop in Kerala. In **Nigeria**, a policy workshop is planned with representatives of the Ministry of Humanitarian affairs, Disaster Management and Social Development, the National Social Safety Net Coordinating Office, and others to discuss uptake of study recommendations in SP. In the **USA**, a dedicated workshop "Policy Modelling meets Policy Practice" has taken place in Washington D.C. at the side of the Annual Modeling and Simulation Conference "ANNSIM 2024" to inform policy representatives of the White House and others about the results of the American case study.[7]

[7] https://annsim.org/wp-content/uploads/2024/05/Flyer_policy-workshop_DC_v8d.pdf; last accessed 15.06.2024.

These dissemination activities of case studies present the classical function of science communicating to policy through publications, briefings, and dedicated information events. Governmental decision-making is spread over the whole policy cycle of agenda setting, policy formulation, policy adoption, policy implementation, and policy evaluation leading again to new agenda setting (Knil & Tosun, 2012). As can be derived from above (see italics), mainly the phases of implementation and evaluation have been targeted so far. The phases of agenda setting, policy formulation, and policy adoption will be targeted by the modelling and simulation work, which is presented in the second volume of AI FORA research (cf. last section of this chapter).

Policy Learning by Participation

However, it is important to note that the participatory approach employed by the case studies in this book also features a new mode of policy advice. The conventional scientific policy advice relationship (Weingart & Lentsch, 2009; Wrasai & Swank, 2007; Jasanoff, 2004; Weaver et al., 2001), the long-familiar alliance of decision-makers from political elites receiving so-called evidence-based policy advice by experts from scientific elites, is changing. Our knowledge societies were used to receive their legitimacy from connecting democratic representation to scientific knowledge.

This was not only to guarantee the postulated rationality of political discourse, but also to produce public policies that were based on the best available evidence for their appropriateness and success. However, complexity features of modern societies challenge the continuing applicability of this conventional relationship between disciplinary science and parliamentary politics and ask for interdisciplinary response. Firstly, the complex nature of today's policy problems requires data amounts, qualities, and accessibility of data that are unprecedented and cannot be satisfied by monodisciplinary scientific expert advice alone. Secondly, as a further feature of social complexity in policy domains such as public policy, true uncertainty exists between suggested policy interventions and their desired effects overtaxing prognostic capacities of conventional analytical approaches. Thirdly, massive, complex, pervasive, and intensely interlinked societal challenges such as AI use for social services require buy-in and governance by all societal interest groups contributing expertise from all inter- and transdisciplinary knowledge fields.

Participatory approaches such as the one of inclusive technology co-design featured in this book address the need to include multiple stakeholders in providing policy-relevant expertise. The "evidence" is created together, and the policies to govern the field are co-created as well.

This means that not only "non-governmental actors engage in new forms of governance with emphasis on stakeholder and citizen participation" (European Commission, 2007) for public co-creation, but that decision-makers from governmental organisations are both clients and participants in the research and innovation

process. This participant role of policy is also postulated by policy concepts themselves, as already quoted and in Chap. 2: "Responsible Research and Innovation (RRI) implies that societal actors (researchers, citizens, policy makers, business, third sector organisations, etc.) work together during the whole research and innovation process in order to better align both the process and its outcomes with the values, needs and expectations of society." This would mean policy learning by shared experience, empathy, and participation rather than by a one-sided information flow from science to government. Case studies in AI FORA tried to include public authorities who need to roll out policies for AI-based social assessment, public agencies that use them in their daily routines, representatives of national governments and administrations, NGOs, and civil society organisations as steering actors of the "Quadruple Helix" in their empirical social research.

This worked well for some countries (mainly Spain and Estonia); for others it proved to be difficult due to various reasons. Where it worked well, this was mostly based on previous relationship-building where researchers and policy representatives had already established trust and commitment by experience in working together for mutual benefit. In other cases where new contacts had to be made, the different functional logics of scientific research and policy practice sometimes clashed: It was difficult for researchers to get access to, attention, and time of policymakers and policy analysts who asked for quick wins, high utility of results for their daily problems, and strict efficiency of the interaction (Jager & Edmonds, 2015), especially in a domain as publicly sensitive as AI use in social service provision.

Inclusive Science Communication

People are in very different positions to appreciate scientific achievements. How could the proverbial "people in the street" become aware of AI FORA's research? Research is often only available to experts, shaped by their interests, and framed by their language.

This acts as a threshold, particularly excluding the general public. Hence, science communication requires inclusive formats that are accessible to non-scientific audiences.

One suggestion how to do this was to break out of the silos of academia by switching media. "Angels and other Cows" (Ahrweiler, 2024)—the AI FORA novel in literary fiction blending genres such as sci-fi, romance, adventure, mystery, and comedy—took the task of inclusive science communication making available research topics, results, and consequences of AI use in the public sector to a broad readership for also attracting non-scientists to academic research. The fictional story of the open access novel introduces, reflects, and discusses main epistemological, sociological, and technical concepts of the research area. Hidden in a travel story around fictional cases in **Spain, Estonia, Germany, India, China, and the**

USA, a plot that works like a "Trojan Horse" deals with the deep questions of AI FORA:

- How does social assessment work?
- How is it used in welfare decisions?
- How is AI work used?
- What role do cultural values play?
- What about bias and discrimination?
- Is AI the crown of socio-technical evolution?
- Is there an alternative to AI?
- What are the risks, limitations, and barriers of AI?
- Why is it important to get involved in AI?
- What would be needed to create "better AI"?
- What can I do myself?

