Proceedings of the Weizenbaum Conference 2023: AI, Big Data, Social Media, and People on the Move

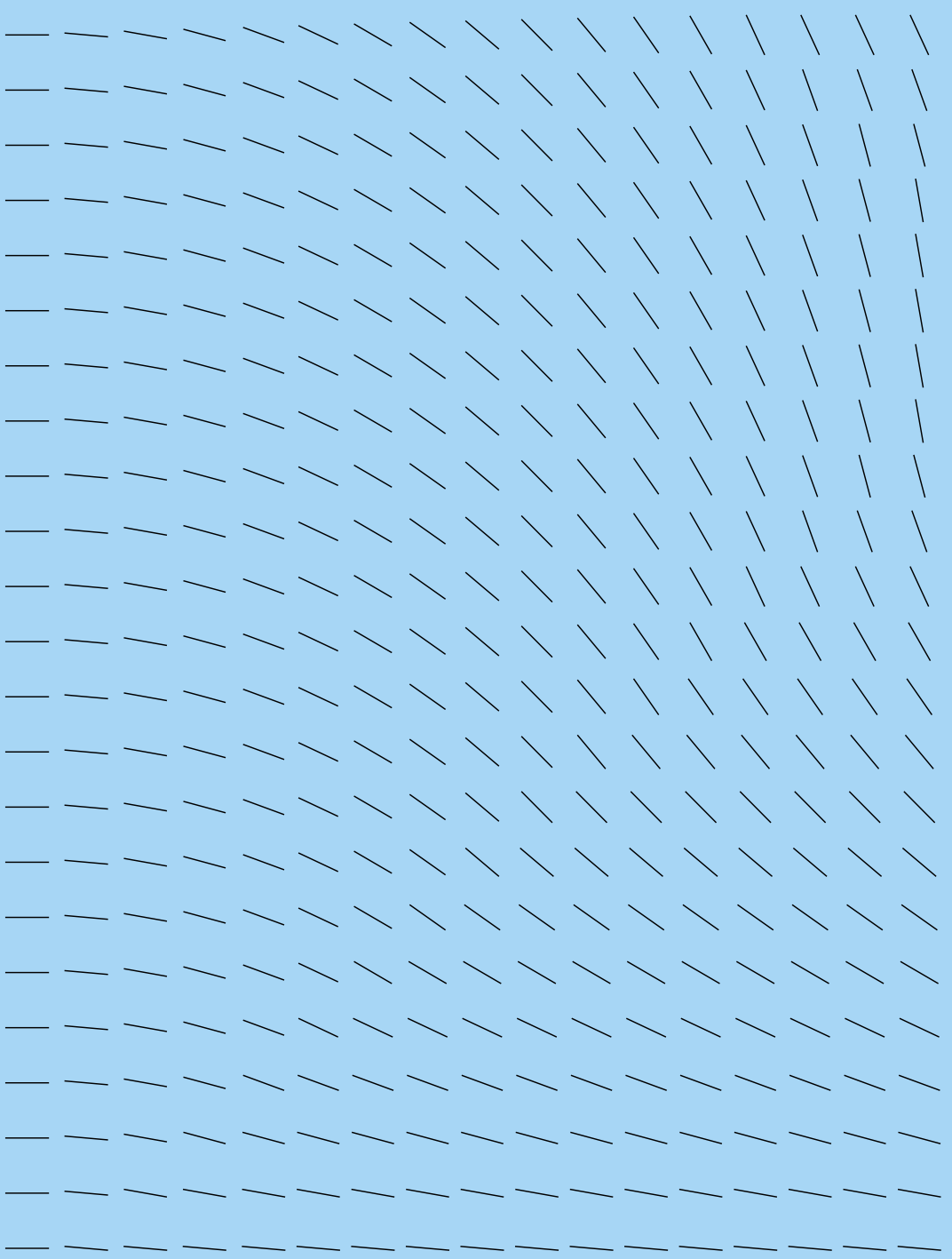
Berendt, Bettina (Ed.); Krzywdzinski, Martin (Ed.); Kuznetsova, Elizaveta (Ed.)

Erstveröffentlichung / Primary Publication
Konferenzband / conference proceedings

This work has been funded by the Federal Ministry of Education and Research of Germany (BMBF) (grant no.: 16DII121, 16DII122, 16DII123, 16DII124, 16DII125, 16DII126, 16DII127, 16DII128 - "Deutsches Internet-Institut").

Empfohlene Zitierung / Suggested Citation:
Berendt, B., Krzywdzinski, . M., & Kuznetsova, E. (Eds.). (2023). *Proceedings of the Weizenbaum Conference 2023: AI, Big Data, Social Media, and People on the Move*. Berlin: Weizenbaum Institute for the Networked Society - The German Internet Institute. https://doi.org/10.34669/wi.cp/5

weizenbaum
institut



Proceedings of the Weizenbaum Conference

AI, Big Data, Social Media, and People on the Move

June 2023

Weizenbaum Institute for the Networked Society –
The German Internet Institute

Proceedings of the Weizenbaum Conference 2023:
AI, Big Data, Social Media, and People on the Move

Proceedings of the Weizenbaum Conference 2023: AI, Big Data, Social Media, and People on the Move

Berlin, 2023

DOI 10.34669/wi.cp/5 \ ISSN 2510-7666

Weizenbaum Institute for the Networked Society –
The German Internet Institute
Hardenbergstraße 32 \ 10623 Berlin \ Tel.: +49 30 700141-001
info@weizenbaum-institut.de \ www.weizenbaum-institut.de

CONFERENCE CHAIRS & VOLUME EDITORS:
Bettina Berendt
Martin Krzywdzinski
Elizaveta Kuznetsova

EDITORIAL MANAGER:
Moritz Buchner

WEIZENBAUM INSTITUTE: The Weizenbaum Institute for the Networked Society – The German Internet Institute is a joint project funded by the Federal Ministry of Education and Research (BMBF) and the State of Berlin. It conducts interdisciplinary and basic research into the digital transformation of society through digitization and provides politicians, business and civil society with evidence- and value-based options for action in order to shape digitization in a sustainable, self-determined and responsible manner.

This work has been funded by the Federal Ministry of Education and Research of Germany (BMBF) (grant no.: 16DII121, 16DII122, 16DII123, 16DII124, 16DII125, 16DII126, 16DII127, 16DII128 – "Deutsches Internet-Institut").

Table of Contents

Short Papers

Abstracts

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

STANDARDIZATION AND HETEROGENIZATION

THE AUTOMATION OF MANAGEMENT AND THE MULTIPLICATION OF LABOUR

Altenried, Moritz
Berliner Institut für Empirische
Integrations- und Migrationsforschung
Humboldt-Universität zu Berlin
Berlin, Germany
moritz.altenried@hu-berlin.de

KEYWORDS

labour; algorithmic management; migration; platforms; gig economy

DOI: 10.34669/wi.cp/5.1

ABSTRACT

Algorithmic management is increasingly used to (semi-)automatically organise, measure and control labour in many sectors and industries. Based on empirical research in the (online and location-bound) gig economy, the paper argues that this digital automation of management allows for the quick and flexible inclusion of a broad range of workers in very diverse situations into production. This is shown, firstly, by the example of crowdwork platforms and their ability to integrate diverse and spatially distributed workers into labour processes. Secondly, the paper analyses the role of migrant labour for the urban gig economy and argues, that here, too, digital technologies and algorithmic management are to be understood as being part and parcel of a multifaceted process of the heterogenization of workforces. This particular effect and quality of algorithmic management and digital standardization is conceptually analysed in the framework of a multiplication of labour.

1 INTRODUCTION

Across the world of work digital technologies are increasingly used to plan, organise, measure and control labour and the labour process. From simple software to sophisticated machine learning applications, these technologies are profoundly transforming labour in contemporary capitalism. Not least in the context of covid-19, the development and implementation of such technologies has been dynamic, in places substituting for stagnating attempts to automate labour (Schaupp 2022a).

The shape and impact of these forms of algorithmic management vary greatly across different companies, sectors, and locations. Thus, it is very hard to give a concise and encompassing overview of the impact of these technologies on labour. It is, however, possible to identify specific tendencies that come with the proliferation of algorithmic management. The goal of this paper is exactly this: To analyse and conceptualise such a tendency. I will argue that forms of algorithmic management that (at least partly) automate the management of labour often have specific effects on the composition of workforces: this automation and digital organisation of management allows for the increasingly flexible and efficient inclusion of heterogeneous workforces into labour processes and supply chains. Migration and mobility are crucial expressions of this heterogeneity, but more factors of demographic and spatio-temporal flexibility come into play. I will describe these effects and affordances of technologies of algorithmic management in the context of a "multiplication of labour" (Mezzadra and Neilson 2013).

Empirically, the paper draws from multi-year ethnographic and qualitative research in different parts of the platform economy: global crowd or cloud work as well the location-bound urban gig economy. Based on comprehensive qualitative research into different platforms, I will show how algorithmic management enables the tightly controlled and standardized cooperation of a huge number of platform workers who can come from different backgrounds, experiences, and situations and who are distributed throughout space. Digitally (and often automatically) managed and standardized work procedures allow for the quick inclusion and remote organisation as well as substitutability and fluctuation of workers and hence contribute to the flexibilization and heterogenization of labour. Based on research in the gig economy in particular sectors and geographical locations, the findings are of course limited in their range, however, as I will argue in in the conclusion, the tendencies we can analyse here are observable way beyond the gig economy across the world of work in digital capitalism.

In the following, I will establish the terms algorithmic management and the multiplication of labour, before I move on to illustrate the interplay of these in the platform economy. The third part is focussed on the online gig economy and its global dynamics of distributed digital production, while the fourth part concentrates on the urban gig economy and the special role of migrant labour. In concluding, I shortly summarise and situate the findings.

2 ALGORITHMIC MANAGEMENT AND THE MULTIPLICATION OF LABOUR

With the term algorithmic management, I want to address a number of technologies designed to partly or completely automate organizational, coordination and control elements of the labour process (Lee et al. 2015; Moore 2017; Beverungen 2017). Instead of getting instructions and supervision directly from (middle) management, workers are given their orders and specifications via digital applications,

which manage, for example, workflows for office workers or navigation routes for delivery drivers. These forms of automated management are often helped by tracking, tracing, and rating and can use "nudging techniques" or elements of gamification. Algorithmic management is hence a broad – and somewhat imprecise – term bringing together a number of different techniques and technologies (Krzywdzinski and Gerber 2021). For the sake of this paper, this broad term shall suffice as I am particularly interested in the effects of automated (hence replicable and cheap/efficient on large scales) management which can indeed reach from direct control to gamified incentives.

Pioneered in but not limited to the gig economy, the extent of usage of algorithmic management varies across sectors and locations as well as the extent to which management processes are completely automated or human management works alongside and with the help of tools of digital management. While these forms of automated management certainly allow for new forms (and often a new granularity) of control over the labour process, they also have gaps and produce new forms and strategies of resistance by workers (Heiland 2022; Altenried and Niebler 2022). In many cases, it is not only (or not even primarily) the level or efficiency of direct control (which can be patchy or low), but the speed and cost-efficiency in the flexible inclusion of diverse workers into production processes that makes algorithmic management a factor in the transformation of work. Before looking at this empirically, I want to introduce the concept of the multiplication of labour to conceptualise this tendency.

Sandro Mezzadra and Brett Neilson describe the "multiplication of labour" as "the parallel operation of the three tendencies—intensification, diversification, and heterogenization of labor—that are increasingly reshaping labor experiences and conditions" (Mezzadra and Neilson 2013, 91-92). With the term, they strive to supplement the familiar term of the division of labour and hint at the heterogeneity of living labour in a time characterized by the increasing coalescing of labour and life, the increasing flexibilization of labour, as well as shifting overlapping production geographies in the ongoing processes of globalization. This centres not only the dynamics of migration in the production of labour markets but focusses on the "productive" role of borders in the constantly ongoing segmentation, fragmentation, temporalization of these markets and their overlapping and unstable borders in contemporary capitalism.

I would argue that the concept is also extremely effective in understanding major dynamics in the transformation of labour driven by digital technology (Altenried 2022). Digital technology, or, more precisely the standardization of tasks, the means of algorithmic management, and surveillance to organize the labour process, as well as the automated measuring of results and feedback often allow for a more efficient, temporal, and flexible inclusion of very heterogenous workforces into production processes. In other words, it is precisely the standardization of work that can profit from and allows for the multiplication of living labour in many ways. Looking at this from the perspective of the mobility of labour, it becomes crucial to research the interaction between algorithmic workplace regimes and migration regimes and how these are co-productive in the creation and transformation of segmented labour markets (Schaupp 2022b).

Platform labour illustrates this in a concentrated form. Here, for example, we can observe this multiplication quite literally in the sense that many combine two or more jobs, are logged into various apps and or combine wage labour and reproductive tasks at the same time. Beyond this first obvious dimension, digital platforms express many of the described tendencies of multiplication as they strive to flexibly and efficiently include workers (often addressed as independent contractors) from very heterogenous backgrounds, often for short amounts of time, into their production process precisely by automating large parts of the organisation and control of the labour process.

3 CROWDWORK: REMOTE ORGANISATION AND SPATIO-TEMPORAL FLEXIBILITY

I want to start the empirical part with a snapshot from my research on the online gig economy: A quote from Daniel, a crowdworker, I interviewed some years ago. He was 27 back then, a student and lived in Berlin-Wedding. He did crowdwork on a number of platforms to make ends meet. Talking about the ways he includes online platform work in his daily life he said: "Food in the oven— half an hour of working; if there is a break between two lectures, I'll quickly write another text on curtains on my laptop." He was supported by his parents, worked as student assistant and still needed €100-200 per months which he tried to earn on platforms whenever he had some free time to spare, his speciality being SEO-optimised product descriptions for online shops such as these curtains.

Today, the online gig economy, often also referred to as cloud or crowdwork, encompasses over two-thousand platforms such as Amazon Mechanical Turk, Freelancer, or Appen. These online labour platforms enact new forms of control and flexibility and serve as decentralized sites of digital production that are crucial to many nodes of the global economy, most notably the production and training of artificial intelligence (AI) (Altenried 2022; Gray and Suri 2019; Schmidt 2022). Training data for AI is only one, if nowadays the most dynamic, sector of crowdwork. Generally, these platforms outsource all kinds of digital work globally and we can see a huge variety between platforms, tasks, worker profiles as well as different forms of labour organisation and control (Krzywdzinski and Gerber 2020; Berg et al. 2018).

In the case of these platforms, the digital organisation and distribution of tasks, automated management, and surveillance and quality control allow for the inclusion of deeply heterogeneous workers without the need to spatially, temporally and subjectively homogenize them. Workers can access platforms from their homes, internet cafés, and even their mobile phones. In this way, the platforms are infrastructures opening up new labour pools previously difficult or impossible to reach as wage labourers and further diversify the workforce. The pause between lectures Daniel talked about is an example for this: a slice of time that has been previously unreachable for wage labour and can now almost seamlessly be integrated in a globally distributed but tightly (and automatically organised) production line. Another important example would be women with care responsibilities who combine reproductive tasks such as care for children or other relatives with crowdwork when they have a few hours or minutes to spare (Berg et al. 2018; Altenried and Wallis 2018; Wallis 2021).

Online platform labour folds onto existing economic geographies and transform them. Crowdwork has become, for example, important in locations with little alternatives in the labour market: from rural North America over urbanising Africa to refugee camps in Lebanon (Graham and Ferrari 2022; Hackl 2022), to name a few examples. Millions of digital home-based workers across the globe log daily into these platforms from their kitchens or living rooms to earn money from the tasks these platforms provide. Digital standardisation and algorithmic management as enacted by digital platforms make "work identifiable, searchable, and tradable at a truly planetary scale. Fixed material infrastructures of computing, international standards, and global payment systems allow the integration into broader systems of production of work that is broken into commodifiable chunks" as Fabian Ferrari and Mark Graham write (2022, 12). Even though the platforms workers come from very different backgrounds and situations and are located in vastly different geographical, cultural and temporal contexts, the algorithmic infrastructure of digital platforms synchronises their labour into a tightly organised production process.

4 URBAN GIG WORK: MIGRATION AND MOBILITY

With a second empirical snapshot, I want to move on to the location-bound gig economy providing services such as cleaning, cab rides or food delivery predominantly in urban areas. The urban gig economy provides a related yet particular impression of the dynamic interaction between automated management and the heterogenization of labour. In August of 2019, I interviewed Bastián, a Chilean food delivery rider in a park in Berlin-Neukölln. We were speaking about his decision to move to Berlin and how he started as a food delivery rider. "I always thought that it was an option working in Deliveroo, even when I was in Chile", he told me. For him and many other young migrants working for gig economy platforms was a known option, even before they arrived in Berlin. Not only amongst the migrants from Chile and Argentina, most of whom come to Berlin on a one-year visa like Bastián, "it's quite known that both Helpling and Deliveroo are the easy jobs to apply to when you come with a visa because you only have one year, and this is very immediately. They…, you don't need that much papers, and you don't need to speak German" as he explained.

The points he mentions already explain many of the reasons why gig economy platforms are a stronghold of migrant labour. Most platforms have a quick and unbureaucratic application process with very few formal requirements concerning qualifications, documents or skills. Many platforms even dispense with application interviews and only ask for a minimum of registration papers, work permits and similar documents and have few mechanisms to control the existence of these papers. For many (especially recent) migrants whose documentation, visa and permits would not suffice at other jobs, digital platforms are thus a quick way to start earning money.

Like Bastián, many migrant workers on digital platforms have numerous qualifications and degrees which are, however, often not accredited in Germany, a fact that contributes to protecting better-paid parts of the labour market against migrant workers. Another major and related problem for many migrants coming to Berlin is the German language. The availability even of precarious and low-skilled jobs becomes scarce and many find that their options diminish substantially without basic German skills. As the gig economy apps often work in several languages and are quite simple to operate, this offers possibilities even to those who speak no German or English.

The easy and quick accessibility of platforms like Deliveroo and the ability to earn money without the knowledge of the language makes those platforms important to many migrants especially in the time immediately after their arrival. As Bastián explained above, the option to work for such platforms is common knowledge among young people from Chile or Argentina wanting to come to Germany. In these cases, digital platforms become part of "migration infrastructures" as Biao Xiang and Johan Lindquist describe the "systematically interlinked technologies, institutions, and actors that facilitate and condition mobility" (Xiang and Lindquist 2014, 124; see also Altenried et al. 2018). While these platforms do not inhibit active brokerage positions such as labour agencies sending workers abroad, they enable new strategies, routes and pathways for migrant workers who base their mobility projects on platform labour and condition their differential, i.e. partial and temporal inclusion into national labour markets.

For platforms like Deliveroo, Helpling, Uber and many others, migrant workers constitute a crucial pool of workers forced to accept unstable and precarious conditions. While the importance of migrant labour especially to the service, gastronomy or taxi sectors of Berlin and many other cities is nothing new, digital platforms express a new special quality here. In fact, the labour model of the digital gig economy is geared almost perfectly towards the exploitation of migrant labour.

The systems of algorithmic management employed by the platforms via their apps allow for the (semi-)automated organisation and control of labour replacing, in large parts, human management while allowing for a new level of granular control and planning. Food delivery riders like Bastián, for example, need only minimal training, language skills or supervision as they are navigated by the app through urban space. These possibilities of digital and automated organisation, instruction and control make it possible and efficient for platforms, to hire workers who are new to a city and do not speak German or English, let them start working immediately and maybe let them go after a few weeks. In such cases, algorithmic management substitutes large amounts of training, forms of supervision, control or building of trust by human managers that would make it hugely in-efficient (and possible risky) for corporations to hire such workers only for a few weeks or months.

In the case of digital platforms, these mechanisms of algorithmic management develop their effect and efficiency in combination with contingent labour arrangements, i.e. the forms of self-employment, short-term or zero-hour contracts, or sub-contracting models found in the platform economy. It is also this very combination that allows platforms to accept a high number of workers as there are few fixed costs and risks are outsourced to the workers. Under these conditions, a high fluctuation in the workforce is no problem but rather part of the calculation of the platforms that can count on a latent reserve army of (migrant) workers who can be allowed into and expelled from the platforms with minimal costs and problems.

Indeed, there are very similar tendencies and logics at play in many European cities and even globally: More often than not, the platforms' workforces are in their majority migrant workers (Altenried, Bojadžijev, and Wallis 2020, Altenried 2021, Ferrari, Graham, and Van Doorn 2020; Gebrial 2022; Greef 2019; Liu 2019; Das and Srravya C 2021). Looking at the ways platforms' recruitment strategies profit from stratified and segmented labour markets that create a multiplicity of migrant situations and a reserve army of workers for the platforms, it becomes clear that without migrant labour, there would be no gig economy as we know it.