The readers are invited for feedback. It is still open whether they want to see "the science behind": The novel, which is also made available as graphic novel for visual experience on social media, precedes the two scientific publications of AI FORA and is published in the same series *Artificial Intelligence, Simulation and Society*. Readers getting in touch for feedback will see results from their engagement: The fourth and concluding book of the AI FORA open access series, which will be another novel evaluating the two scientific volumes, will publish any inputs, discussions, and feedback.

Assessing Plans: What Is Next?

The book has demonstrated that inclusive and participatory designs are essential not only to give voice to citizens in the process of developing and implementing AI applications but also in identifying persistent patterns of social injustice such as masking, redlining, or data-inferred biases. It had presented empirical research on AI-based social assessment in nine case study countries. The focus was on existing configurations, on relevant actors, networks, and resources, on current assessment practices, on their relation to values, culture, and context, on bias and discrimination, on general performance issues and deficits, and on stakeholders' experiences with AI use.

The participatory approach revealed some ideas for more desired systems—for "Better AI". Interactive formats at multi-stakeholder workshops, especially at Safe Spaces workshops, brought forth the culturally shaped and heterogeneous value perspectives of the local social groups.

As a central component, participants played "serious games" for co-designing better AI systems (see above). In summary, the participatory reconstruction and review of existing systems in case studies resulted in a call for more contextualised, value-sensitive, responsive, and dynamic AI systems to be co-designed starting

from existing systems that were perceived as problematic to more desired systems from the point of view of societal stakeholders.

Staying in line with the approach chosen, this needs a participatory anticipation, projection, and realisation of the desired systems. Accordingly, AI FORA has developed a participatory modelling strategy that was designed to support the transition from existing to desired systems.

Case studies see a great chance in bringing data from empirical research to models and simulations for a better future. The case study in Spain will build an ABM to act as a kind of theorem checking device for the assessment algorithm in place in the empirical system under investigation, in this case social assessment in Catalunya. The ABM will not only ensure that the ruleset is coherent and complete, but let it act as an informed starting point for devising a better algorithm by stakeholder involvement. The case study in Estonia will use modelling to shift the focus more explicitly to citizens and their requirements concerning AI systems in the context of (un) employment services; particular attention will be directed towards potentially vulnerable groups to gain a deeper understanding of their interactions with and needs related to AI systems. The case study in Germany will focus with an ABM on the tipping points for agency of refugees (for more and quicker integration into society and the job market) and legitimacy of administrative decisions (for accountability and correctness of bureaucratic procedures in granting asylum) and the trade-offs between these two policy objectives. Iran's case study will construct an agent-based model (ABM) informed by insights gained from the analysis of TSP in Iran to explore future scenarios involving the implementation of various policy options. The case study in India expects that modelling will explore the pivotal role of deep learning in the development of community-specific vulnerability prediction models in the Indian scenario, which will be an innovation in approach to community-based interventions. The case study in Nigeria plans to model a new national SP policy framework with an AI strategy and carry out a micro-simulation targeting an analysis of an AI-enabled social cash transfer programme in Nigeria. Further analysis of serious games and related ABM will be conducted for case study questions of Ukraine and China.

However, research will not stop at just identifying more desirable social assessment routines. For each case study country, a machine learning (ML) system will be developed for prototyping AI social assessment systems. This approach will involve the creation of synthetic populations for case study countries, generating synthetic outcomes based on improved assessment rules, training the neural networks, and evaluating the effectiveness of more desired AI systems. This will be a final important result for policy as the main target audience of research: Unintended consequences and ineffective systems can be avoided, saving on costly development and ensuring that AI for social service delivery is responsive, fair, and beneficial to case study societies. Policy modelling can be used for co-designing AI systems with stakeholders leveraging acceptance and public ownership; it can be used as ex ante evaluation for testing and prototyping AI systems before implementation reducing risks and costs; it can be used for scenario analysis and asking what-if questions

reducing uncertainty, and it might even be used to directly create training data, e.g. using a micro-simulation approach, reducing process complexity.

AI FORA's results will show how AI will change public service provision and its underlying societal value systems in the future, and what will be likely effects and impacts for living conditions and well-being of international societies. Research will find out which policies are necessary and appropriate to prevent or support certain scenarios, and how to create better, i.e. more responsible, AI technology that engages with societal norms and values of stakeholders. Volume II will present results of AI FORA's participatory modelling strategy.

Acknowledgements Research has been funded by the German VolkswagenStiftung under grant agreement number 98 560. Additional funding for open access publication by the same funder and matching open access funding by Johannes Gutenberg University, Mainz, Germany, are gratefully acknowledged.