5 CONCLUSION

I would argue that both examples show how technologies of algorithmic management (ranging from simple functions to AI applications and very complex software) in the ways they are used in the platform economy participate in a flexibilization and heterogenization of workforces that I have described as a multiplication of labour. This is a multifaceted process encompassing the level of global production geographies and the shifting division of labour as well as the everyday lives of platform workers.

Platforms serve as distributed "digital factories" (Altenried 2022) that can, as in the case of crowdwork, coordinate tens of thousands of spatially distributed digital workers into tightly and automatically organised production processes without the need, however, to temporally, spatially, or subjectively homogenise them as, say, a Fordist factory needed to do it. Most of today's urban gig platforms, on the other hand, are based on predominantly migrant and often highly mobile workforces whose quick, flexible and temporal inclusion into the platforms labour process is predicated upon technologies of automated management.

Clearly, platforms are very vivid example of these tendencies and particular development in the world of work. However, the tendency I have described as the interplay between algorithmic management and the multiplication of labour becomes visible across many sites and in different forms. In

Amazon's warehouses, to take an example from outside the gig economy, the various technologies of standardization and algorithmic management reduce training times and increase control possibilities, thereby allowing flexible and short-term solutions in the recruitment of labour to satisfy the contingencies of supply chains for business peaks such as the weeks before Christmas when the workforce in many warehouses doubles. Seasonal labour, short-term contracts, and outsourced labour are important components of the labour regime in Amazon's distribution centres and proliferating beyond Amazon across different sectors and locations where we could find many more examples for the interplay of algorithmic management and the multiplication of labour for which the gig economy is an important laboratory.

REFERENCES

1. Altenried, Moritz. 2021. "Mobile workers, contingent labour: Migration, the gig economy and the multiplication labour." *Environment and Planning A: Economy and Space*. Online first, November 2021.
2. ---. 2022. *The Digital Factory: The Human Labor of Automation*. University of Chicago Press.
3. Altenried, Moritz, Manuela Bojadžijev, Sandro Mezzadra, Leif Höfler, and Mira Wallis. 2018. "Logistical Borderscapes: Politics and Mediation of Mobile Labor in Germany after the "Summer of Migration"." *South Atlantic Quarterly* 117 (2): 291-312.
4. Altenried, Moritz, Manuela Bojadžijev, and Mira Wallis. 2020. Platform Im/mobilities: Migration and the Gig Economy in Times of Covid-19. *Routed*, Issue 10. https://www.routedmagazine.com/platform-immobilities
5. Altenried, Moritz, and Valentin Niebler. 2022. "Fragmentierte Arbeit, verallgemeinerter Konflikt: Alltägliche Auseinandersetzungen in der Plattformarbeit." In *Widerstand im Arbeitsprozess. Eine arbeitssoziologische Einführung*, edited by Heiner Heiland and Simon Schaupp, 277-300. Bielefeld: transcript Verlag.
6. Altenried, Moritz, and Mira Wallis. 2018. "Zurück in die Zukunft: Digitale Heimarbeit." *Ökologisches Wirtschaften* 4/2018: 24-26.
7. Berg, Janine, Marianne Furrer, Ellie Harmon, Uma Rani, and M Six Silberman. 2018. *Digital labour platforms and the future of work.Towards Decent Work in the Online World*. Geneva: ILO.
8. Beverungen, Armin. 2017. "Algorithmisches Managements." In *Nach der Revolution: Ein Brevier digitaler Kulturen*, edited by Jörg Metelmann Timon Beyes, Claus Pias, 51-63. Berlin: Tempus.
9. Das, Kaarika, and Srravya C. 2021. Between Platform and Pandemic: Migrants In India's Gig Economy. *Futures of Work*, Issue 18. https://futuresofwork.co.uk/2021/03/16/caught-between-the-platform-and-the-pandemic-locating-migrants-in-indias-gig-economy/
10. Ferrari, Fabian, Mark Graham, and Niels Van Doorn. 2020. "Migration and migrant labour in the gig economy: An intervention." *Work, Employment and Society,* online first July 2022.
11. Gebrial, Dalia. 2022. "Racial platform capitalism: Empire, migration and the making of Uber in London." *Environment and Planning A: Economy and Space*, online first August 2022.
12. Graham, Mark, and Fabian Ferrari. 2022. *Digital Work in the Planetary Market*. Cambridge, Mass.: MIT Press.
13. Gray, M.L., and S. Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston and New York: Houghton Mifflin Harcourt.
14. Greef, Kimon de. 2019. "One More Way to Die: Delivering Food in Cape Town's Gig Economy". *New York Times*, August 24, 2019. https://www.nytimes.com/2019/08/24/world/africa/south-africa-delivery-deaths.html
15. Hackl, Andreas. 2022. "Digital Livelihoods in Exile: Refugee Work and the Planetary Digital Labor Market." In *Digital Work in the Planetary Market*, edited by F. Ferrari and M. Graham, 97-114. Cambridge, Mass.: MIT.

16. Heiland, Heiner. 2022. "Algorithmische Gegenmacht. Algorithmisches Management und Widerstand." In *Widerstand im Arbeitsprozess. Eine arbeitssoziologische Einführung*, edited by Heiner Heiland and Simon Schaupp, 301-326. Bielefeld: transcript Verlag.
17. Krzywdzinski, Martin, and Christine Gerber. 2020. *Varieties of platform work. Platforms and social inequality in Germany and the United States* (Weizenbaum Series, 7). Berlin: Weizenbaum Institute for the Networked Society.
18. ---. 2021. "Between automation and gamification: forms of labour control on crowdwork platforms." *Work in the Global Economy* 1 (1-2): 161-184.
19. Lee, Min Kyung, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. "Working with machines: The impact of algorithmic and data-driven management on human workers." *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 1603–1612
20. Liu, Hong Yu. 2019. Migrant workers in the digital market: China's platform economy. *The Asia Dialogue*, August 13, 2019. https://theasiadialogue.com/2019/08/13/migrant-workers-in-the-digital-market-chinas-platform-economy/
21. Mezzadra, Sandro, and Brett Neilson. 2013. *Border as Method, or, the Multiplication of Labor*. Durham and London: Duke University Press.
22. Moore, Phoebe V. 2017. *The Quantified Self in Precarity: Work, Technology and What Counts*. London: Routledge.
23. Schaupp, Simon. 2022a. "COVID-19, economic crises and digitalisation: How algorithmic management became an alternative to automation." *New Technology, Work and Employment*, online first May 2022.
24. ---. 2022b. "Algorithmic Integration and Precarious (Dis)Obedience: On the Co-Constitution of Migration Regime and Workplace Regime in Digitalised Manufacturing and Logistics." *Work, Employment and Society*, 36(2): 310–327.
25. Schmidt, Florian A. 2022. "The Planetary Stacking Order of Multilayered Crowd-AI Systems." In *Digital Work in the Planetary Market*, edited by Fabian Ferrari and Mark Graham, 137-156. Cambridge, Mass.: MIT Press.
26. Wallis, Mira. 2021. "Digital Labour and Social Reproduction–Crowdwork in Germany and Romania." *spheres: Journal for Digital Cultures* (6): 1-14.
27. Xiang, Biao, and Johan Lindquist. 2014. "Migration infrastructure." *International Migration Review* 48 (1): 122-S14.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

WHY DOES THE AI SAY THAT I AM TOO FAR AWAY FROM THE JOB MARKET?

Berman, Alexander
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Gothenburg, Sweden
alexander.berman@gu.se

KEYWORDS

public employment services; decision-support systems; explainability; machine learning; artificial intelligence; explainable artificial intelligence

DOI: 10.34669/wi.cp/5.2

1 INTRODUCTION

As artificial intelligence (AI) is increasingly being deployed in various domains such as healthcare (Qayyum et al., 2021), finance (Dastile, Celik & Potsane, 2020) and public welfare (Saxena et al., 2020; Carney, 2020), there is a growing need for understanding how stakeholders are affected by AI (Vaassen, 2022) and how to design and present explanations of AI-based decisions in ways that humans can understand and use (Miller, 2019). This paper contributes to these efforts by examining an AI-based decision-support system (DSS) launched by the Swedish Public Employment Service (PES) in 2020. Specifically, the study investigates to what extent the studied system enables affected jobseekers to understand the basis of AI-assisted decisions, to negotiate or contest dispreferred decisions, and to use the AI as a tool for increasing their job chances.

The rest of the paper is organised as follows: Section 2 situates the study in relation to previous work. Section 3 presents the empirical material, including technical information about the studied DSS. The main contribution of the paper is then presented in section 4 which elaborates weaknesses and limitations in the explainability of the system and how they could be addressed. Finally, section 5 offers some conclusions.

2 RELATED WORK

Previous studies have investigated the use of AI and algorithms in the context of PES from the perspectives of accuracy and discrimination (Desiere, Langenbucher & Struyven, 2019; Desiere & Struyven, 2021), norms and values embedded in algorithms (Sztandar-Sztanderska & Zielenska, 2020), austerity politics (Allhutter et al., 2020), caseworkers' attitudes and strategies (Assadi & Lundin, 2018; Sztandar-Sztanderska and Zielenska, 2022) and legal certainty (Carlsson, forthcoming). Few previous works have analysed explainability in relation to PES; exceptions include Niklas et al. (2015) who investigated the transparency of a Polish algorithmic profiling system and Zejnilovic et al. (2021) who studied the effects of explanations on caseworkers' decisions. An important basis for the present study is Scott et al.'s (2022) investigation of jobseekers' needs and desires in relation to algorithmic DSS. This paper extends previous work by technically describing the new Swedish AI-based DSS and by analysing explainability from the perspective of jobseekers' needs and interests.

3 CASE DESCRIPTION

The material presented below is based on public sources (cited where relevant) and information received from the agency via email (Nov 2021 – May 2023).

3.1 General information

In 2019, the Swedish government decided that a statistical tool[1] should be developed as an integrated part of the operations of the Public Employment Service (PES) in order to improve consistency and accuracy of labour-market related assessments, and thereby improve efficiency of resource

[1] The term "statistical assessment support tool" is used by both the government and the Swedish PES. In this paper, the terms "AI-based" and "statistical" are used interchangeably.

allocation.[2] Subsequently, the employment initiative Prepare and Match was launched in 2020 and rolled out nationally in 2021 (Hansson et al., 2022). The initiative enables enrolled jobseekers to get support, e.g. in the form of training or guidance, from a chosen provider. Decisions about access to the initiative are based on outputs from an AI-based DSS. The function of the AI is to assess the jobseeker's distance to the job market, with the purpose of targeting the employment agency's resources to those individuals that are most likely to find a job through the initiative.

Caseworkers are instructed to primarily adhere to the automated recommendation. Overruling a negative decision is difficult since it requires contacting a special working group within the agency. Interviews with caseworkers have indicated that some of them are reluctant to use this option since the working group rarely admits exceptions from automated recommendations (Bennmarker et al., 2021).[3]

3.2 Statistical model and decision algorithm

Decisions about access to the employment initiative are partly based on a statistical estimate of the jobseeker's probability of finding a job within 6 months. The statistical analysis encompasses 26 variables pertaining to personal information, including age, gender and education, as well as previous unemployment activities. It also involves data about the jobseeker's postal area, including levels of unemployment, income, education and citizenship (Bennmarker et al., 2021).

The statistical model is a neural network[4] trained on historical data consisting of 1.1 million profiles collected over a period of 10 years. The model estimates probabilities for 14 different future employment statuses; the DSS uses the sum of two of the outputs, corresponding to the probability of being employed within 6 months, either permanently or on fixed-term/part-time (Bennmarker et al., 2021).

The statistically estimated probability is combined with the jobseeker's current unemployment duration using threshold functions into three possible outcomes (Arbetsförmedlingen, 2020; see figure 1):

- Too near the job market – the jobseeker is deemed capable of finding a job with minor help, such as digital services
- Suitable for Prepare and Match
- Too far away from the job market – the jobseeker needs further investigation and other kinds of support

The thresholds between different outcomes are subject to political or administrative decisions related to e.g. available resources.

The system's accuracy, measured as the fraction of historical data points that are assigned an adequate decision (i.e. positive decision for jobseeker without job after 6 months, and vice versa), is 68%. Accuracy differs across sub-populations and decisions; the lowest accuracy is reported for negative decisions for jobseekers with disabilities (F1=17%) (Böhlmark, Lundström & Ornstein, 2021).

[2] https://www.esv.se/statsliggaren/regleringsbrev/?RBID=20264 (Accessed Jan 19, 2022)

[3] This can be contrasted with an earlier Swedish system, where caseworkers were instructed to consider the recommended decision carefully but to also use their professional judgement (Assadi & Lundin, 2018).

[4] The neural net has 64 inputs, two hidden layers and 14 outputs.

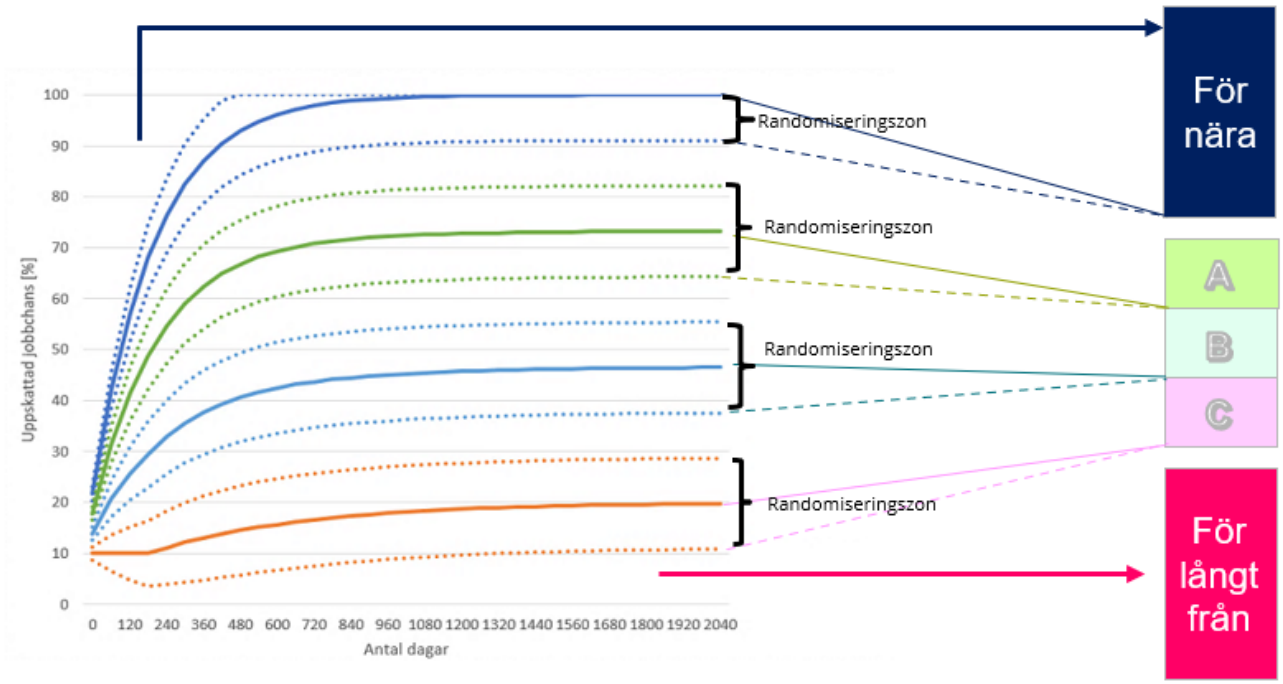


Figure 1. Thresholds and outcomes. Relationship between estimated probability of finding a job ("uppskattad jobbchans"), number of days of current unemployment ("antal dagar"), and outcome (too near ("för nära") or too far away from ("för långt ifrån") the labour market, or positive decision). For example, if job chance is estimated at 50% and unemployment duration is 360 days, the jobseeker is recommended access to the employment initiative. (Levels A-C affect the amount of compensation that providers receive.) Note that this illustration is not presented to jobseekers as part of any explanation. Reprinted with permission from Arbetsförmedlingen.

3.3 Explanations

Decisions about access to the employment initiative are communicated to the jobseeker in a meeting with a caseworker. Towards caseworkers, the recommended decision is shown in the case management system and is accompanied by a ranking of the 10 most important factors. The decision is also sent as a letter to the jobseeker and presented to the jobseeker when logged in at the agency's website. Towards jobseekers, only the top 4 most important factors are listed.

A suggested phrasing of the decision is automatically generated by the case management system (Arbetsförmedlingen, 2020). Below is an example of a positive decision (my translation):

> By comparing your information with statistics we have tried to assess how near you are the job market. Our assessment is that you will get the best help from a supervisor at one of the providers within the initiative Prepare and Match. In your case it was primarily the following factors that contributed to the assessment: Your unemployment duration, Your unemployment history, Your city of residence and Working time.

Unemployment duration is always presented as the most important factor. The rest of the factors are ranked using a method called LIME (Ribeiro, Singh & Guestrin, 2016). Given an input (i.e. data for a jobseeker at a particular point in time), LIME creates a simplified model by systematically investigating outcomes for various modifications of the input. For example, if age is assigned a high rank by LIME in a given case, it means that in situations similar to the case at hand, a different age tends to cause a different outcome.

Neither the estimated probabilities of different outcomes, the system's accuracy, or the thresholds and their influence on decisions are communicated to jobseekers or caseworkers. Implications of omitting such information will be discussed in section 4.

4 EXPLAINABILITY ANALYSIS

Previous studies of the DSS have shown that agency officials find it difficult to understand the basis for specific decisions, sometimes referring to the system as a black box (Bennmarker et al., 2021; Carlsson, forthcoming). The analysis below may illuminate why this is the case, although it takes the perspective of affected jobseekers rather than caseworkers. Following Scott et al. (2022), the analysis focuses on jobseekers' interests in *intelligibility* (outputs from system should be understandable) and *empowerment* (system should empower the jobseeker e.g. by providing actionable information).

4.1 Opaque internal logic

The statistical model is a neural network which, due to its non-linear processing and complex interactions between variables, is fairly opaque. Consequently, it is difficult even for AI experts with full access to the model to understand how the model reaches its judgements. This circumstance underpins many of the other issues raised below.

4.2 Unreliable explanation method

A common approach for explaining predictions by opaque models is to create a simpler, interpretable model that approximates the opaque model on a case-by-case basis, and then get explanations from the "surrogate" model instead. The agency uses one of the most popular techniques of this kind, called LIME (Ribeiro, Singh & Guestrin, 2016), to rank importance of factors.

While LIME and similar methods can give some insight into how an opaque model operates, the methods have been shown to be unstable: different explanations can be generated for the same prediction. Furthermore, since LIME and similar methods are approximate, explanations are not always faithful with respect to the outcomes that they are supposed to explain (Amparore, Perotti & Bajardi, 2021). In other words, the potential intelligibility afforded by approximate explanations comes at the cost of unreliability.

4.3 Misleading importance attribution for unemployment duration

In addition to unreliability issues associated with the explanation method as such, the special treatment of unemployment duration raises additional concerns. Towards jobseekers, the list of factors is presented as case-specific ("*In your case* it was primarily the following factors..."). However, this is misleading in the sense that current duration of unemployment is programmed to always appear first in the list. Furthermore, the special treatment leads to potentially inaccurate explanations, since the importance of unemployment duration may vary from case to case. As illustrated by figure 1, the effect of unemployment duration on decisions diminishes as duration increases. For example, we can consider a jobseeker that is deemed too far away from the job market, has been unemployed for 2000 days and is near the decision threshold (i.e. the estimated probability of finding a job is slightly below 20%). In such a situation, a positive decision would have required a much shorter unemployment

duration, or just a slight increase in estimated probability of finding a job (i.e. a potentially small change among other factors). In other words, there may exist cases where unemployment duration is less important than other factors.

4.4 Limited usefulness

Beyond issues regarding unreliability, previous work has shown that outputs from LIME and similar explanation methods can be difficult to interpret (Dieber & Kirrane, 2020). If city of residence is presented with a higher rank than working time for a negative outcome, what does this mean? Technically, the answer is that changing city of residence is more likely to lead to a positive outcome than changing working time. However, this information has limited value. For example, it does not explain *how* city of residence or working time would need to change in order to yield a positive decision.

Generally speaking, factor rankings do not enable the kind of counterfactual or contrastive reasoning that are common in human explanations. Research in linguistics and psychology has shown that humans tend to explain events in terms of conditions that would cause another event to occur (Miller, 2019). In the context of the current case, a counterfactual explanation for a negative decision could be expressed as: "If you would seek a full-time rather than part-time employment, your chances of finding a job would likely increase and you would be considered near enough the job market to get help within the initiative Prepare and Match". As argued by Wachter, Mittelstadt & Russell (2017), counterfactual explanations not only convey why or how a particular decision was reached, but also provide grounds to contest a decision and guidance on how to receive a different (e.g. more desired) outcome in the future. A similar recommendation is made by European Parliamentary Research Service in relation to automated decision-making, arguing that "data subjects who did not obtain the decision they hoped for should be provided with the specific information that most matters to them, namely, with the information on *what values for their features* determined in their case an unfavourable outcome" (Sartor & Lagioia, 2020, emphasis mine).

Since counterfactual explanations depend on notions of actionability that may differ between subjects (Rudin, 2019) – for example, switching from part-time to full-time work may be more feasible for some jobseekers than others – counterfactual explanations may require some kind of interaction between system and jobseeker (see section 4.6).

4.5 Choice of model

Several of the issues discussed above boil down to the opacity of the statistical model. Two international comparisons can illustrate how a more transparent model could potentially mitigate these issues. The Danish PES has used a decision tree with only five variables and very few interactions between variables. For example, if a jobseeker is unconfident about finding a job, the model predicts a 83% risk of future unemployment, regardless of other factors; if the jobseeker is more optimistic, the model uses three additional factors (age, previous employment rate and migration status) to categorise risk of unemployment into three different probabilities.[5] The Polish PES has used an algorithm with 24 questions scored from 0 (highest employability) to 8. Depending on the total score, the

[5] https://star.dk/media/12514/2020_01_31_beskrivelse_-_profilafklaringsvaerktoej_til_dagpengemodtagere.pdf (Accessed Feb 17, 2023)

jobseeker is categorised into one of three profiles (Sztandar-Sztanderska & Zielenska, 2020). In both the Danish and Polish case, the simplicity of the model eliminates the need of an additional explanation method; the models more or less "explain themselves". For example, if a Danish jobseeker wants to know why the model makes a particular prediction, a caseworker can show the decision tree in its entirety and highlight the path at hand. Seeing the entire decision tree also enables counterfactual reasoning, since it is easy to see how an alternative path leads to a different outcome. Similarly, if a Polish jobseeker wants to know how to increase his/her job chances (according to the algorithm), this information is directly contained in the scoring of individual questions. Note however that this requires that the scoring criteria are disclosed, which has not been the case with the Polish system (Niklas et al., 2015).

Is a simple and interpretable model, such as a small decision tree or a scoring algorithm, as accurate as a more opaque neural network? Comparing accuracy across countries is difficult, since the data varies between the countries. However, the Swedish PES has experimented with two models that are much simpler and explainable than the deployed one, and whose accuracy can be compared to the deployed model using the same data. The simplest model (a linear regressor), has an accuracy of 66%, comparable with the 68% for the deployed model (Ornstein & Thunström, 2021); a slightly more sophisticated model (small decision tree + 6 linear regressors) has an accuracy of 74% (Helgesson & Ornstein, 2021), i.e. *better* than the deployed model. This suggests that a simpler model can fulfil the stated goals – consistency and accuracy – equally well, or even better, than an opaque model, without the negative consequences for explainability that an opaque model brings about. This finding also resonates with some previous work on the relationship between accuracy and interpretability (Rudin, 2019).

4.6 Interactivity

Philosophical, cognitive, and social studies of explanations tend to emphasise their social nature: explanations involve transfer of knowledge in an interaction between an explainer and an explainee (Miller, 2019). In line with this, some scholars emphasise the potential values of interactive explanations (Miller, 2019; Arya et al., 2019; Weld & Bansal, 2019, Simkute et al. 2021; Lakkaraju et al. 2022; Berman & Howes, 2022; Cheng et al., 2019). For example, interactivity can enable stakeholders to ask "what-if" questions for hypothetical circumstances, without any need for simplified approximations (Wachter et al., 2017). If a jobseeker is denied access to the employment initiative, the possibility to ask questions such as "What if I move to Stockholm?" or "What if I get a university degree?" can help the jobseeker to not only understand how the AI makes its judgements, but also to use this understanding to negotiate or contest a decision. To the extent that the AI has learned something relevant about employability, exploration of hypothetical circumstances also enables the AI to be used as a coach for getting advise on how to get nearer the job market (Scott et al., 2022).

Supporting hypothetical questions is technically trivial; users only need to be equipped with a graphical interface which allows exploring how modifying the input affects the output. Interactivity could in principle also enable more open-ended counterfactual questions such as "What would motivate a positive decision in my case?", where the feasibility of changes in circumstances can be addressed in a dialogue between the system and jobseeker (Berman et al., 2022).

4.7 Accuracy

As mentioned in section 3.3, the accuracy of the system is not communicated to jobseekers or caseworkers, despite the fact that accuracy is far from perfect and varies greatly across different subpopulations and decisions. This makes it difficult for jobseekers to assign adequate degrees of trust in the AI. For example, to the extent that the AI can be used for getting actionable advise, knowledge about accuracy enables jobseekers to assess the reliability of the advise. Accuracy information also helps jobseekers to assess how appealable their case is; in situations where the AI is less accurate, there might be more room for negotiation.

To mitigate this, stakeholders could be provided with performance indicators for the relevant subpopulation and decision. For example, if a jobseeker with disabilities is rejected access to the initiative, the system could provide a reservation about its high uncertainty.

4.8 Thresholds and confidence estimates

As described in section 3.2, outcomes are partly governed by thresholds that are continuously adjusted by the agency. For example, if the agency lowers the threshold for positive decisions, some jobseekers may obtain a positive decision as a direct consequence of the changed threshold. However, the thresholds are mentioned neither in explanations for specific decisions or in general information to the public on the agency's web site. Arguably, concealing some of the factors that underpin decisions impedes jobseekers' ability to understand the basis for the decisions.

Furthermore, the probabilities of future employment statuses estimated by the statistical model are not communicated to stakeholders. As with accuracy information (see section 4.7), this makes it difficult for jobseekers to assess how much individual assessments can be trusted. Arguably, jobseekers have an interest in knowing if their decision is considered straightforward and univocal, or if it is a borderline case with high uncertainty. For example, if the model predicts a job chance of 5% for jobseeker A and 20% for jobseeker B, then both jobseekers are deemed too far away from the job market (assuming that they are both long-term unemployed). Nevertheless, jobseeker B is very near the threshold for a positive decision, and should therefore be in a more negotiable situation.

In this regard, explainability could potentially be enhanced by showing a simplified variant of figure 1, where the relevant region of the decision landscape has been zoomed in and/or highlighted. Additionally, the probability of making the correct decision given the current thresholds can be calculated and presented. For example, the confidence value would be near 50% for person B (indicating a very low confidence of recommending the right decision), while it would be higher for person A.

5 CONCLUSIONS

This case study of an AI-based decision-support system deployed by the Swedish Public Employment Service has shown that its justifications of decisions lack important information and are unreliable, potentially misleading and difficult to interpret. These weaknesses in explainability may affect jobseekers by influencing the caseworkers' decision-making; if caseworkers had access to more intelligible and reliable explanations, this might have affected their trust in the AI in either direction from case to case, and thereby also the final decisions. First and foremost, however, the study has

highlighted how the weaknesses impede jobseekers' ability to understand, negotiate and contest dispreferred decisions, and to get advise on how to increase their employment chances.

The good news is that many of the highlighted issues could be mitigated by replacing the current opaque statistical model with a simpler, more interpretable one; this would address jobseekers' interests and needs without necessarily impairing other desiderata. Increasing the degree of interactivity could also serve jobseekers' needs, potentially without replacing the statistical model.

It is important to note that the jobseeker perspective adopted in the present study is based on insights from previous research involving jobseekers in somewhat different contexts (Scott et al., 2022). In future work, it would be useful to empirically study the extent to which jobseekers find provided explanations intelligible and useful, e.g. using questionnaires and interviews. (Such studies could potentially also involve alternative, e.g. more interactive, forms of explanations.) It would also be interesting to collect and analyse caseworker-jobseeker conversations and study their strategies in relation to the AI.

Finally, it should be stressed that explainability is only one of many aspects to consider when assessing a decision-support system (other aspects include e.g. fairness). Nonetheless, this study may contribute to a better understanding of how choice of statistical model and design of explanations can impact the value and usefulness of an AI-based decision-support system from the perspective of those that are directly affected by the decisions; these insights may be relevant in other domains as well.

ACKNOWLEDGEMENTS

This work was supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

REFERENCES

1. Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. "Algorithmic profiling of job seekers in Austria: how austerity politics are made effective." *Frontiers in Big Data*, 5.
2. Amparore, Elvio, Alan Perotti, and Paolo Bajardi. 2021. "To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods." *PeerJ Computer Science* 7:e479.
3. Arbetsförmedlingen. 2020. *Arbetsförmedlingens handläggarstöd.* Dnr Af-2020/0016 7459.
4. Arya, Vijay, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques." *arXiv preprint arXiv:1909.03012.*
5. Assadi, Anahita, and Martin Lundin. 2018. "Street-level bureaucrats, rule-following and tenure: How assessment tools are used at the front line of the public sector." *Public Administration* 96 (1): 154–170.
6. Bennmarker, Helge, Martin Lundin, Tove Mörtlund, Kristina Sibbmark, Martin Söderström, and Johan Vikström. 2021. *Krom – erfarenheter från en ny matchningstjänst med fristående leverantörer inom arbetsmarknadspolitiken.* Institutet för arbetsmarknads- och utbildningspolitisk utvärdering (IFAU), July.
7. Berman, Alexander, Ellen Breitholtz, Jean-Philippe Bernardy, and Christine Howes. 2022. "Explaining Predictions with Enthymematic Counterfactuals." In *Proceedings of the 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22*, 95–100.

8. Berman, Alexander, and Christine Howes. 2022. “"Apparently acousticness is positively correlated with neuroticism". Conversational explanations of model predictions.” In *Proceedings of SEMDIAL 2022 (DubDial)*.
9. Böhlmark, Anders, Tom Lundström, and Petra Ornstein. 2021. *Träffsäkerhet och likabehandling vid automatiserade anvisningar inom Rusta och matcha. En kvalitetsgranskning.* Arbetsförmedlingen analys.
10. Carlsson, Vanja. Forthcoming.
11. Carney, Terry. 2020. “Artificial intelligence in welfare: Striking the vulnerability balance?” *Monash University Law Review* 46 (2): 23–51.
12. Cheng, Hao-Fei, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. “Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders.” In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12.
13. Dastile, Xolani, Turgay Celik, and Moshe Potsane. 2020. “Statistical and machine learning models in credit scoring: A systematic literature survey.” *Applied Soft Computing* 91:106263.
14. Desiere, Sam, Kristine Langenbucher, and Ludo Struyven. 2019. “Statistical profiling in public employment services,” *OECD Social, Employment and Migration Working Papers*, no. 224, https://doi.org/10.1787/b5e5f16e-en.
15. Desiere, Sam, and Ludo Struyven. 2021. “Using artificial intelligence to classify jobseekers: the accuracy-equity trade-off.” *Journal of Social Policy* 50 (2): 367–385.
16. Dieber, Jürgen, and Sabrina Kirrane. 2020. *Why model why? Assessing the strengths and limitations of LIME.* https://doi.org/10.48550/ARXIV.2012.00093.
17. Hansson, Ewa, Gioia Luigetti, Martin Waara, and Stefan Öster. 2022. *ESF-projekt Kundval rusta och matcha. Slutrapport.* Arbetsförmedlingen.
18. Helgesson, Petter, and Petra Ornstein. 2021. *Vad avgör träffsäkerheten i bedömningar av arbetssökandes stödbehov? En undersökning av förutsättningarna för statistiska bedömningar av avstånd till arbetsmarknaden, med fokus på betydelsen av inskrivningstid.* Arbetsförmedlingen analys.
19. Lakkaraju, Himabindu, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. “Rethinking Explainability as a Dialogue: A Practitioner’s Perspective.” *arXiv preprint arXiv:2202.01875.*
20. Miller, Tim. 2019. “Explanation in artificial intelligence: Insights from the social sciences.” *Artificial intelligence* 267:1–38.
21. Niklas, Jędrzej, Karolina Sztandar-Sztanderska, Katarzyna Szymielewicz, A Baczko-Dombi, and A Walkowiak. 2015. *Profiling the unemployed in Poland: social and political implications of algorithmic decision making.* Fundacja Panoptykon.
22. Ornstein, Petra, and Hanna Thunström. 2021. *Träffsäkerhet i bedömningen av arbetssökande. En jämförelse av arbetsförmedlare och ett statistiskt bedömningsverktyg.* Arbetsförmedlingen analys.
23. Qayyum, Adnan, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. 2021. “Secure and Robust Machine Learning for Healthcare: A Survey.” *IEEE Reviews in Biomedical Engineering* 14:156–180. https://doi.org/10.1109/RBME.2020
24. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. KDD ’16. San Francisco, California, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778
25. Rudin, Cynthia. 2019. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nature Machine Intelligence* 1 (5): 206–215.
26. Sartor, Giovanni, and Francesca Lagioia. 2020. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.* European Parliamentary Research Service.
27. Saxena, Devansh, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. “A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376229

28. Scott, Kristen M, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. “Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective.” In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2138–2148.
29. Simkute, Auste, Ewa Luger, Bronwyn Jones, Michael Evans, and Rhianne Jones. 2021. *“Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable.”* Journal of Responsible Technology 7-8:100017. https://doi.org/10.1016/j.jrt.2021.100017
30. Sztandar-Sztanderska, Karolina, and Marianna Zielenska. 2020. “What Makes an Ideal Unemployed Person? Values and Norms Encapsulated in a Computerized Profiling Tool.” *Social Work & Society* 18(1) (May).
31. Sztandar-Sztanderska, Karolina, and Marianna Zielenska. 2022. “When a Human Says “No” to a Computer: Frontline Oversight of the Profiling Algorithm in Public Employment Services in Poland.” *Sozialer Fortschritt* 71 (6-7): 465–487.
32. Vaassen, Bram. 2022. “AI, Opacity, and Personal Autonomy.” *Philosophy & Technology* 35 (4): 88.
33. Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual explanations without opening the black box: Automated decisions and the GDPR.” *Harv. JL & Tech.* 31:841.
34. Weld, Daniel S, and Gagan Bansal. 2019. “The challenge of crafting intelligible intelligence.” *Communications of the ACM* 62 (6): 70–79.
35. Zejnilovic, Leid, Susana Lavado, Carlos Soares, Íñigo Martınez De Rituerto De Troya, Andrew Bell, and Rayid Ghani. 2021. “Machine Learning Informed Decision-Making with Interpreted Model’s Outputs: A Field Intervention.” In *Academy of Management Proceedings*, 2021:15424. 1. Academy of Management Briarcliff Manor, NY 10510.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

AI AND INEQUALITY IN HIRING AND RECRUITING

A FIELD SCAN

Dinika, Adio-Adet
Bremen International Graduate School of Social Sciences (BIGSSS),
Distributed AI (DAIR) Institute
Bremen, Germany
adinika@uni-bremen.de

Sloane, Mona
University of Virginia
Charlottesville, USA
mona.sloane@virginia.edu

KEYWORDS

artificial intelligence, recruiting, inequality, STEM, gender, bias

DOI: 10.34669/wi.cp/5.3

ABSTRACT

This paper provides a field scan of scholarly work on AI and hiring. It addresses the issue that there still is no comprehensive understanding of how technical, social science, and managerial scholarships around AI bias, recruiting, and inequality in the labor market intersect, particularly vis-à-vis the STEM field. It reports on a semi-systematic literature review and identifies three overlapping meta themes: productivity, gender, and AI bias. It critically discusses these themes and makes recommendations for future work.

1 INTRODUCTION

Artificial intelligence (AI) has taken a strong foothold in the human resources (HR) management domain, and recruiting specifically: the global market of AI-driven tools used in recruiting is expected to grow to $695 million (A2Z Market Research 2022), seemingly addressing automation needs of recruiting professionals across the hiring funnel. These developments run parallel to two other major phenomena: an ever raging "war on talent" in the science, technology, engineering, and mathematics (STEM) fields that maps onto the most promising fields of technology innovation and global competition, most recently the semiconductor industry (Shein 2023); and the escalation of global inequality (Savage 2021; Piketty 2014).

While regulatory concerns around enhancing the technical workforce rub shoulders with efforts to curb AI bias in recruiting, there still is no comprehensive understanding of how technical, social science, and managerial scholarships around AI bias, recruiting, and inequality in the labor market intersect, and importantly what similar or distinct narratives emerge. Therefore, the objective of this paper is to provide a comprehensive overview of the research conducted on these intersecting topics, to synthesize the knowledge, and to identify key themes that can provide new avenues for interdisciplinary research on AI, inequality, and recruiting.

2 METHODS

Since the goal of this paper is to accurately map existing scholarly works on AI bias, recruiting, and inequality in STEM across more than one discipline, a literature review is the most appropriate method. As the intersection of AI bias, recruiting, and inequality in STEM is a novel topic, a semi-systematic approach to literature review was chosen as this method provides an *overview* of a research area. Furthermore, the semi-systematic approach focuses primarily on research articles as source material rather than, for example, quantitative datasets from past studies (Snyder 2019). Also labeled a narrative review approach, it allows the examination of *topics* that have been conceptualized differently by different research communities across diverse disciplines (Wong et al. 2013). To provide a field scan, rather than an exhaustive literature review, the semi-systematic approach mandates a meta view that does not aim to review every single paper on any given topic, but that aims at reviewing a topic by way of examining how a topic has developed across a selection of disciplinary domains (Snyder 2019).

For this paper, we used a three-step strategy for our semi-systematic literature review. First, we conducted a high-level topical search across social science and technical fields to identify the most relevant domains for our more focused literature review. We made this decision based on papers available per topic (AI bias, recruiting, and inequality in STEM) per domain. Technical scholarship, mostly in computer science, social science, and management emerged as the dominant domains. We analyzed the collected papers by way of qualitative coding to identify key themes. In a second step, we expanded our literature search by way of keyword search along the themes and across these three domains and collected further relevant scholarly papers that were available. Our final dataset included 56 papers. In the third and last step, we read and analyzed the papers, summarizing and further coding the data to condense themes which we then clustered into the three meta themes we discuss below: productivity, gender, and AI bias. It is important to note that these themes are not neatly distinct but overlap and converge at times, and that cited papers are representative of wider discourses. It is also

important to note that due to language restrictions (all papers had to be available in English to be accessible by the research team), selection bias occurred.

3 FINDINGS

3.1 Productivity

The first meta theme of *productivity* describes the major concern of making all things recruiting more "efficient" by way of introducing AI. It echoes well-known narratives of techno-solutionism deployed in the general discourse around work and automation. Generally, the use of AI in recruitment is already on the rise, with 98% of Britain's Fortune 500 companies using automated hiring systems (AHSs) to onboard employees (Graham et al. 2020). In 2019, 99% of Fortune 500 companies used Applicant Tracking Systems (ATS) such as Workday, Taleo, SuccessFactors, BrassRing, and iCIMS (Hu 2019), to recruit and onboard new employees. While ATS have been around for a long time, they are increasingly equipped with AI-driven tools, such as resume screeners and candidate rankers (Manatal 2022), adding on to the already full AI-toolbox that recruiters use consisting of natural language processing (NLP) tools to write more "inclusive" job descriptions and ads, targeted advertising for placing job ads on various platforms and in different outlets, AI-driven job and talent search platforms to search for suitable job candidates, video interviewing software, AI-driven personality testing, automated skills assessment, or chatbots.

The latter in particular has been hailed as increasing efficiency in in candidate communication, for example by providing real-time feedback, addressing inquiries, and consistently engaging with the candidates throughout the hiring process (Brishti and Javed 2020). In particular, chatbots are perceived by recruiters as improving accessibility and lowering the application threshold (Koivunen et al. 2022), not least because they can answer a candidate's questions and address their concerns before they even apply (Nawaz and Gomes 2019). Similarly, chatbots have been depicted as a "quick and easy way" to improve efficiency and performance due to their 24/7-availability (Zamora 2017), helping recruiters understand the experience of a candidate, and to automate the more administrative side of recruiting, such as scheduling interviews.