References

Ahrweiler, P. (2017): Simulationsexperimente realexperimenteller Politik – der Gewinn der Zukunftsdimension im Computerlabor. In: S. Böschen, M. Gross, & W. Krohn (Eds.) *Experimentelle Gesellschaft*. Nomos Verlagsgesellschaft, edition. Sigma (pp. 199–237).

Ahrweiler, P. (2020). Interdisciplinary approach to AI governance research. In: *AI governance in 2019. A year in review*. Shanghai Institute for the Science of Science (ed.). Shanghai, June 2020: 25-27.

Ahrweiler, P. (2024). *Angels and other cows. A celestial adventure into AI worlds, the social good, and unknown connections*. Springer.

Ahrweiler, P., & Neumann, M. (2021). AI governance for the people. In: *AI governance in 2020. A year in review*. Shanghai Institute for the Science of Science (ed.). Shanghai, June 2021: 41-43.

Ahrweiler, P., Gilbert, N., Bicket, M., Sabater Coll, A., Luque Capellas, B., Wurster, D., Siqueiros, J., & Spaeth, E. (2024). Gamification and Simulation for Innovation. In: C. Elsenbroich, & H. Verhagen (Eds.) Advances in Social Simulation. Springer Proceedings in Complexity. Springer, Cham: 121–136.

Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M.C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. In Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), 1–8. https://doi.org/10.1145/3551624.3555290

Cserpes, B., Bindreiter, S., Forster, J., & I. Schuster (2019). *ICT-solutions for MICADO*. Migrant Integration Cockpits and Dashboards. EU Report. Horizon 2020 – DT-MIGRATION-06-2018 | 822717. https://www.micadoproject.eu/wp-content/uploads/sites/18/2020/07/MICADO_1.3_ICT-Solutions-for-MICADO_revised-version_2020-06-30.pdf (last accessed 16.06.2024).

David, M. (Ed.). (2010). *Methods of interpretive sociology*. Sage.

European Commission. Directorate-General for Research and Innovation. (2007). *Taking European knowledge society seriously*. Retrieved December 19, 2023, from https://op.europa.eu/en/publication-detail/-/publication/5d0e77c7-2948-4ef5-aec7-bd18efe3c442/language-en

Gan, Y., Ma, J., Wu, J., Chen, Y., Zhu, H., & Hall, B. J. (2022). Immediate and delayed psychological effects of province-wide lockdown and personal quarantine during the COVID-19 outbreak in China. *Psychological Medicine, 52*(7), 1321–1332. https://doi.org/10.1017/S0033291720003116

Hofstede, G. (2003). *Culture's consequences: Comparing values, behaviours, institutions and organizations across nations*. Sage.
Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.
Jager, W., & Edmonds, B. (2015). Policy making and modelling in a complex world. In M. Janssen, M. Wimmer, & A. Deljoo (Eds.), *Policy practice and digital science* (pp. 57–74). Springer.
Jasanoff, S. (Ed.). (2004). *States of knowledge: The co-production of science and social order*. Routledge.
Kitcher, P. (1993). *The advancement of science. Science without legend, objectivity without illusions*. Oxford University Press.
Knill, C., & Tosun, J. (2012). *Public policy: A new introduction*. Palgrave Macmillan.
Knorr-Cetina, K. (1981). *The manufacture of knowledge. An essay on the constructivist and contextual nature of science*. Berkeley.
Ren, X. (2020). Pandemic and lockdown: A territorial approach to COVID-19 in China, Italy and the United States. *Eurasian Geography and Economics, 61*(4–5), 423–434. https://doi.org/10.1080/15387216.2020.1762103
Strahringer, S., & Leyh, C. (Eds.). (2017). *Gamification and serious games: Grundlagen, Vorgehen und Anwendungen*. Wiesbaden/Heidelberg.
Szczepanska, T., Antosz, P., Berndt, J. O., Borit, M., Chattoe-Brown, E., Mehryar, S., Meyer, R., Onggo, S., & Verhagen, H. (2022). GAM on! Six ways to explore social complexity by combining games and agent-based models. *International Journal of Social Research Methodology, 25*, 541–555.
Tredget, D. (2002). The "benedict rule" and its significance for the world of work. *Journal of Managerial Psychology, 17*(3), 219–229.
Treyger, E., Taylor, J., Kim, D., & Holliday, M. A. (2023). *Assessing and suing an algorithm: Perceptions of algorithmic decisionmaking*. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2100-1.html
Waibel, D., Peetz, T., & Meier, F. (2021). Valuation constellations. *Valuation Studies, 8*(1), 33–66.
Weaver, K., Stares, P., & Kokusai Koryu Senta, N. (Eds.). (2001). *Guidance for Governance: Comparing alternative sources of public policy advice*. Tokyo.
Weingart, P., & Lentsch, J. (Eds.). (2009). *Scientific Advice to policy-making: International comparison, Opladen & Farmington hills*. Barbara Budrich.
Wrasai, P., & Swank, O. H. (2007). Policy "makers, advisers, and reputation". *Journal of Economic Behavior and Organization, 62*(4), 579–590.