This falls in line with common narratives around increased efficiency and reduced cost in AI-driven recruiting (Singh and Finn 2003; Okolie and Irabor 2017), due to what is perceived as a streamlining of the hiring process, especially in the context of pre-selecting suitable candidates (Derous and De Fruyt 2016). The latter has become more challenging for recruiters as technology has significantly lowered the application threshold and increased the application volumes, leaving HR practitioners with a growing amount of data they must take into account (Guo et al. 2021). This means that recruiters are expected to use technology to increase their productivity but are also facing enlarged workloads *due* to technology. Yet, recruiting professionals see using AI-enabled software as an efficient way of processing candidate data and, thus, as a pathway for introducing or advancing candidates from broader and more diverse pools (L. Li et al. 2021). Invoking the idea of tech neutrality, scholars have also suggested that AI-driven tools are less prone to bias and can be more impartial than human recruiters (Upadhyay and Khandelwal 2018), partially by manipulating AI to avoid bias (Black and van Esch 2020).

3.2 Gender

The second meta theme of *gender* emerges specifically vis-à-vis the STEM labor market and a lack of diversity across STEM jobs. Here, it appears to be undisputed that the "pipeline problem" – i.e., not having sufficient input and retention of STEM students – maps most strongly onto a stark gender divide (Hill, Corbett, and St. Rose 2010; Beede et al. 2011), leading to notoriously small and homogenous talent pools in the STEM fields. For example, in 2019, 73% of all STEM workers in the US were men (US Census Bureau 2021). Various theories percolate around the cause of such strongly sustained gender divisions in the STEM field. These range from a lack of girls' social identity with mathematics (Akin, Santillan, and Valentino 2022) to access to education and educational choices (Hanson and Krywult-Albańska 2020; Bertrand 2020). Scholars have outlined that children display equal interest in mathematics, regardless of gender, in primary and secondary school (Riegle-Crumb et al. 2012) but then diverge in middle school (Akin, Santillan, and Valentino 2022; Seo, Shen, and Alfaro 2019). It has been argued that a deliberate investment in the up-scaling of enrolment of women into STEM programs can have a positive influence on retaining women in STEM-related jobs (Botella et al. 2019). Flowing from these concerns is a present and overwhelming narrative of needing to support women's careers in STEM.

Treated somewhat separate to this body of work is scholarship on the harmful effects of gender stereotyping on education and the composition of the workforce, and the STEM workforce specifically. Here, it has been argued that relative poorer performance of girls and women in mathematics is gender-constructed (Bertrand 2020) which is, for example, evidenced in mathematics teachers' implicit stereotyping having a measurable negative effect on girls' performance in mathematics (Carlana 2019) or in primary school teachers' biases favoring boys, which has been demonstrated to have a positive effect on boys' in-class achievements and enrollments in advanced-level mathematics courses with a corresponding negative effect on girls (Lavy and Sand 2015). These processes reinforce socially inappropriate roles for women and men, with material effects on women's STEM careers, particularly in the academe, as they are considered less able than men by important institutional players such as grant reviewers for the US National Institute of Health (Magua et al. 2017).

It can be argued that gender stereotyping is symptomatic of wider systems of oppression and (gender) inequality that are so infrastructural to the organization of social life that they take hold long before people enter any form of education (Gomez-Herrera and Koeszegi 2022). Historians have shown that these systems materialize in narratives around ability, skill, and power that have excluded women from ascending to more powerful positions alongside the rise of computer, even though "computing", originally, was a high-skill, low reputation role typically occupied by women (Hicks 2017). The knock-on effects of this "gender shift" reinforce the "vicious cycle of digital inequality" in which inequalities and gender stereotypes in society underpin segregation in society and the professions, leading to technologies amplifying (gender) inequality (Gomez-Herrera and Koeszegi 2022). The AI field is a prime example of this dynamic with women accounting for only 22% of all AI and computer science higher degree programs in North America in 2019, and currently only 26% of the data and AI workforce being classified as women (Deloitte 2022).

When putting these findings in context with the composition of the field of human resource management, a field which is currently undergoing rapid technological change, including in recruiting, an equally stark gender divide emerges. In the US, the occupation of "human resource manager" is comprised of over 80% women. The opposite is true for the tech industry that is fueling the AI-fication

of recruiting, the "computer and mathematical occupations", which has an only 26% share of women (U.S. Bureau of Labor Statistics 2021).

3.3 AI Bias

The third and last meta theme that emerges at the intersection of equity in AI, recruiting, and STEM in the fields of computer science, social science, and management scholarship is the theme of *AI bias*. The potentially discriminatory effects of AI in general have, by now, been aptly demonstrated. For the purposes of this paper, these can best be schematized by mapping them onto two dominant AI techniques: computer vision and natural language processing (NLP). Computer vision systems set out to replicate elements of the human vision system and train computers to identify and parts of the complexity of the human vision system and enabling computers to locate and classify objects in images and videos (Mihajlovic 2021). In the context of recruiting, computer vision-based AI is, for example, used in extracting text from the image of a CV, or for the automated analysis of virtual interviews. NLP sets out to model human language by way of combining computational linguistics with statistical analysis, machine learning and deep learning in order to "understand" the meaning of written or spoken speech (IBM n.d.; Yse 2019). NLP used in recruiting includes AI-driven systems used to write job ads, as well as candidate search systems, resume parsers, pre-screening processes, chatbots, and more (Recruiter.com n.d.).

Evidence of intersectional discrimination in computer vision has prominently been proposed by (Buolamwini and Gebru 2018) who demonstrated that facial recognition technologies show disproportionately higher inaccuracy rates for women with darker skin tones. Unevenly distributed false positives in facial recognition technology amplify racial discrimination (Najibi 2020), for example in policing in the United States (Crockford 2020; Perkowitz 2021) and in education. Facial recognition technology used in online proctoring during the Covid-19 pandemic has been shown to be biased against students with certain skin tones and genders (Yoder-Himes et al. 2022). Similarly, word embedding, a framework used in NLP, replicate societal bias and provide pathways for perpetuating sexist tropes (Bolukbasi et al. 2016), as well as perpetuate historic biases more generally (Caliskan 2019).

In the context of recruiting, researchers have found that resume search engines working with text data and demographic features can produce rankings that disadvantage some candidates (Chen et al. 2018). Others have outlined how the automation of hiring by way of algorithmic systems can facilitate and obfuscate employment discrimination (Ajunwa 2019), especially in the context of hiring platforms (Ajunwa and Greene 2019) and algorithmic systems used for workforce management (Ajunwa 2020). Issues of validity in personality-assessment tools used in recruiting have been demonstrated by an interdisciplinary team of scholars conducting a stealth audit (Rhea et al. 2022) while investigative journalists have highlighted how hiring AI increasingly works as "black box" gatekeeper in the hiring process (Schellmann 2022), including in public agencies (Varner 2021). A more nascent body of work examines how recruiters use and make sense, and often only reluctantly embrace, various AI tools (L. Li et al. 2021).

The latter body of work connects to older scholarships on *human* bias in recruiting. For example, well known studies have shown strong bias against African-American-sounding names in the application process (Bertrand and Mullainathan 2004) and underlined the formation of ethnic bias in resume screening (Derous and Ryan 2019) which can lead to job candidates from racial minorities to engage

in "résumé whitening" (Kang et al. 2016). Work on gender discrimination in hiring, similarly, has long demonstrated how recruiting bias disproportionately affects women (Birkelund et al. 2022; Barron et al. 2022), particularly women of childbearing age (K. K. Li et al. 2022). Interestingly, the issue of human bias in recruiting has been used as the main argument *for* seemingly "neutral" AI applications in HR more broadly (Raghavan et al. 2020), promising to decrease gender discrimination specifically (Pisanelli 2022), often couched in narratives of "scientism" (Vassilopoulou et al. 2022). More recent studies, however, have shown that these types of claims are misleading, misconstruing AI technology as neutral and misunderstanding the dynamics of gender and race (Drage and Mackereth 2022).

4 DISCUSSION

To provide new avenues for interdisciplinary research on AI, inequality, and recruiting, it is helpful to have a clear understanding of how the three meta themes of productivity, gender, and AI bias emerge at the intersection of technical, social science, and managerial scholarship. However, it is equally important to critically discuss these themes to chart their limitations and make future work more effective.

Whilst the term "bias" has been productive for highlighting both the allocative and the representational harms that AI can cause (Barocas, Hardt, and Narayanan 2021), it also has been critiqued as being conditioned on an inherently normative process and as not being connected well across disciplines (Blodgett et al. 2020). It also tends to skew conversations around AI harm towards training data rather than societal inequalities (Sloane 2019), organizational decision making (Moss and Metcalf 2020; Sloane and Zakrzewski 2022; Rakova et al. 2021), and algorithm and models themselves (O'Neil 2016; Zou and Schiebinger 2018). In the context of recruiting and hiring specifically, a narrow focus on bias also precludes a much needed critical examination of the potentially discriminatory assumptions baked into AI (Sloane, Moss, and Chowdhury 2022), as well as locates bias in either people or technologies, rather than in socio-technical systems. This precludes a closer examination of how socio-technical bias occurs in the hiring funnel. We propose that to address this issue, a closer examination of socio-technical systems is critical. Bias could emerge from these systems because of the interactions and relationships between these social and technical components, not merely between individuals and technologies. Therefore, a practice-based approach that focuses on how technologies are used and made sense of in discretionary decision making is vital. Future research in this area should include investigations into the role of organizational structures, work processes, technological implementation, continuous data input and interpretation and of bias mitigations trainings. Such an investigation will be paramount for enhancing the understanding of harm produced by socio-technical systems in HR.

Similarly, scholars have outlined the limitations of the "pipeline problem", demonstrating that "improving the pipeline" does not necessarily improve discriminatory workplace cultures in STEM institutions (Rankin 2022), or increase diversity in the workforce (Dickey 2021; Bui and Miller 2016). To the contrary, it has been argued that ramping up the enrollment of women in STEM clusters runs the risk of labeling women as "affirmative enrollments" (Heilman, Block, and Stathatos 1997) or of framing gender as binary or one-dimensional. One could also argue that the gender and racial stereotyping that occurs as a function of social stratification in society is amplified by the representational harm that is propagated through label and unlabeled data that AI models are trained on. Indeed, scholars have demonstrated that demographic information about individuals can be inferred from online

data without said individuals explicitly relaying such information (Karimi et al. 2016; Fiscella and Fremont 2006). This highlights the need of technical education of HR and specifically recruiting professionals so that they themselves are literate in the potential generation of bias in socio-technical systems.

Whilst increased productivity has been framed as major driver for AI-adoption in recruiting, we still know very little about e-recruiting in general (Chapman and Gödöllei 2017) and specifically how professional recruiters actually use AI in their professional practice, and if there indeed is an increase in productivity ushered in by AI. What is known, however, is that HR practitioners remain critical of the technology, lacking a trust in data accuracy and decrying an inadequate level of control over algorithmic candidate matches (L. Li et al. 2021). There appears to be a clear need for a more decided engagement of HR professionals in shaping choices around AI in the professional practice of recruiting, not least to circumnavigate what is perceived as threats of largescale automation (Anthony 2021; Charlwood and Guenole 2022).

5 CONCLUSION

This paper has addressed the issue that there still is no comprehensive understanding of how technical, social science, and managerial scholarships around AI bias, recruiting, and inequality in the labor market intersect, particularly vis-à-vis the STEM field. It has reported on a semi-systematic literature review and concluded that currently three overlapping themes dominate: productivity, gender, and AI bias. It has detailed each theme before critically discussing the findings. The key take-away from this study is that the overwhelmingly female and white profession of HR and recruiting is substantially changed through the introduction of AI, which is initiated and driven by the predominantly male and majority white "computer and mathematical occupations" (U.S. Bureau of Labor Statistics 2021). Here, a "gender flipping" (Hicks 2017) occurs that sees men slotted into feminized jobs, here by way of the technology itself, as well as a further racial stratification of the HR industry. Future work should focus further on critically examining this dynamic across an even wider spectrum of disciplines (such gender, critical race, and disability studies) to inform applied AI design, HR management, and policymaking. Such an approach could, for example, be facilitated by way of using social practice theory, or example by way of taking a practice-based approach (Shove, Pantzar, and Watson 2012) (Sloane and Moss 2022).

ACKNOWLEDGEMENTS

This was in part supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, as well as the NYU Center for Responsible AI at the Tandon School of Engineering at New York University.

REFERENCES

1 A2Z Market Research. 2022. “AI Recruitment Market: All the Stats, Facts, and Data Youll Ever Need to Know.” Digital Journal. October 7, 2022. https://www.digitaljournal.com/pr/ai-recruitment-market-all-the-stats-facts-and-data-youll-ever-need-to-know

2 Ajunwa, Ifeoma. 2019. “An Auditing Imperative for Automated Hiring.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.3437631

3 ———. 2020. “The ‘Black Box’ at Work.” *Big Data & Society* 7 (2): 205395172096618. https://doi.org/10.1177/2053951720938093

4 Ajunwa, Ifeoma, and Daniel Greene. 2019. “Platforms at Work: Automated Hiring Platforms and Other New Intermediaries in the Organization of Work.” In *Work and Labor in the Digital Age*, edited by Steve P. Vallas and Anne Kovalainen, 33:61–91. Research in the Sociology of Work. Emerald Publishing Limited. https://doi.org/10.1108/S0277-283320190000033005

5 Akin, V., S. T. Santillan, and L. Valentino. 2022. “Strengthening the STEM Pipeline for Women: An Interdisciplinary Model for Improving Math Identity.” *PRIMUS* 0 (0): 1–24. https://doi.org/10.1080/10511970.2022.2032506

6 Anthony, Callen. 2021. “When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies.” *Administrative Science Quarterly* 66 (4): 1173–1212. https://doi.org/10.1177/00018392211016755

7 Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2021. *Fairness and Machine Learning*.

8 Barron, Kai, Ruth Ditlmann, Stefan Gehrig, and Sebastian Schweighofer-Kodritsch. 2022. “Explicit and Implicit Belief-Based Gender Discrimination: A Hiring Experiment.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4097858

9 Beede, David N., Tiffany A. Julian, David Langdon, George McKittrick, Beethika Khan, and Mark E. Doms. 2011. “Women in STEM: A Gender Gap to Innovation.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.1964782

10 Bertrand, Marianne. 2020. “Gender in the Twenty-First Century.” *AEA Papers and Proceedings* 110 (May): 1–24. https://doi.org/10.1257/pandp.20201126

11 Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94 (4): 991–1013. https://doi.org/10.1257/0002828042002561

12 Birkelund, Gunn Elisabeth, Bram Lancee, Edvard Nergård Larsen, Javier G Polavieja, Jonas Radl, and Ruta Yemane. 2022. “Gender Discrimination in Hiring: Evidence from a Cross-National Harmonized Field Experiment.” *European Sociological Review* 38 (3): 337–54. https://doi.org/10.1093/esr/jcab043

13 Black, J. Stewart, and Patrick van Esch. 2020. “AI-Enabled Recruiting: What Is It and How Should a Manager Use It?” *Business Horizons*, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, 63 (2): 215–26. https://doi.org/10.1016/j.bushor.2019.12.001

14 Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. “Language (Technology) Is Power: A Critical Survey of ‘Bias’ in NLP.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–76. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

15 Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” In *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

16 Botella, Carmen, Silvia Rueda, Emilia López-Iñesta, and Paula Marzal. 2019. “Gender Diversity in STEM Disciplines: A Multiple Factor Problem.” *Entropy* 21 (1): 30. https://doi.org/10.3390/e21010030

17 Brishti, Juthika Kabir, and Ayesha Javed. 2020. *THE VIABILITY OF AI-BASED RECRUITMENT PROCESS : A Systematic Literature Review*. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-172311

18 Bui, Quoctrung, and Claire Cain Miller. 2016. “Why Tech Degrees Are Not Putting More Blacks and Hispanics Into Tech Jobs.” *The New York Times*, February 25, 2016, sec. The Upshot. https://www.nytimes.com/2016/02/26/upshot/dont-blame-recruiting-pipeline-for-lack-of-diversity-in-tech.html

19 Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

20 Caliskan, Aylin. 2019. “Applying the Right Relationship Marketing Strategy through Big Five Personality Traits.” *Journal of Relationship Marketing* 18 (3): 196–215. https://doi.org/10.1080/15332667.2019.1589241

21 Carlana, Michela. 2019. “Implicit Stereotypes: Evidence from Teachers’ Gender Bias*.” *The Quarterly Journal of Economics* 134 (3): 1163–1224. https://doi.org/10.1093/qje/qjz008

22 Chapman, Derek S., and Anna F. Gödöllei. 2017. “E-Recruiting.” In *The Wiley Blackwell Handbook of the Psychology of the Internet at Work*, 211–30. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119256151.ch11

23 Charlwood, Andy, and Nigel Guenole. 2022. “Can HR Adapt to the Paradoxes of Artificial Intelligence?” *Human Resource Management Journal* n/a (n/a). https://doi.org/10.1111/1748-8583.12433

24 Chen, Le, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. “Investigating the Impact of Gender on Rank in Resume Search Engines.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. CHI ’18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3173574.3174225

25 Crockford, Kade. 2020. “How Is Face Recognition Surveillance Technology Racist?” *American Civil Liberties Union* (blog). June 16, 2020. https://www.aclu.org/news/privacy-technology/how-is-face-recognition-surveillance-technology-racist

26 Deloitte. 2022. “State of AI in the Enterprise 2022.” Deloitte. https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-ai-2022.html

27 Derous, Eva, and Filip De Fruyt. 2016. “Developments in Recruitment and Selection Research.” *International Journal of Selection and Assessment* 24: 1–3. https://doi.org/10.1111/ijsa.12123

28 Derous, Eva, and Ann Marie Ryan. 2019. “When Your Resume Is (Not) Turning You down: Modelling Ethnic Bias in Resume Screening.” *Human Resource Management Journal* 29 (2): 113–30. https://doi.org/10.1111/1748-8583.12217

29 Dickey, Megan Rose. 2021. “Examining the ‘Pipeline Problem’ | TechCrunch.” February 14, 2021. https://techcrunch.com/2021/02/14/examining-the-pipeline-problem/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABphVHeZI3Z28qKWF3Fste5tTPdOfdwsoC0HQBIdbUizE24yJ71BE3rvZoW_vTzK-6ql8MDKutA6rDDDjW0jr1V-8gRgaR5VaQgYCFkt9wHJKc1G7bDGRtjxVBZrU0I8Y9MUu8aE_1_xoPsi22me6RdscsdtRtN8-YL6_NtkJM9-

30 Drage, Eleanor, and Kerry Mackereth. 2022. “Does AI Debias Recruitment? Race, Gender, and AI’s ‘Eradication of Difference.’” *Philosophy & Technology* 35 (4): 89. https://doi.org/10.1007/s13347-022-00543-1

31 Fiscella, Kevin, and Allen M. Fremont. 2006. “Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity.” *Health Services Research* 41 (4 Pt 1): 1482–1500. https://doi.org/10.1111/j.1475-6773.2006.00551.x

32 Gomez-Herrera, Estrella, and Sabine Koeszegi. 2022. “A Gender Perspective on Artificial Intelligence and Jobs: The Vicious Cycle of Digital Inequality.” *Bruegel | The Brussels-Based Economic Think Tank*. https://www.bruegel.org/working-paper/gender-perspective-artificial-intelligence-and-jobs-vicious-cycle-digital-inequality

33 Graham, Logan, Abigail Gilbert, Joshua Simons, Anna Thomas, and Helen Mountfield. 2020. “Artificial Intelligence in Hiring: Assessing Impacts on Equality -.” Institute for the Future of Work. https://www.ifow.org/publications/artificial-intelligence-in-hiring-assessing-impacts-on-equality

34 Guo, Feng, Christopher Gallagher, Tianjun Sun, Saba Tavoosi, and Hanyi Min. 2021. “Smarter People Analytics with Organizational Text Data: Demonstrations Using Classic and Advanced NLP Models.” *Human Resource Management Journal*, December. https://doi.org/10.1111/1748-8583.12426

35 Hanson, Sandra L., and Małgorzata Krywult-Albańska. 2020. “Gender and Access to STEM Education and Occupations in a Cross-National Context with a Focus on Poland.” *International Journal of Science Education* 42 (6): 882–905. https://doi.org/10.1080/09500693.2020.1737341

36 Heilman, Madeline E., Caryn J. Block, and Peter Stathatos. 1997. “The Affirmative Action Stigma of Incompetence: Effects of Performance Information Ambiguity.” *The Academy of Management Journal* 40 (3): 603–25. https://doi.org/10.2307/257055

37 Hicks, Mar. 2017. Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing. 1st edition. Cambridge, MA: MIT Press

38 Hill, Catherine, Christianne Corbett, and Andresse St. Rose. 2010. “Why So Few? Women in Science, Technology, Engineering, and Mathematics.” *American Association of University Women*. American Association of University Women. https://eric.ed.gov/?id=ED509653

39 Hu, James. 2019. “Report: 99% of Fortune 500 Companies Use Applicant Tracking Systems.” Jobscan. November 7, 2019. https://www.jobscan.co/blog/99-percent-fortune-500-ats/

40 IBM. n.d. “What Is Natural Language Processing?” Accessed February 14, 2023. https://www.ibm.com/topics/natural-language-processing

41 Kang, Sonia K., Katherine A. DeCelles, András Tilcsik, and Sora Jun. 2016. “Whitened Résumés: Race and Self-Presentation in the Labor Market.” *Administrative Science Quarterly* 61 (3): 469–502. https://doi.org/10.1177/0001839216639577

42 Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. “Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods.” In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW ’16 Companion*, 53–54. https://doi.org/10.1145/2872518.2889385

43 Koivunen, Sami, Saara Ala-Luopa, Thomas Olsson, and Arja Haapakorpi. 2022. “The March of Chatbots into Recruitment: Recruiters’ Experiences, Expectations, and Design Opportunities.” *Computer Supported Cooperative Work (CSCW)* 31 (3): 487–516. https://doi.org/10.1007/s10606-022-09429-4

44 Lavy, Victor, and Edith Sand. 2015. “On The Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers’ Stereotypical Biases.” w20909. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w20909

45 Li, King King, Lunzheng Li, Wei Si, and Zhibo Xu. 2022. “Childbearing Age and Gender Discrimination in Hiring Decisions: A Large-Scale Field Experiment.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4199754

46 Li, Lan, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. “Algorithmic Hiring in Practice: Recruiter and HR Professional’s Perspectives on AI Use in Hiring.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–76. Virtual Event USA: ACM. https://doi.org/10.1145/3461702.3462531

47 Magua, Wairimu, Xiaojin Zhu, Anupama Bhattacharya, Amarette Filut, Aaron Potvien, Renee Leatherberry, You-Geon Lee, et al. 2017. “Are Female Applicants Disadvantaged in National Institutes of Health Peer Review? Combining Algorithmic Text Mining and Qualitative Methods to Detect Evaluative Differences in R01 Reviewers’ Critiques.” *Journal of Women’s Health (2002)* 26 (5): 560–70. https://doi.org/10.1089/jwh.2016.6021

48 Manatal. 2022. “The Role of AI in Recruitment ATS.” 2022. https://www.manatal.com/blog/role-ai-recruitment

49 Mihajlovic, Ilija. 2021. “Everything You Ever Wanted To Know About Computer Vision. Here’s A Look Why It’s So Awesome.” Medium. September 24, 2021. https://towardsdatascience.com/everything-you-ever-wanted-to-know-about-computer-vision-heres-a-look-why-it-s-so-awesome-e8a58dfb641e

50 Moss, Emanuel, and Jacob Metcalf. 2020. “Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies.” Report. Data & Society Research Institute. https://apo.org.au/node/308440

51 Najibi, Alex. 2020. “Racial Discrimination in Face Recognition Technology.” Harvard University. *Science in the News* (blog). October 24, 2020. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

52 Nawaz, Nishad, and Anjali Mary Gomes. 2019. “Artificial Intelligence Chatbots Are New Recruiters.” *International Journal of Advanced Computer Science and Applications (IJACSA)* 10 (9). https://doi.org/10.14569/IJACSA.2019.0100901

53 Okolie, Ugo Chuks, and Ikechukwu Emmanuel Irabor. 2017. “E-Recruitment: Practices, Opportunities and Challenges.” *European Journal of Business and Management* 9 (11): 116–22.

54 O’Neil, Cathy. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. 1st edition. New York: Crown.

55 Perkowitz, Sidney. 2021. “The Bias in the Machine: Facial Recognition Technology and Racial Disparities.” *MIT Case Studies in Social and Ethical Responsibilities of Computing*, no. Winter 2021 (February). https://doi.org/10.21428/2c646de5.62272586

56 Piketty, Thomas. 2014. *Capital in the Twenty First Century*. Translated by Arthur Goldhammer. Cambridge Massachusetts: Belknap Press: An Imprint of Harvard University Press.

57 Pisanelli, Elena. 2022. “A New Turning Point for Women: Artificial Intelligence as a Tool for Reducing Gender Discrimination in Hiring.” SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4254965

58 Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. “Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices.” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–81. FAT* ’20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

59 Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. “Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices.” *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1): 7:1-7:23. https://doi.org/10.1145/3449081

60 Rankin, Joy Lisi. 2022. “Misogyny and the Making of the Tech Fratriarchy.” *JCMS: Journal of Cinema and Media Studies* 61 (4): 175–80. https://doi.org/10.1353/cj.2022.0054

61 Recruiter.com. n.d. “An Introduction to NLP and How It Is Transforming Recruitment.” Recruiter.Com. Accessed February 14, 2023. https://www.recruiter.com/recruiting/an-introduction-to-nlp-and-how-it-is-transforming-recruitment/

62 Rhea, Alene, Kelsey Markey, Lauren D’Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, and Julia Stoyanovich. 2022. “Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit.” In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 572–87. AIES ’22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3514094.3534189

63 Riegle-Crumb, Catherine, Barbara King, Eric Grodsky, and Chandra Muller. 2012. “The More Things Change, the More They Stay the Same? Prior Achievement Fails to Explain Gender Inequality in Entry Into STEM College Majors Over Time.” *American Educational Research Journal* 49 (6): 1048–73. https://doi.org/10.3102/0002831211435229

64 Savage, Mike. 2021. The Return of Inequality: Social Change and the Weight of the Past. Cambridge, MA: Harvard University Press.

65 Schellmann, Hilke. 2022. “Finding It Hard to Get a New Job? Robot Recruiters Might Be to Blame.” *The Guardian*, May 11, 2022, sec. US news. https://www.theguardian.com/us-news/2022/may/11/artitifical-intelligence-job-applications-screen-robot-recruiters

66 Seo, Eunjin, Yishan Shen, and Edna C. Alfaro. 2019. “Adolescents’ Beliefs about Math Ability and Their Relations to STEM Career Attainment: Joint Consideration of Race/Ethnicity and Gender.” *Journal of Youth and Adolescence* 48 (2): 306–25. https://doi.org/10.1007/s10964-018-0911-9

67 Shein, Esther. 2023. “Semiconductor Industry’s Growing Talent Shortage: How to Recruit Skilled STEM Talent.” TechRepublic. January 30, 2023. https://www.techrepublic.com/article/semiconductor-industry-talent-shortage-how-recruit-skilled-stem-talent/

68 Shove, Elizabeth, Mika Pantzar, and Matt Watson. 2012. *The Dynamics of Social Practice: Everyday Life and How It Changes*. SAGE Publications.

69 Singh, Parbudyal, and Dale Finn. 2003. "The Effects of Information Technology on Recruitment." *Journal of Labor Research* 24 (3): 395–408. https://doi.org/10.1007/s12122-003-1003-4

70 Sloane, Mona. 2019. "Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice." In *Proceedings of the Weizenbaum Conference 2019 "Challenges of Digital Inequality - Digital Education, Digital Work, Digital Life,"* 9. Berlin. https://doi.org/10.34669/wi.cp/2.9

71 Sloane, Mona, and Emanuel Moss. 2022. "Introducing a Practice-Based Compliance Framework for Addressing New Regulatory Challenges in the AI Field." SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=4060262

72 Sloane, Mona, Emanuel Moss, and Rumman Chowdhury. 2022. "A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability." *Patterns* 3 (2). https://doi.org/10.1016/j.patter.2021.100425

73 Sloane, Mona, and Janina Zakrzewski. 2022. "German AI Start-Ups and 'AI Ethics': Using A Social Practice Lens for Assessing and Implementing Socio-Technical Innovation." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 935–47. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533156

74 Snyder, Hannah. 2019. "Literature Review as a Research Methodology: An Overview and Guidelines." *Journal of Business Research* 104 (November): 333–39. https://doi.org/10.1016/j.jbusres.2019.07.039

75 Upadhyay, Ashwani Kumar, and Komal Khandelwal. 2018. "Applying Artificial Intelligence: Implications for Recruitment." *Strategic HR Review* 17 (5): 255–58. https://doi.org/10.1108/SHR-07-2018-0051

76 U.S. Bureau of Labor Statistics. 2021. "Employed Persons by Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity." 2021. https://www.bls.gov/cps/cpsaat11.htm

77 US Census Bureau. 2021. "Women Are Nearly Half of U.S. Workforce but Only 27% of STEM Workers." Census.Gov. January 26, 2021. https://www.census.gov/library/stories/2021/01/women-making-gains-in-stem-occupations-but-still-underrepresented.html

78 Varner, Maddy. 2021. "Public Agencies Are Buying Up AI-Driven Hiring Tools and 'Bossware' – The Markup." December 23, 2021. https://themarkup.org/news/2021/12/23/public-agencies-are-buying-up-ai-driven-hiring-tools-and-bossware

79 Vassilopoulou, Joana, Olivia Kyriakidou, Mustafa F. Özbilgin, and Dimitria Groutsis. 2022. "Scientism as Illusio in HR Algorithms: Towards a Framework for Algorithmic Hygiene for Bias Proofing." *Human Resource Management Journal*, January. https://doi.org/10.1111/1748-8583.12430

80 Wong, Geoff, Trish Greenhalgh, Gill Westhorp, Jeanette Buckingham, and Ray Pawson. 2013. "RAMESES Publication Standards: Meta-Narrative Reviews." *BMC Medicine* 11 (1): 20. https://doi.org/10.1186/1741-7015-11-20.

81 Yoder-Himes, Deborah R., Alina Asif, Kaelin Kinney, Tiffany J. Brandt, Rhiannon E. Cecil, Paul R. Himes, Cara Cashon, Rachel M. P. Hopp, and Edna Ross. 2022. "Racial, Skin Tone, and Sex Disparities in Automated Proctoring Software." *Frontiers in Education* 7. https://www.frontiersin.org/articles/10.3389/feduc.2022.881449

82 Yse, Diego Lopez. 2019. "Your Guide to Natural Language Processing (NLP)." Medium. April 30, 2019. https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

83 Zamora, Jennifer. 2017. "Rise of the Chatbots: Finding A Place for Artificial Intelligence in India and US." In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion*, 109–12. IUI '17 Companion. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3030024.3040201

84 Zou, James, and Londa Schiebinger. 2018. "AI Can Be Sexist and Racist — It's Time to Make It Fair." *Nature* 559 (7714): 324–26. https://doi.org/10.1038/d41586-018-05707-8

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

DIGITAL ACCOUNTABILITY

THE UNTAPPED POTENTIAL OF PARTICIPATION WHEN USING TECHNOLOGY IN HUMANITARIAN ACTION

Düchting, Andrea
Centre for Humanitarian Action
Berlin, Germany
andrea.duechting@fellow.chaberlin.org

KEYWORDS

digitalisation; digital transformation; accountability; participation; humanitarian

DOI: 10.34669/wi.cp/5.4

1 INTRODUCTION

Over the past decades, digital technologies have seen a massive increase in use and have profoundly shaped the humanitarian sector. Their exponential growth has greatly increased the amount of data to be managed and accelerated the speed with which information travels (ALNAP 2022; OCHA 2021). This growth triggered discussions around the efficiency of necessary humanitarian services to respond to rising needs and sector-wide funding cuts. The request for more evidence-based programming, improved coordination, and increased accountability pushed many humanitarian organisations to 'go digital'.

The COVID-19 pandemic, Venezuelan migration crisis and ongoing Ukraine response especially influenced the way humanitarian organisations digitalise. At the same time, questions were raised on how best to communicate with local actors and affected populations, including people on the move. The use of digital tools like mobile messaging apps, social media or AI-based solutions are increasingly discussed to leverage the potential of more effective aid delivery whilst seamlessly enhancing accountability and doing no digital harm.

Despite its potential, humanitarian practices look different. Existing opportunities to engage people and make their voices heard through real-time information sharing and two-way communication are hardly recognised. Various forms of humanitarian feedback mechanisms have been introduced to increase accountability but reality shows that they are mostly used for one-way information sharing or showing impact to donors (ALNAP 2022; Ground Truth Solutions et al. 2022; Owl Re 2022; CHS Alliance et al. 2015)

In sum, massive amounts of personal and non-personal data are collected for various purposes, often without systematically using the data and considering long-term aspects related to data management and governance. To avoid doing harm, humanitarian organisations focus more on data security than processes to inform affected people about the full usage of their data and their rights. Yet, some organisations started to raise questions about people's involvement in data and technology-related decision making and to discuss approaches to co-creating with affected people and building data agency. The debate about digital accountability in the humanitarian sector, however, remains limited and is only picking up slowly (Cieslik et al. 2022; Currion 2022; et al. 2022; Vinck et al. 2022; Ada Lovelace Institute 2021; Hilhorst et al. 2021; OCHA 2021; Madianou 2019; van Solinge 2019; Jacobsen et al. 2018; Madianou et al. 2016).

Humanitarian practitioners are generally aware of the need to improve transparency when processing people's data and the need for an honest discussion about power dynamics in an offline and online sphere. They usually differentiate between legally imposed data subject rights and rights-based approaches that allow tackling programme quality issues and power imbalances. Organisations might be committed to put people at the centre of (data-related) decision-making and (digital) programme design but without systematically embedding digital technologies in overall organisational processes. For this to happen, change and mindset shifts are needed in addition to political willingness. Raising awareness on digital accountability, building digital capacities and capabilities have the potential to avoid doing digital harms while increasing people's data agency (Cieslik et al. 2022; Schächtele et al. 2022; Vinck et al. 2022; Ada Lovelace Institute 2021; Bryant 2021; CDAC Network 2021; OCHA 2021; Williamson 2020; Madianou 2019; Jacobsen et al. 2018; Greenwood et al. 2017; Madianou et al. 2016; Sandvik et al. 2014).

The paper examines the tension between digital technologies, participation, and accountability by exploring their interlinkages, benefits, and challenges. It analyses the ways in which humanitarian actors hold themselves responsible and accountable when using digital technologies and shows ways in which affected people can hold organisations to account.

2 METHOD IN BRIEF

The paper is based on two main questions:

(1) How do humanitarian organisations use digital technology to strengthen the participation of and accountability to affected populations?

(2) How do humanitarian organisations hold themselves responsible when using digital technologies? In turn, how can affected people hold organisations accountable when using digital technologies?

The literature review comprised documents like academic papers, operational reports, guidance notes, strategies, and webpages about the use of technology in humanitarian action, digital transformation, and accountability. Several in-group discussions following Chatham House Rule served in shaping and validating the research.

In addition, 22 qualitative, dialogue interviews were conducted with diverse humanitarian stakeholders. Due to the sensitivity of the topic, the interviews were not recorded and no interviewees are cited. Interview memos were drafted and shared with interviewees for their reference and potential rectification. The memos were structured and analysed as per the following sections: Reasons for going digital, digital transformation, digital accountability, challenges, and vision.

The paper was further informed by two conceptual frameworks: Arnstein's Ladder of Participation (1969) and the Core Humanitarian Standard on Quality and Accountability (2015). The Ladder of Participation originates from the discussion about increasing citizen participation and describes eight levels and three categories of participation including non-participation, tokenisms and citizen power (Arnstein 1969).

The Core Humanitarian Standard (CHS) is a core value in the humanitarian sector encompassing Nine Commitments aiming at a principled, accountable and high-quality support for future system change (ALNAP 2022; Hilhorst et al. 2021; CHS Alliance et al. 2015). The CHS is used as the main accountability framework for measuring quality and effectiveness of humanitarian action by putting affected people at the centre. The research particularly focused on Commitments Four and Five (CHS Alliance et al. 2015).

3 DIGITAL PARTICIPATION AND ACCOUNTABILITY IN HUMANITARIAN ACTION

Humanitarian organisations use digital technologies to engage people and increase accountability by sharing real-time information and asking people's feedback. Interactive tools like social media and mobile messaging apps alongside digital tools for managing feedback data continue to be on the rise. According to Lough, "social media is likely to play an increasingly prominent role for affected people in current and future crises [and] it is not a phenomenon humanitarian actors can continue to sidestep" (Lough 2022, 7). Lough proved that digital communication tools are particularly important to

people on the move, no matter if across borders or within countries, and used for news consumption and communication with family and friends but hardly with humanitarians. Latest studies further reiterate the need for an in-depths analysis of different types of technologies aiming at better understanding their opportunities, trade-offs and risks, and fostering digital inclusion and accountability to those already marginalised and left behind (Lough 2022; CDAC Network 2021; Bryant 2021; OCHA 2021; Madianou 2019; Madianou et al. 2016; Sandvik et al. 2014).

3.1 Digital Technologies for Information Sharing and Collecting Feedback

In addition to traditional forms of participation and collecting feedback, humanitarian organisations currently apply a mix of digital and non-digital approaches to inform people, collect their feedback and ask about their satisfaction. These vary from helpdesks, suggestion boxes to toll-free hotlines, mobile messaging apps, social media and AI-based solutions like chat- and voicebots.

In comparison to most offline approaches, digital tools are mainly used for one-way information sharing and sometimes for rumours tracking but hardly considered to actively consult affected people. Many humanitarian organisations prefer technologies like hotlines, IVR or SMS but hesitate to use mobile messaging apps, social media, not to speak about bots. When applied, they are hardly considered as two-way communication channels and mainly used for limited purposes like information campaigns or sharing programme updates.

The focus on one-way communication channels is mainly due to data protection and privacy concerns, resource constraints as well as unclear roles and responsibilities amongst different teams. Its limitation to one-way information-sharing thus reinforces a tokenistic involvement of people and leaves humanitarian organisations to continue using traditional ways for two-way communication, assuming this preference is mutually shared by affected people.

Interviewees further referred to a cultural and ethical divide when using mobile messaging apps and social media platforms. They are mostly opted against and restricted for data protection reasons, even though affected people might have chosen those tools as preferred option to receive information and communicate with humanitarian stakeholders. Instead, numerous in-house tools are developed and introduced to affected people without necessarily diversifying communication channels and fostering digital inclusion of diverse population segments (e.g. youth, persons with disabilities).

3.2 Legal and Social Accountability in Humanitarian Action

The ethical dimension of accountability and using technologies for accountability purposes include questions on how technologies are best applied when impacting those whose data is processed and who are meant to benefit from using such technologies. It goes beyond giving account (e.g. informing affected people about the technology and data processing activities) and taking account (e.g. collecting feedback, involving affected people in design and decision-making). It raises questions around responsibility, transparency, and ownership. In other words, digital technologies can contain new accountability needs but also reveal important accountability gaps (Hilhorst et al. 2021; Jacobsen et al. 2018).

Pizzi et al. differentiate between social and legal accountability from a technology perspective: “Social accountability requires that the public have been made aware of [the] systems and have adequate digital literacy to understand their impact. Legal accountability requires having legislative and

regulatory structures in place to hold those responsible for bad outcomes to account" (Pizzi et al. 2020, 173). Following this logic, affected people need to be made aware and capacitated to understand the impact of using technologies and take informed decisions. Regulatory frameworks like data protection are one aspect to increase organisational responsibility and data subject rights, humanitarian principles and human rights frameworks further highlight the need for a principled approach.

Due to its nature, legal accountability is mostly considered as a 'must have' by humanitarian organisations. It mainly refers to compliance aspects that are requested by donors and decision-makers, including data protection and privacy regimes like GDPR, national legislation in addition to organisational policies. While non-governmental organisations are bound to such laws, international organisations are generally exempted and follow best industry standards instead.

Most of the interviewees considered the collection of consent as a good practice to increase legal accountability while acknowledging that people are hardly made aware about the full scope of technology and their rights, thus questioning the consent to be really meaningful. While consent is an important cornerstone of data protection and data governance, "it is increasingly viewed as insufficient on its own to foster accountability" (Global Partnership for Sustainable Development Data 2022, 36f). For the consent to be meaningful, affected people need to understand the purpose for using the technology and their data-related rights. Instead, power asymmetries and digital literacy levels influence affected people's decision to share or not to share their personal data in return of assistance (i.e. data for aid) (Veron 2022; Bryant 2021; Holloway, Al Masri, and Abu Yahia 2021; ICRC and Brussels Privacy Hub 2020; Greenwood et al. 2017).

Social accountability is more characterised as an ethical question and a 'nice to have' (e.g. digital literacy, data agency, design justice). While digital transformation and the use of technology is generally driven by efficiency, many organisations do have an aspiration to address long-standing power asymmetries with digital technology. The importance of trust and trustful relationships were repeatedly mentioned. Informing people about programme design, technologies, and data as well as their right to express their opinion and raise complaints are the very basic for creating trustful relationships, in digital and non-digital sphere (Bryant 2022; Martin et al. 2022; Ground Truth Solutions and OCHA 2022; Owl Re 2022; Barbelet, Bryant, and Willitts-King 2020; Madianou et al. 2016).

Beyond audit requirements, humanitarian organisations approach accountability from programme quality lenses with limited leverage to change digitalisation processes or the use of technology at organisational level. When talking about digital technologies in accountability, it seems like humanitarian actors need to start talking about the transformative bit of digitalisation leading to system change and a debate about accountability 2.0, as one of the interviewees called it.

3.2 Case Study: Humanitarian Organisations in Ukraine's Digital Ecosystem

To better understand the linkages between digital technologies and accountability, the ongoing humanitarian crisis in Ukraine is particularly interesting to look at. Humanitarian organisations are part of a functioning ecosystem with a civil society which, in comparison to many other humanitarian and migration crises, is digitally literate and knowledgeable about their data rights. The fact that people are digitally connected and used to digital services pushed many humanitarian organisations to its limits. New ways of informing and communicating with affected people through chat- and voicebots are explored but sceptically viewed by humanitarians who are not used to work in such digitised

environments. The ongoing crisis, thus, feels like a reality check for the humanitarian system and raises operational as well as ethical questions around digital transformation and communication technology for increasing digital participation and accountability (Calp Network 2022; Ground Truth Solutions et al. 2022; Grunewald 2022; Humanitarian Outcomes 2022).

The experience of Ukraine showcases the importance of digital literacy in claiming data rights and taking informed decisions. When people understand their rights, they are in a position to raise concerns and ask about their data. Some interviewees shared the experience of people claiming their data to be updated or erased but difficult to respond to as most organisations missed the relevant processes like transparent data flows to track down all data points.

In addition, hotlines and data systems were set-up fast but humanitarian organisations were soon overwhelmed with the sheer number of incoming calls and requests. Interviewees confirmed that feedback mechanisms, in theory, could be used for claiming data subject rights but were rarely used. In Ukraine, people did raise data concerns and many organisations had to realise that their systems were not fit for purpose. While processes and systems are legally compliant, they fail the operational reality check. This raises practical as well as ethical questions around the humanitarian system's ability and willingness around digital accountability and its operationalisation.

The humanitarian crisis in Ukraine is an interesting example to question current digital accountability practices and highlight the importance of digital literacy and people-centred approaches. When affected people are used to navigating digital tools and claiming their data rights, they do not question but demand digital services and hold organisations to account.

4 KEY FINDINGS

The analysis confirmed that digital technologies are indeed a viable option to strengthen the participation of and accountability to affected people. To leverage the full potential, technologies however need to be embedded in long-term transformation processes aiming at people's increased decision-making or 'citizen power' as it is called in Arnstein's Ladder of Participation. It is not only a matter of using digital tools for specific business processes but integrating technology in systematic ways that trigger mindset shifts and system change.

While affected people worldwide use digital tools to communicate with each other, this is not the case with humanitarian actors. When choosing digital tools as preferred communication channel, people's choices often conflict with data protection and privacy concerns challenging organisations to fulfil their full commitment to respect people's preferences for participating in humanitarian response and sharing feedback. Humanitarian organisations hence prioritise potential risks over actual benefits.

Digital tools are mainly used for sharing information and only few organisations apply digital tools for two-way communication with affected people. Resource constraints, privacy concerns, and political willingness are the main bottlenecks to exploring new ways of engaging affected people in a virtual space, leaving trade-offs like misinformation and disinformation widely unnoticed. The tension of tokenistic activities versus decision-making power further increases when digital technologies come into play as digital transformation adds another layer of complexity to longstanding power relations on the one hand and the dilemma of replicating offline problems to an online environment on the other.

While digital technologies in humanitarian action have the potential to contain new accountability needs, they also reveal important accountability gaps. Legal accountability is primarily associated with compliance requirements and collecting meaningful consent, and social accountability is still in its infancy and yet to be explored. Affected people are rarely consulted in technology-related decision-making and remain stuck at the tokenistic level of information sharing and consultation.

As per Arnstein's Ladder of Participation, empowering people refers to trustworthy partnerships and 'citizen control' reflected in data agency and stewardship concepts which are yet to be explored and introduced to accountability standards like the CHS. The humanitarian crisis in Ukraine highlights the importance of digital literacy for people to digitally engage and control their data. New approaches need to be considered to increase digital accountability alongside people-centred approaches in technology choices and a whole-system approach to raising awareness about new digital responsibilities. Simple answers are needed to address complex issues and the dilemma of increasingly replicating offline challenges into an online environment.

ACKNOWLEDGEMENTS

The research paper is part of the Data and Digitalisation Project aiming at increasing digital literacy of international humanitarian actors. The research and the author's involvement were funded by the German Federal Foreign Office.

REFERENCES

1. Ada Lovelace Institute. 2021. 'Participatory Data Stewardship. A Framework for Involving People in the Use of Data'. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/.
2. ALNAP. 2022. 'The State of the Humanitarian System'. London: ALNAP/ODI. sohs.alnap.org.
3. Arnstein, Sherry R. 1969. 'A Ladder Of Citizen Participation'. *Journal of the American Institute of Planners* 35 (4): 216–24. https://doi.org/10.1080/01944366908977225.
4. Barbelet, Veronique, John Bryant, and Barnaby Willitts-King. 2020. '"All Eyes Are on Local Actors": Covid-19 and Local Humanitarian Action'. HPG (ODI).
5. Bryant, John. 2021. 'Digital Mapping and Inclusion in Humanitarian Response'. Working Paper. HPG Working Paper. London: ODI. https://odi.org/en/publications/digital-mapping-and-inclusion-in- humanitarian-response.
6. ———. 2022. 'Digital Technologies and Inclusion in Humanitarian Response'. HPG Report. London: ODI. https://cdn.odi.org/media/documents/Digital_inclusion_synthesis.pdf.
7. Calp Network. 2022. 'Registration, Targeting and Deduplication: Emergency Response inside Ukraine'. Thematic Paper. Calp Network. https://www.calpnetwork.org/publication/registration-targeting-and-deduplication-emergency-response-inside-ukraine-thematic-paper/?utm_source=CALP+Network+Master+List&utm_campaign=94dd42b618-EMAIL_CAMPAIGN_2022_10_12_10_56&utm_medium=email&utm_term=0_debc722091-94dd42b618-372359747.
8. CDAC Network. 2021. 'Peer Pressure: How Deepening Digital Access Is Transforming Communication as Aid'. Webinar presented at the HNPW 2021, Online. https://www.youtube.com/watch?v=jn1b5NbUzpU.
9. CHS Alliance, The Sphere Project, and Groupe URD. 2015. 'CHS Guidance Notes and Indicators'. Guidance. https://corehumanitarianstandard.org.
10. Cieslik, Katarzyna, and Dániel Margócsy. 2022. 'Datafication, Power and Control in Development: A Historical Perspective on the Perils and Longevity of Data'. *Progress in Development Studies* 22 (4): 352–73. https://doi.org/10.1177/14649934221076580.
11. Düchting, Andrea. 2023. Digital Accountability: The Untapped Potential of Participation when Using Digital Technology in Humanitarian Action. Berlin: Centre for Humanitarian Action. https://www.chaberlin.org/en/publications/digital-accountability-2/.
12. Greenwood, Faine, Caitlin Howarth, Danielle Escudero Poole, Nathaniel A. Raymond, and Daniel P. Scarnecchia. 2017. 'The Signal Code: A Human Rights Approach to Information During Crisis'. 1. Harvard Humanitarian Initiative. https://hhi.harvard.edu/publications/signal-code-human-rights-approach-information-during-crisis.
13. Ground Truth Solutions, and OCHA. 2022. 'Listening Is Not Enough. People Demand Transformational Change in Humanitarian Assistance. Global Analysis Report'. Analysis. Ground Truth Solutions. https://groundtruthsolutions.org/wp-content/uploads/2022/12/GTS_Global-Analysis_November-2022_website.pdf.
14. Grunewald, Francois. 2022. 'Real Time Evaluation of the Humanitarian Response to the Crisis Resulting from the War in Ukraine. July 24th - August 18th, 2022'. Real Time Evaluation. Groupe URD. https://ukraineresponse.alnap.org/help-library/real-time-evaluation-of-the-humanitarian-response-to-the-crisis-resulting-from-the-war.
15. Hilhorst, Dorothea, Samantha Melis, Rodrigo Mena, and Roanne van Voorst. 2021. 'Accountability in Humanitarian Action'. *Refugee Survey Quarterly* 40 (4): 363–89. https://doi.org/10.1093/rsq/hdab015.
16. Holloway, Kerrie, Reem Al Masri, and Afnan Abu Yahia. 2021. 'Digital Identity, Biometrics and Inclusion in Humanitarian Response to Refugee Crises'. HPG Working Paper. London: ODI. https://cdn.odi.org/media/documents/Digital_IP_Biometrics_case_study_web.pdf.

17. Humanitarian Outcomes. 2022. ‘Enabling Local Response: Emerging Humanitarian Priorities in Ukraine. March-May 2022’. Rapid Review. https://www.humanitarianoutcomes.org/projects/Ukraine-review-2022.
18. ICRC, and Brussels Privacy Hub. 2020. *Handbook on Data Protection in Humanitarian Action. 2nd Ed. Edited by Christopher Kuner and Massimo Marelli.* 2. Auflage. Geneva, Brussels: , Brussels Privacy Hub. https://www.icrc.org/en/data-protection-humanitarian-action-handbook.
19. Jacobsen, Katja Lindskov, and Kristin Bergtora Sandvik. 2018. ‘UNHCR and the Pursuit of International Protection: Accountability through Technology?’ *Third World Quarterly* 39 (8): 1508–24. https://doi.org/10.1080/01436597.2018.1432346.
20. Lough, Oliver. 2022. ‘Social Media and Inclusion in Humanitarian Response’. Working Paper. HPG Working Paper. London: ODI. https://odi.org/en/publications/social-media-and-inclusion-in- humanitarian-response.
21. Madianou, Mirca. 2019. ‘Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises’. *Social Media + Society* 5 (3): 205630511986314. https://doi.org/10.1177/2056305119863146.
22. Madianou, Mirca, Jonathan Corpus Ong, Liezel Longboan, and Jayeel S. Cornelio. 2016. ‘The Appearance of Accountability: Communication Technologies and Power Asymmetries in Humanitarian Aid and Disaster Recovery’. *Journal of Communication* 66 (6): 960–81. https://doi.org/10.1111/jcom.12258.
23. Martin, Aaron, Gargi Sharma, Siddharth Peter de Souza, Linnet Taylor, Boudewijn van Eerd, Sean Martin McDonald, Massimo Marelli, Margie Cheesman, Stephan Scheel, and Huub Dijstelbloem. 2022. ‘Digitisation and Sovereignty in Humanitarian Space: Technologies, Territories and Tensions’. *Geopolitics*, March, 1–36. https://doi.org/10.1080/14650045.2022.2047468.
24. OCHA. 2021. ‘From Digital Promise to Frontline Practice: New and Emerging Technologies in Humanitarian Action’. https://www.unocha.org/sites/unocha/files/OCHA%20Technology%20Report.pdf.
25. Owl Re. 2022. ‘Humanitarian Accountability Report 2022’. Annual Report. Geneva: CHS Alliance.
26. Pizzi, Michael, Mila Romanoff, and Tim Engelhardt. 2020. ‘AI for Humanitarian Action: Human Rights and Ethics’. *International Review of the Red Cross* 102 (913): 145–80. https://doi.org/10.1017/S1816383121000011.
27. Sandvik, Kristin Bergtora, Maria Gabrielsen Jumbert, John Karlsrud, and Mareile Kaufmann. 2014. ‘Humanitarian Technology: A Critical Research Agenda’. *International Review of the Red Cross* 96 (893): 219–42. https://doi.org/10.1017/S1816383114000344.
28. Schächtele, Kai, Ingo Dachwitz, Felix Zimmermann, Chris Köver, Christine Meissler, Martina Hahn, and Sven Hilbig. 2022. ‘Atlas der Zivilgesellschaft: Freiheitsrechte unter Druck. Schwerpunkt Digitalisierung’. Berlin: Brot für die Welt. https://www.brot-fuer-die-welt.de/themen/atlas-der-zivilgesellschaft/.
29. Solinge, Delphine van. 2019. ‘Digital Risks for Populations in Armed Conflict: Five Key Gaps the Humanitarian Sector Should Address’. *ICRC Humanitarian Law & Policy* (blog). 12 June 2019. https://blogs.icrc.org/law-and-policy/2019/06/12/digital-risks-populations-armed-conflict-five-key-gaps-humanitarian-sector/.
30. Veron, Pauline. 2022. ‘Digitalisation in Humanitarian Aid: Opportunities and Challenges in Forgotten Crises’. 143. Ecdpm Briefing Note. Brussels: ECDPM. https://ecdpm.org/wp-content/uploads/Digitalisation-humanitarian-aid-ECDPM-Briefing-note-143-2022.pdf.
31. Vinck, Patrick, Emmanuel Letouzé, Tatiana Goetghebuer, Maria Antonia Bravo, Romain Fourmy, Kevin Henkens, Jeroen Peers, and Stephen Matthew. 2022. ‘Strategic Evaluation of WFP’s Use of Technology in Constrained Environments’. Strategic Evaluation OEV/2020/002. Centralized Evaluation Report. World Food Programme (WFP), Aide à la Décision Èconomique (ADE). https://www.wfp.org/publications/strategic-evaluation-wfps-use-technology-constrained-environments.
32. Williamson, Jazmin. 2020. ‘Ensuring Accountability to Affected Populations in Humanitarian Settings: “Holding Humanitarian Organizations Accountable to People.”’ *Independent Study Project (ISP) Collection*, April. https://digitalcollections.sit.edu/isp_collection/3295.

33. Worthington, Robert, and Andrea Düchting. 2023. Enabling Dignified Humanitarian Assistance through Safe Data Sharing. Landscape Mapping. Geneva: International Federation of the Red Cross and Red Crescent Society. https://interoperability.ifrc.org/.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

EXPLORING THE GERMAN-LANGUAGE TWITTERSPHERE

NETWORK ANALYSIS OF DISCUSSIONS ON THE SYRIAN AND UKRAINIAN REFUGEE CRISES

Kiyak, Sercan
KU Leuven
Leuven, Belgium
sercan.kiyak@kuleuven.be

De Coninck, David
KU Leuven
Leuven, Belgium
david.deconinck@kuleuven.be

Mertens, Stefan
KU Leuven
Leuven, Belgium
stefan.mertens@kuleuven.be

d'Haenens, Leen
KU Leuven
Leuven, Belgium
leen.dhaenens@kuleuven.be

KEYWORDS

migration; social network analysis; Twitter; Syria; Ukraine; Germany

DOI: 10.34669/wi.cp/5.5

ABSTRACT

This study conducts a comparative analysis of Twitter communication networks relating to the Syrian and Ukrainian refugee crises. Employing a network analysis approach, the study uses approximately 660,000 tweets to gain insights into the online discussion communities surrounding these crises. Tweets specifically discussing Syrian refugees were collected between 2015 and 2023, while those about Ukrainians were harvested from 2022 to 2023, utilizing the full-archive search endpoint of the Twitter API. By transforming retweets into communication networks between users, the study investigates the community structure within these networks. The findings reveal that the online anti-refugee community is smaller in size, more active, highly interconnected, and transcends national boundaries, in contrast to the opposing communities. These results underscore the need for increased social media engagement of pro-refugee voices and improved moderation practices to foster a more inclusive virtual public sphere.

1 INTRODUCTION

Migration is a central topic in political discussions across European countries, leading to polarisation (Damstra et al., 2021). Right-wing populist politicians exploit the fear surrounding migrants to gain electoral advantage and propose policies to reduce their numbers. In contrast, liberal and progressive politicians prioritise the humane treatment and human rights of migrants, advocating for integrationist policies (Bossetta, 2018). The media played a role in negatively framing Syrian refugees in 2015, particularly in countries with right-leaning governments, leading to a "systemic and persistent" de-individualisation of their image (Georgiou & Zaborowski, 2017, p. 3). People's media preferences, consumption habits, and attitudes towards migrants were found to be connected (De Coninck et al., 2019; Debrael et al., 2021). Furthermore, public opinions on migration vary significantly and are polarised across Europe (d'Haenens et al., 2019). Hence, the dynamic relationship between politicians, media, and public opinion is not fixed and evolves over time and in response to ongoing events (De Coninck et al., 2022).

Utilising social media platforms for analysing latent public opinion about contested topics, such as migration, not only provides valuable insights but also enables the detection of viral sources and the spread of information, contributing to combating misinformation and fostering a more inclusive public discourse. Twitter is frequently used in analysing social media platforms for political communication studies (Seabold et al., 2015; van Klingeren et al., 2021). Twitter data serves as a valuable and extensive information source for computational social science (Verbeke et al., 2017). By detecting viral sources and monitoring information dissemination on online social networks, it becomes possible to combat misinformation (Tambuscio et al., 2018) and to contribute to a more inclusive public discourse (Ahmed et al., 2020). Through collecting trace data from platforms like Twitter and applying computational research methods, we can gain insights into the underlying public opinion and how these opinions are communicated within online networks (Freelon, 2020).

Previous research on refugee crises has primarily focused on examining the impact of textual or visual content on social media platforms (Chouliaraki, Lilie et al., 2017; d'Haenens et al., 2019; McCann et al., 2023; Nerghes & Lee, 2019; Ozerim & Tolay, 2021; Öztürk & Ayvaz, 2018). Some network-based approaches also exist (Institute for Strategic Dialogue, 2021; Nerghes & Lee, 2018; Pöyhtäri et al., 2021). This new study adds a partition-based network analysis of the online political communication about refugees from a comparative perspective to the literature. More in particular, our research investigates the retweet networks related to Syrian and Ukrainian refugees on Twitter (from now on, N1 and N2, respectively).[1] There are several studies of political message networks on Twitter (Stegmeier et al., 2019), their temporal changes (Nasralah et al., 2022) and in different national twitterspheres (Fincham, 2019) and different languages (Smyrnaios & Ratinaud, 2017; Yao et al., 2022)). Prior research on political communication networks indicates that these networks are divided into two main camps (Galeazzi, 2022). However, these camps consist of different user clusters (Freelon 2020). Our first goal is to disclose these clusters in the networks.

RQ1: What are the community structures of retweet networks related to these refugee crises?

[1] We want to highlight here two ethical challenges for our research: 1) In this study we aim at analysing communication networks, however we do not assume or claim that both Syrian and Ukrainian refuge crises are identical events. Both have different cultural and historical dynamics that are beyond the limits of our communication-based research. 2) Although we refer to these events concerning human mobility as 'crises', we do not claim they should be considered as such. The narrative of crises about human mobility can have detrimental effects for social inclusion and communication (Sommer, 2022). We will nevertheless use the term 'crisis' for lack of a better term to refer to these events.

Additionally, the literature on social media analysis reveals that far-right parties and movements effectively use social media platforms to advance their anti-refugee agenda (Åkerlund, 2022; Schroeder, 2018).We ask ourselves whether there any differences between the two networks in this regard.

RQ2: What are the activity and engagement levels of the main pro- and anti-refugee communities[2] *in both networks?*

Finally, we will exploit a feature of our dataset to explain the results of our RQ2. Concretely, we will analyse the national composition of main clusters to investigate the transnationalityof the leading communities in debate (Stoltenberg, 2021). The study of interactions between similar political movements and groups across national borders in the virtual public sphere is facilitated by digital methods (Dahlberg-Grundberg et al., 2016; Merrill & Copsey, 2022). Thanks to our German-language Twitter dataset, we have the geolocation of users from mainly three countries: Germany, Austria, and Switzerland. We investigate if the communication networks are divided along national lines or if there are any cross-national collaborations in network clusters. If so, do they happen equally on each side of the polarisation?

RQ3: What is the national composition of the clusters? Which are more transnational?

2 RESEARCH DESIGN AND DATA

Social network analysis (SNA) is a method that is widely applied for the analysis of polarisation on online social media platforms (Adamic & Glance, 2005; Al Amin et al., 2017; Esteve-Del-Valle, 2022; Feller et al., 2011; Garimella, 2018; Urman, 2020). Unsurprisingly, online discussions about migration are the subject of many SNA studies (Dehghan & Bruns, 2022; Vilella et al., 2020; Yoo, 2019). We adopt a mixed-methods approach to the social networks championed by recent studies in communication sciences (D'angelo et al., 2016; Freelon, 2020; Froehlich et al., 2020, 2020; Yousefi Nooraie et al., 2020). Furthermore, we aimed to conduct SNA for communication science purposes and focused on the partitions and their interrelations (Freelon, 2020; Freelon et al., 2015, 2016). Furthermore, we used descriptive network statistics, visual analysis methods (Jacomy, 2021) and data-driven but reflective analysis of the textual features of user and tweet metadata (Dehghan et al., 2020) to understand the complex social network structures.

We will adopt an exploratory approach to find changes and patterns between these communication networks related to the Ukrainian and Syrian refugee crises. We will focus on the structure of the retweet networks as they signify positive relations between users and approval of message content most unambiguously (Ahn & Park, 2015; Freelon, 2020). Our workflow was as follows: First, we collected the dataset from Twitter API. Second, we applied custom Python scripts to our dataset to clean the data and scrape the location attribute of user data. Third, we constructed networks (graphs) and detected the communities (subgraphs, partitions) and influential users (based on indegree centrality). Fourth, we (qualitatively) labeled the communities based on the description and retweet data. Fifth, we analysed meso-level network structures using 1) community size, 2) the internal ties as an

[2] Given the strong anti-refugee disinformation and propaganda in online social networks, we will refer to any group and tweet that does not explicitly aim at expanding this agenda as pro-refugee in this research. Pro-refugee in this study should not be understood as an ideological position but rather as a non-anti-refugee position.

indicator of retweet activity and connectedness of the community and 3) subgraph average degree to quantify the level of engagement per user of a given cluster. Finally, we exploited the user geolocation data to investigate the national composition of the communities.

Our analysis of retweet networks involved the utilisation of various tools and techniques. Firstly, we constructed the networks by representing users as nodes and retweets as directed and weighted (the number of retweets between users) ties. To conduct partition-based network analysis, we used Python and its Pandas, NetworkX, and TSM packages (Freelon, 2020). Subsequently, we visualised the networks using Gephi, a widely adopted tool for visual network analysis (Bastian et al., 2009). Gephi facilitates the transformation of networks into visual maps, employing force-directed layout algorithms to position related nodes in close proximity (Jacomy et al., 2014). Community detection is a vital part of network analysis of political communication (Münch, 2019). Our study employed the "Louvain" algorithm, renowned for its efficiency (Blondel et al., 2008). The algorithm works based on detecting sets of nodes (users) exhibiting dense interconnections, indicating homophily. To identify influential nodes, we employed the weighted indegree centrality measure.

Considerable deliberation was invested in formulating our tweet dataset construction methodology. We deliberately extended our investigation beyond 2015 for the Syrian Refugee Crisis, considering the enduring impact and ongoing debates surrounding this issue. Therefore, our data collection encompassed the period from 2015 to 2023 for this event. We focused on gathering data from 2022 to 2023 for the Ukrainian case. To ensure the inclusion of relevant tweets concerning each refugee debate, we adopted a filtering method that entailed capturing tweets containing keywords associated with human migration, coupled with ethnic markers such as 'Syrian' or 'Ukrainian,' all in the German language. The query strings used for data collection were made accessible through the GitHub repository of the first author.[3] These refined queries specifically targeted data directly relevant to our study, excluding indirectly related events such as the Syrian Civil War or the invasion of Ukraine.[4] The collected dataset is presented below for reference.

	SYRIAN CASE	UKRAINIAN CASE	TOTAL
TOTAL MESSAGES	318,338	342,634	660,972
RETWEETS	214,683	238,934	453,617
USERS	92,673	101,192	193,865

Table 1. The collected dataset

The dataset is used to generate a retweet network as described below:

	SYRIAN-NETWORK (N1)	UKRAINIAN-NETWORK (N2)
NODES	55,160	66,885
EDGES	213,031	237,378

Table 2. Node and edge sizes of the generated networks

[3] https://github.com/sercankiyak/GermanTwittersphereMigrationSNA

[4] It is important to acknowledge that our commitment to explicit and stringent criteria implies that our dataset does not capture all Twitter communication on the topic. It is possible for users to allude to these groups without explicitly mentioning ethnicity or referring to migrants. In fact, they can do it without any words such as visuals or gestures or emojis. In short, our keywords generated an amalgamated dataset of political communication that is limited in size but highly accurate.

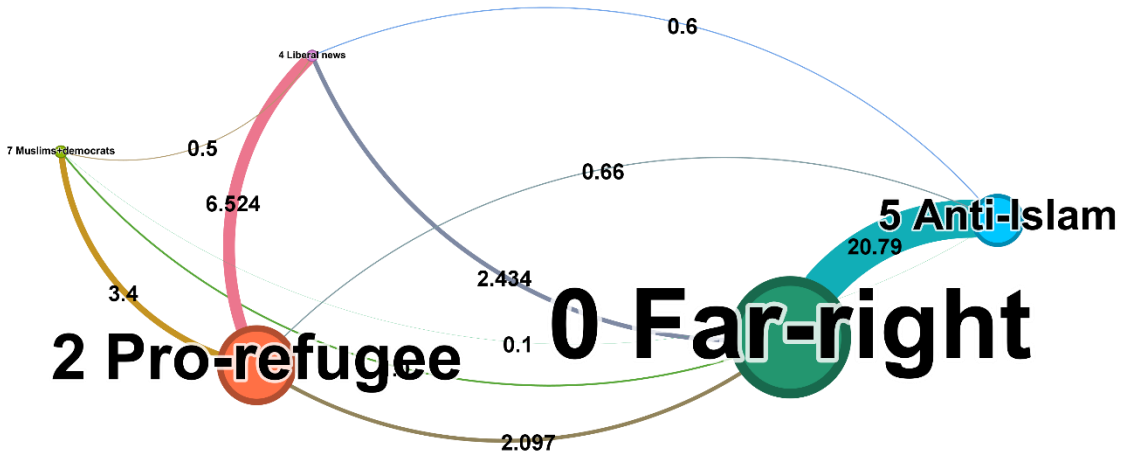
3 RESULTS

Our analysis of N1 resulted in Tables 4 and Graph 1 below.

COMMUNITY	POPULATION	INTERNAL TIES	AVG. DEGREE
PRO-REFUGEE (2)	14,181	36,558	2.58
FAR-RIGHT (0)	10,080	54,025	5.36
ANTI-ISLAM (5)	8,579	25414	2.96
MUSLIMS+DEMOCRATS(7)	7,079	8731	1.23
MEDIA (4)	6,001	8400	1.40
AFD+NEWS (11)	2537	4,207	1.66
SOLIDARITY+NGOS (21)	2026	3371	1.66
INTERNATIONALIST(38)	1863	3369	1.81
ACTIVIST (13)	1416	1461	1.03
MIXED (25)	1398	1809	1.29

Table 4. Size, internal ties and average degree scores for the top 10 detected communities in N1. Community labels consist of two parts: 1) Label assigned by the researchers and 2) the arbitrary number assigned by the algorithm (kept for future reference and convenience). The communities are listed from the largest to smallest in terms of their user population. The internal ties represent the volume of internal retweeting activity. The average degree quantifies the (internal) retweet per user. It shows the average user engagement and promotion of their community by users (irrespective of the community size). The highest numbers in each column are highlighted.

Regarding community sizes and identifying central nodes within N1, our analysis reveals several notable findings. Firstly, the pro-refugee community (2) emerges as the largest group within the network, encompassing pro-refugee NGOs, media outlets, politicians, and their respective supporters. In contrast, community 0 primarily consists of the anti-migrant populist party Alternative für Deutschland (AfD), alongside other nationalist and conservative opinion leaders. Interestingly, the third largest community (5) exhibits a distinct anti-Islam stance that parallels the nationalist community (0). Community 5 also shows a significant average subgraph degree, indicating strong per-user engagement. Unsurprisingly these communities exhibit a similar stance against Syrian refugees. Community 7, conversely, consists of influencers who are Muslims, individuals from migrant backgrounds, and politicians who express empathy toward them. Notably, within the fifth largest community (4), central accounts predominantly belong to liberal or left-wing media entities. After the 5th community, the population size dips significantly (from 6000 to 2500). Consequently, we decided to focus on the top 5 communities in N1. The external ties among communities can be as informative as internal ties. They are visualised below as Graphs 1 and 2.



Graph 1. The community network structure of the top 5 communities in N1. The size of each node, representing a community, is determined by its internal ties rather than the number of nodes within the community. Consequently, larger nodes correspond to communities with higher rates of retweeting among their members. The thickness of the edges connecting nodes indicates the strength of connections between two communities (number of connecting edges divided by all edges).

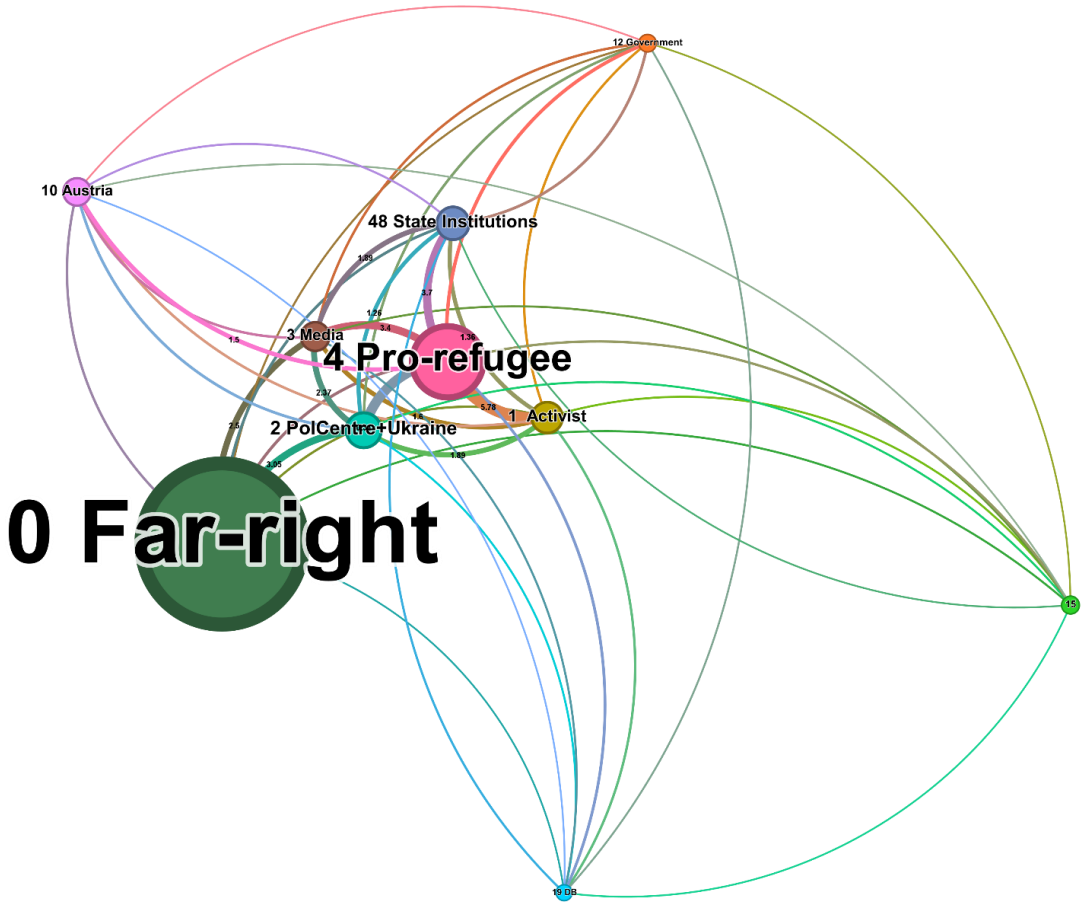
The strong tie between the anti-refugee communities (far-right (0) and anti-Islam (5)) indicate a strong connection and retweet activity between these communities. The retweet activity of pro-refugee communities on Twitter is relatively weak compared to their opposition in the case of Syrian refugees in terms of both internal and external ties. Table 5 and Graph 2 below concern the N2.

COMMUNITY	POPULATION	INTERNAL TIES	AVG. DEGREE
FAR-RIGHT (0)	17,283	85,560	4.95
PRO-REFUGEE (4)	14,703	33,787	2.30
UKRAINIAN+POL.CENTRE(2)	7,560	12,803	1.69
ACTIVIST (1)	7,057	10,522	1.49
GOVERNMENT (48)	5,869	11,433	1.95
MEDIA (3)	4,939	9,159	1.85
AUSTRIAN (10)	3,693	8,483	2.30
GOVERNMENT (12)	2,089	2,803	1.34
SWISS (15)	1,896	2,945	1.55
PUBLIC INSTITUTION (19)	1,796	1,955	1.09

Table 5. Size, internal ties and average degree scores for the top 10 detected communities in N2. See Table 4 above for the explanations.

Table 5 provides insights into the composition of N2, revealing a substantial polarisation between two main opposing groups. The largest group, denoted as community 0, consists of influencers affiliated with AfD politicians, conservative opinion leaders, and their followers. In contrast, community 4 represents the largest pro-refugee group within this network. Community 2 is characterised by pro-Ukrainian accounts featuring CDU parliament members and journalists from conservative-leaning media outlets such as Welt. The fourth biggest community consists of influencer-activists who support Ukrainian refugees. The following smaller communities comprise government, mainstream media, and institutional accounts. Notably, we observe two smaller communities of pro-Ukrainian

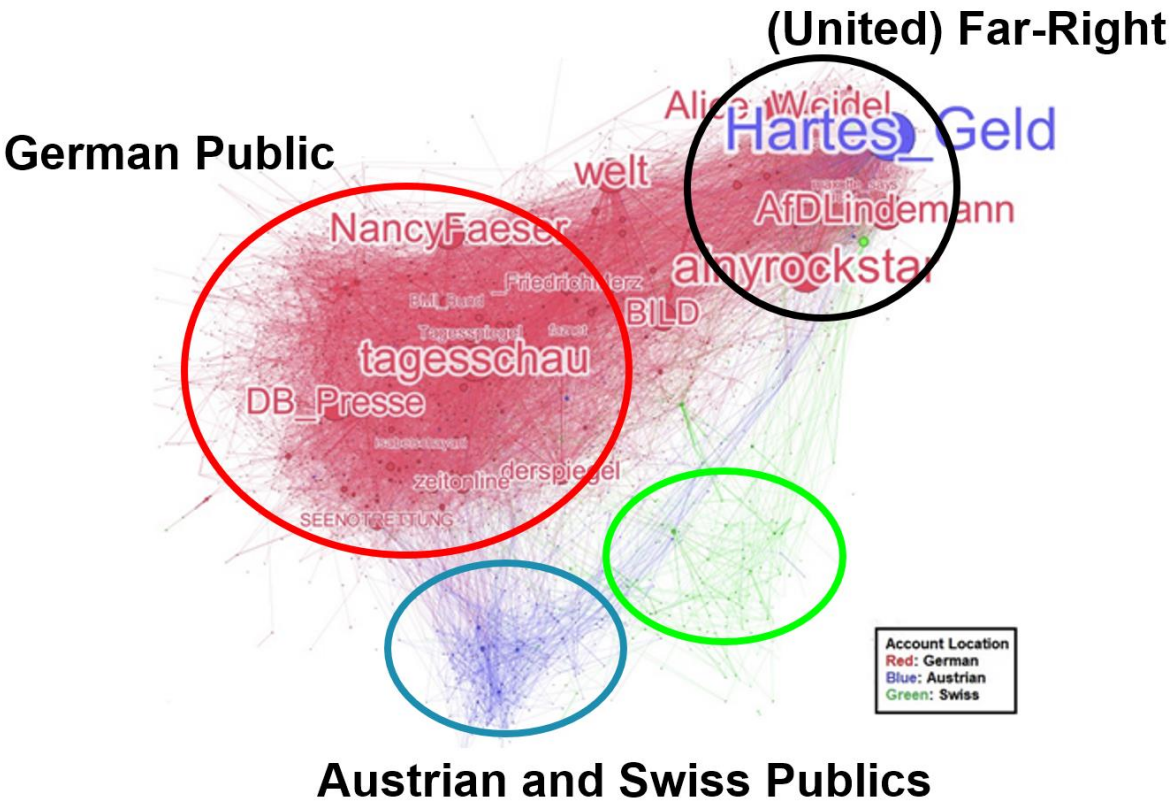
refugee accounts from Austrian (10) and Swiss users (15). However, we do not find anti-refugee Austrian or Swiss communities; we will explain why that might be the case below.



Graph 2. The community network structure of the top 10 communities in N2. See Graph 1 above for the explanations.

The analysis of Graph 2 reveals that the far-right community (0) is the most active one in N2. It is relatively consolidated and highly active in spreading its messages. On the other hand, the pro-refugee communities display smaller node sizes (internal activity) and weaker ties (external retweet connections) compared to N1. This shows sparse connections among the pro-refugee group, which is not helped by the weak connections of Austrian (10) and Swiss (15) to the main German pro-refugee community (4).

These observations answer RQ1 and RQ2. For the former, we found significant community-level retweeting behaviour in both networks and a polarised communication network as described above. Our analysis showed that while the pro-refugee communities are more sparsely connected when amalgamated, they constitute more than half of the users in the network. Conversely, we observed a notably higher level of activity and engagement from the anti-refugee clusters. This result holds particular significance since the anti-refugee communities do not consistently constitute the largest communities. We hypothesise that cross-national collaboration may be one contributing factor to this phenomenon of a highly active anti-refugee community.



Graph 3. A visualisation of N2 coloured by geolocation. Red indicates German, blue indicates Austrian and green indicates Swiss users and their retweets. The circle and labels were added manually to highlight the spatial differences. This graph is limited to users with location data.

	FAR-RIGHT (COM 0)	PRO-REFUGEE (COM 4)
GERMAN %	90	94.6
AUSTRIAN %	5.21	2.79
SWISS %	4.79	2.61

Table 6. Nationality Distribution of the top 2 communities in N2 in percentages.

Graph 3 shows the locations of far-right (0), German pro-refugee (4), Austrian pro-refugee (10) and Swiss pro-refugee (15) communities. Table 6 shows that compared to the main pro-refugee community (0), the anti-refugee community (5) exhibits a more transnational composition. This trend is also expressed in the analysis of its top nodes; while there were no non-German users in the top 50 central users (indegree) in community 0, there were five non-German users among the top 50 far-right users. Alongside their high indegree centrality within the graph, these users included the *hartes_geld*, the node with the highest indegree centrality in the graph, who is from Austria. Therefore, our results indicate that far-right groups engage and benefit from transnational communication and support more than the pro-refugee groups (RQ3).

4 DISCUSSION AND CONCLUSION

The present study has investigated the ongoing online communication networks on Twitter with regard to the Syrian and Ukrainian refugee crises (N1 and N2), focusing on retweets and community structure. The results indicate that the anti-refugee community displays higher activity levels despite not always having the largest numbers. Additionally, while constituting the majority on Twitter, the pro-refugee users are loosely connected with significantly fewer ties between themselves, suggesting less individual engagement and activity on social networks and a weaker community. While N1 showed two anti-refugee clusters (AfD and anti-Islam), in N2, the anti-refugee group is more consolidated, indicating more isolation and growing polarisation in recent years. Finally, in N2, the same community showed more transnational ties compared to the main pro-refugee community. These findings are consistent with previous research that is conducted in different national contexts regarding anti-migration communities being (what wefer to as) “a loud minority" phenomenon (Vilella et al. 2020; Dehghan and Bruns 2022).

Unfortunately, in this study, we had to focus on the explicit and strict criteria for our data, and we could not investigate the tweet contents. Moreover, we did not engage with the temporality of N1 and changes in the Twitter networks. Despite these weaknesses, our research contributes to the study of online communication about migrants by investigating the network structure and diffusion of information on Twitter. Furthermore, it highlights the importance of transnationality for analyzing virtual public discussions. It is a promising direction for future research, and it can help us avoid "methodological nationalism," whose critique highlights the challenges a researcher needs to navigate while studying nations and national public spheres (Wimmer & Schiller, 2003). The digital trace data and social networks of communication open new avenues to use this concept that came out of migration and transnationality studies. Regarding policy suggestions, our findings underscore the necessity of implementing measures to foster improved online discourse surrounding migration. Firstly, it is imperative to enhance moderation efforts aimed at curtailing the dissemination of hateful and misleading content, which could be effectively amplified by far-right factions. Secondly, pro-refugee civil society organisations and public institutions need to enhance their social media presence. In particular, the initiation of transnational campaigns promoting inclusivity and communication through social media channels hold the potential to counteract far-right activism and propaganda.

ACKNOWLEDGEMENTS

This research as part of the OPPORTUNITIES project has received funding from the European Union’s Horizon 2020 Research & Innovation programme under Grant Agreement no. 101004945.

REFERENCES

1. Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. Proceedings of the 3rd International Workshop on Link Discovery, 36–43. https://doi.org/10.1145/1134271.1134277
2. Ahmed, W., Marin-Gomez, X., & Vidal-Alaball, J. (2020). Contextualising the 2019 E-Cigarette Health Scare: Insights from Twitter. International Journal of Environmental Research and Public Health, 17(7), Article 7. https://doi.org/10.3390/ijerph17072236
3. Ahn, H., & Park, J.-H. (2015). The structural effects of sharing function on Twitter networks: Focusing on the retweet function. Journal of Information Science, 41(3), 354–365.

4. Al Amin, M. T., Aggarwal, C., Yao, S., Abdelzaher, T., & Kaplan, L. (2017). Unveiling polarization in social networks: A matrix factorization approach. IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, 1–9. https://doi.org/10.1109/INFOCOM.2017.8056959
5. Åkerlund, M. (2022). *Far right, right here: Interconnections of discourse, platforms, and users in the digital mainstream* [PhD Dissertation, Umeå University]. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-191942
6. Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. Association for the Advancement of Artificial Intelligence.
7. Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10). http://arxiv.org/abs/0803.0476
8. Bossetta, M. (2018). The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. Journalism & Mass Communication Quarterly, 95(2), 471–496.
9. Chouliaraki, Lilie, Georgiou, Myria, Zaborowski, Rafal, & Oomen, W.A. (2017). The European' migration Crisis' and the Media: A Cross-European Press Content Analysis [Project Report]. London School of Economics and Political Science.
10. d'Haenens, L., Joris, W., & Heinderyckx, F. (2019). Images of Immigrants and Refugees in Western Europe: Media Representations, Public Opinion and Refugees' Experiences. Leuven University Press.
11. Dahlberg-Grundberg, M., Lundström, R., & Lindgren, S. (2016). Social media and the transnationalization of mass activism: Twitter and the labour movement. *First Monday*.
12. Damstra, A., Jacobs, L., Boukes, M., & Vliegenthart, R. (2021). The impact of immigration news on anti-immigrant party support: Unpacking agenda-setting and issue ownership effects over time. Journal of Elections, Public Opinion and Parties, 31(1), 97–118.
13. D'angelo, A., Ryan, L., & Tubaro, P. (2016). Visualization in Mixed-Methods Research on Social Networks. Sociological Research Online, 21(2), 148–151.
14. De Coninck, D., Matthijs, K., Debrael, M., Cock, R. D., & d'Haenens, L. (2019). Unpacking Attitudes on Immigrants and Refugees: A Focus on Household Composition and News Media Consumption. Media and Communication, 7(1), Article 1.
15. De Coninck, D., Mertens, S., & d'Haenens, L. (2022). Cross-country comparisons of the media impact on anti-immigrant attitudes. Deliverable 7.3, for Work Package 7 of the HumMingBird project.
16. Debrael, M., d'Haenens, L., De Cock, R., & De Coninck, D. (2021). Media use, fear of terrorism, and attitudes towards immigrants and refugees: Young people and adults compared. International Communication Gazette, 83(2), 148–168.
17. Dehghan, E., & Bruns, A. (2022). The Dynamics of Polarisation in Australian Social Media: The Case of Immigration Discourse. In D. Palau-Sampio, G. López García, & L. Iannelli (Eds.), Advances in Public Policy and Administration (pp. 57–73). IGI Global. https://doi.org/10.4018/978-1-7998-8057-8.ch004
18. Dehghan, E., Bruns, A., Mitchell, P., & Moon, B. (2020). Discourse-analytical studies on social media platforms: A data-driven mixed-methods approach. Produsing Theory in a Digital World 3.0: The Intersection of Audiences and Production in Contemporary Theory. Vol. 3., 159–177.
19. Esteve-Del-Valle, M. (2022). Homophily and Polarization in Twitter Political Networks: A Cross-Country Analysis. Media and Communication, 10(2).
20. Feller, A., Kuhnert, M., Sprenger, T., & Welpe, I. (2011). Divided They Tweet: The Network Structure of Political Microbloggers and Discussion Topics. Proceedings of the International AAAI Conference on Web and Social Media, 5(1), Article 1.
21. Fincham, K. (2019). Exploring Political Journalism Homophily on Twitter: A Comparative Analysis of US and UK Elections in 2016 and 2017. Media and Communication, 7(1), 213–224.
22. Freelon, D. (2020). Partition-Specific Network Analysis of Digital Trace Data: Research Questions and Tools. In B. Foucault Welles & S. González-Bailón (Eds.), The Oxford Handbook of Networked Communication (pp. 89–110). Oxford University Press.

23. Freelon, D., Lynch, M., & Aday, S. (2015). Online Fragmentation in Wartime: A Longitudinal Analysis of Tweets about Syria, 2011– 2013. The Annals of the American Academy of Political and Social Science, 659, 166–179.
24. Freelon, D., McIlwain, C. D., & Clark, M. D. (2016). Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice. Center for Media & Social Impact.
25. Froehlich, D. E., Van Waes, S., & Schäfer, H. (2020). Linking Quantitative and Qualitative Network Approaches: A Review of Mixed Methods Social Network Analysis in Education Research. Review of Research in Education, 44(1), 244–268.
26. Galeazzi, A. (2022). Opinion Mining and Clusters Detection in Online Public Debates: A Quantitative Analysis [PhD Thesis]. https://hdl.handle.net/11379/555016
27. Garimella, K. (2018). Polarization on Social Media [PhD Thesis, Aalto University]. https://aalto-doc.aalto.fi:443/handle/123456789/29708
28. Georgiou, M., & Zaborowski, R. (2017). Council of Europe Report: Media Coverage of The "Refugee Crisis": A Cross-European Perspective. Council of Europe.
29. Institute for Strategic Dialogue. (2021). The networks and narratives of anti-refugee disinformation in Europe. Institute for Strategic Dialogue.
30. Jacomy, M. (2021). Situating Visual Network Analysis. Aalborg Universitetsforlag.
31. Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE, 9(6), e98679.
32. McCann, K., Sienkiewicz, M., & Zard, M. (2023). The role of media narratives in shaping public opinion toward refugees: A comparative analysis. International Organization for Migration (IOM), Geneva, 72.
33. Merrill, S., & Copsey, N. (2022). Retweet solidarity: Transatlantic Twitter connectivity between militant anti-fascists in the USA and UK. Social Movement Studies, 0(0), 1–21.
34. Münch, F. V. (2019). Measuring the networked public: Exploring network science methods for large scale online media studies [PhD Thesis]. Queensland University of Technology.
35. Nasralah, T., Elnoshokaty, A., El-Gayar, O., Al-Ramahi, M., & Wahbeh, A. (2022). A comparative analysis of anti-vax discourse on twitter before and after COVID-19 onset. Health Informatics Journal, 28(4), 146045822211358.
36. Nerghes, A., & Lee, J.-S. (2018). The Refugee/Migrant Crisis Dichotomy on Twitter: A Network and Sentiment Perspective. Proceedings of the 10th ACM Conference on Web Science, 271–280. https://doi.org/10.1145/3201064.3201087
37. Nerghes, A., & Lee, J.-S. (2019). Narratives of the refugee crisis: A comparative study of mainstream-media and twitter. Media and Communication, 7(2 Refugee Crises Disclosed), 275–288.
38. Ozerim, M. G., & Tolay, J. (2021). Discussing the Populist Features of Anti-refugee Discourses on Social Media: An Anti-Syrian Hashtag in Turkish Twitter. Journal of Refugee Studies, 34(1), 204–218. Scopus.
39. Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. Telematics and Informatics, 35(1), 136–147.
40. Pöyhtäri, R., Nelimarkka, M., Nikunen, K., Ojala, M., Pantti, M., & Pääkkönen, J. (2021). Refugee debate and networked framing in the hybrid media environment. International Communication Gazette, 83(1), 81–102.
41. Seabold, S., Rutherford, A., De Backer, O., & Coppola, A. (2015). The Pulse of Public Opinion: Using Twitter Data to Analyze Public Perception of Reform in El Salvador [Working Paper]. World Bank. https://doi.org/10.1596/1813-9450-7399
42. Smyrnaios, N., & Ratinaud, P. (2017). The Charlie Hebdo Attacks on Twitter: A Comparative Analysis of a Political Controversy in English and French. Social Media + Society, 3(1), 205630511769364.
43. Sommer, R. (2022). Narrative Dynamics and Migration: Centrifugal vs. Centripetal Forces (OPPORTUNITIES). University of Wuppertal.
44. Stegmeier, J., Schünemann, W. J., Müller, M., Becker, M., Steiger, S., & Stier, S. (2019). Multi-method Discourse Analysis of Twitter Communication: A Comparison of Two Global Political Issues. In R. Scholz (Ed.),

Quantifying Approaches to Discourse for Social Scientists (pp. 285–314). Springer International Publishing. https://doi.org/10.1007/978-3-319-97370-8_10

45. Stoltenberg, D. (2021). Translocal Networked Public Spheres: Spatial Arrangements of Metropolitan Twitter. https://osf.io/m5aqk/
46. Tambuscio, M., Oliveira, D. F. M., Ciampaglia, G. L., & Ruffo, G. (2018). Network segregation in a model of misinformation and fact-checking. Journal of Computational Social Science, 1(2), 261–275.
47. Urman, A. (2020). Context matters: Political polarization on Twitter from a comparative perspective. Media, Culture & Society, 42(6), 857–879.
48. van Klingeren, M., Trilling, D., & Möller, J. (2021). Public opinion on Twitter? How vote choice and arguments on Twitter comply with patterns in survey data, evidence from the 2016 Ukraine referendum in the Netherlands. Acta Politica, 56(3), 436–455.
49. Verbeke, M., Berendt, B., d'Haenens, L., & Opgenhaffen, M. (2017). Critical news reading with Twitter? Exploring data-mining practices and their impact on societal discourse. Communications, 42(2), 127–149.
50. Vilella, S., Lai, M., Paolotti, D., & Ruffo, G. (2020). Immigration as a Divisive Topic: Clusters and Content Diffusion in the Italian Twitter Debate. Future Internet, 12(10), Article 10.
51. Wimmer, A., & Schiller, N. G. (2003). Methodological Nationalism, the Social Sciences, and the Study of Migration: An Essay in Historical Epistemology. The International Migration Review, 37(3), 576–610.
52. Yao, Q., Li, R. Y. M., & Song, L. (2022). Carbon neutrality vs. Neutralité carbone: A comparative study on French and English users' perceptions and social capital on Twitter. Frontiers in Environmental Science, 10.
53. Yoo, J. J. (2019). Opinion leaders on Twitter immigration issue networks: Combining agenda-setting effects and the two-step flow of information [PhD Thesis]. https://doi.org/10.26153/tsw/5355
54. Yousefi Nooraie, R., Sale, J. E. M., Marin, A., & Ross, L. E. (2020). Social Network Analysis: An Example of Fusion Between Quantitative and Qualitative Methods. Journal of Mixed Methods Research, 14(1), 110–124.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

THE PROBLEMS OF THE AUTOMATION BIAS IN THE PUBLIC SECTOR

A LEGAL PERSPECTIVE

Ruschemeier, Hannah
University of Hagen
Hagen, Germany
hannah.ruschemeier@fernuni-hagen.de

KEYWORDS

AI-bias; ADM-decisions; GDPR; discrimination; AI-act PES

DOI: 10.34669/wi.cp/5.6

ABSTRACT

The automation bias describes the phenomenon, proven in behavioural psychology, that people place excessive trust in the decision suggestions of machines. The law currently sees a dichotomy—and covers only fully automated decisions, and not those involving human decision makers at any stage of the process. However, the widespread use of such systems, for example to inform decisions in education or benefits administration, creates a leverage effect and increases the number of people affected. Particularly in environments where people routinely have to make a large number of similar decisions, the risk of automation bias increases. As an example, automated decisions providing suggestions for job placements illustrate the particular challenges of decision support systems in the public sector. So far, the risks have not been sufficiently addressed in legislation, as the analysis of the GDPR and the draft Artificial Intelligence Act show. I argue for the need for regulation and present initial approaches.

1 INTRODUCTION

Digital transformation has led to the ubiquity of algorithmic decision-making. Public and private actors are replacing previously purely human decision-making processes with processes that integrate computer systems or pre-structuring human decisions with automated decision suggestions. In the administrative sector in particular, this increasing use gives rise to numerous legal issues.[1] So-called Algorithmic-Decision-Making-Systems (ADM) make existentially important decisions that affect people's lives, such as the distribution of child benefits,[2] the allocation of university places[3] or support measures in employment services. [4] There are countless articles about the benefits and risks of fully automated decisions. Societal risks,[5] potential for discrimination,[6] ethical questions,[7] legal redress deficit[8] and constitutional requirements such as legitimisation[9] and transparency[10] are the subject of an ongoing debate. The general aim of using technology and automation in public administration is to make decision-making processes more effective, efficient, rational, or neutral. The factual basis as a formerly analogue reality of life has been datafied and thus supposedly rendering it calculable. But numerous examples of algorithmic errors[11] due to inadequate databases, programming errors, or incorrect application have shown that the digitalisation of decision-making processes is neither efficient, desirable, nor compatible with the principles of the rule of law[12] and the protection of fundamental rights in all areas.[13]

In this context, the supposedly clear dichotomy between human and machine decision-making processes suggests a clear distribution of risks to the detriment of full automation,[14] whereas human decisions seem to be sufficiently constrained by existing normative structures such as justification requirements, bias rules, time periods, etc. to legitimise the content of the decision. The risks of these hybrid decision-making processes are not covered by the regulatory systems, for example, there are no provisions for decision-support systems in German administrative law, only a reference to fully automated administrative acts. I focus here on the case of decision-support systems in the administrative sector which produce automated proposals for human decision-makers.

---

[1] (Ruschemeier 2023 i.E.).

[2] https://www.autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf.

[3] (Martini et al. 2020).

[4] (Allhutter et al. 2020), (Scott et al. 2022).

[5] Eg. (Pasquale 2015).

[6] Eg. (Wachter 2019).

[7] Eg. (Mühlhoff 2020).

[8] (Martini, Ruschemeier, and Hain 2021).

[9] Eg. (Liu, Lin, and Y.-J. Chen 2019).

[10] Eg. (Burrell 2016).

[11] (O'Neil 2017).

[12] (Kolain and Ruschemeier 2023).

[13] (Nink 2021).

[14] See the specific norms for algorithmic administrative acts in German Administration Law: § 35a Administrative Procedures Act, § 88 (4) Tax Code; § 31a Social Code X. (Kahneman et al. 2016) argue the opposite, namely that algorithms make better decisions than humans.

Human decisions follow a limited rationality.[15] In an increasingly complex world, it seems more and more difficult to consider all the relevant factors when making decisions.[16] The use of algorithm-based systems to alleviate human decision-making processes is therefore ubiquitous. However, the focus of (scientific) discussions is mainly on visions of "artificial intelligence", robotics, and fully automated decision-making processes. Thus, the risks of algorithmic decision support are not sufficiently reflected in legal and institutional terms, but are in fact widespread.[17]

In between these two poles of fully automated and fully human decision-making lies the as-yet unestablished (legal) category of decision-support, which also seems ubiquitous in everyday life. Digital devices, platforms, or other network-based services are constantly suggesting decisions: from operating modalities and default settings to recommendation algorithms for advertising and content. These decision-support systems become problematic when they are subject to different parameters than purely human or fully automated decisions, parameters which are previously unknown and lead to anomalies, such as the automation bias.[18]

Automation bias is the phenomenon, well established in behavioural psychology, in which people trust the suggestions and decisions of machines against their better judgement.[19] With the widespread use of such systems, for example in education or benefits administration, such biases develop a leverage effect and multiply the number of people affected. Especially in environments where people have to make a large number of similar decisions on a routine basis, the risk of automation bias increases, and expert knowledge alone is not a sufficient defence. At present, the problem of automation bias is not addressed by legislation. External factors such as time pressure or the effort required to check the algorithm can encourage such behaviour.[20]

I will show the legal difficulties inherent in automation bias using the case study of the ASM algorithm (2) and then look at the current legal situation (3) before considering the need for regulation (4).

2 CASE STUDY: THE AUSTRIAN AMS-ALGORITHM AND SIMILAR SYSTEMS

The Austrian ASM-algorithm[21] has been used in the placement of unemployed people and was designed to help case workers make more efficient use of resources in the long term.[22] For this purpose,

[15] (Kahneman 2011).

[16] (Ruschemeier 2022).

[17] (Wachter, Mittelstadt, and Russell 2021).

[18] Some studies (Alon-Barkat and Busuioc 2023) also find other human biases beyond overreliance on machine output, such as selective adherence, i.e. adherence to machine output that systematically differs across sensitive demographic dimensions of decision subjects.

[19] Recently in the context of judicial reviews: (Kazim and Tomlinson 2023). Regarding public administration: (Alon-Barkat and Busuioc 2023); (Green and Y. Chen 2019). See also: (Bailey and Scerbo 2007); (Lyell and Coiera 2017b); (Parasuraman, Molloy, and Singh 1993); (Indramani L. Singh, Anju L. Singh, and Proshanto K. Saha 2007); (Rovira, McGarry, and Parasuraman 2007); (Snow 2021).

[20] (Lyell and Coiera 2017a); (Pilniok 2022).

[21] Arbeitsmarktchancen-Assistenz-System des Arbeitsmarktservices (AMS) Österreich; https://iab.de/iab-veranstaltungen/einblicke-in-das-arbeitsmarktchancen-assistenz-system-der-sogenannte-ams-algorithmus-des-arbeitsmarktservice-ams-oesterreich/.

[22] Currently, the application is suspended due to an ongoing court case. (Der Standard 2022).

the programme divided jobseekers into three groups according to their calculated chances on the labour market: high, medium and low. The AMS-model considered personal characteristics such as age, gender, education, health limitations, caring responsibilities, education, and citizenship.[23] On the basis of the published model, it can now be seen that the model deducts certain points for being a woman (0.14 points) as well as for potentially having care responsibilities as a woman (0.15 points). This reveals that the existence of care responsibilities alone played a role in a woman's future job chances. All people over 30 are also penalised on the basis of their age alone, and the penalty is even more drastic from the age of 50 (0.7 points). People with 'health problems' are also penalised by the system (0.67 points), as are people from non-EU countries.

In addition to the obvious potential for discrimination, other legal problems arose: the competent data protection authority (DPA) prohibited the further use of the programme, as it constituted illegal profiling and a violation of Art. 22 of the GDPR which prohibits fully automated individual decisions with legal effect or similar impairments.[24] The DPA argued that while the final decision rested with the person responsible according to internal PES guidelines, these internal instructions have no "external effect" and are therefore not binding on the authority concerned. In this respect, the affected persons cannot refer to them in a legally effective manner and thus cannot demand a review. The fact that in some cases the counselling time allocated was only ten minutes speaks in favour of a routine acceptance. Furthermore, the DPA argued that it could be assumed that counsellors would increasingly rely[25] on the decision of the AMS as a result of COVID-19. As genuine supervision by a human being is therefore "not bindingly ordered (in the sense of a legal guarantee) for all individual cases and thus not fully guaranteed", Article 22 of the GDPR should be applied with reference to the guidelines of the Article 29 Working Party,[26] which assumes an "automated decision" in the sense of Article 22 GDPR in cases in which automatically calculated results are routinely simply adopted.

In the ensuing legal dispute, however, the Austrian Federal Administrative Court proceeded on the basis of a purely formal assessment of the decision-making structure and did not address the risks of automation bias. The court argued that the assessment was only carried out by the relevant consultants using the model and that routine adoption did not carry any great significance. An appeal against the decision has been submitted, but no decision has been made so far, and the AMS-model is not in use.[27] A similar system has been used in Poland, where statistical analysis has also shown that caseworkers made changes to the automated classification of jobseekers in only 0.58% of the cases examined.[28]

Since, according to the court, this was not a fully automated decision within the meaning of Art. 22 GDPR, despite the many indications suggesting the case handlers simply relied on the suggestions of the system, the protective mechanisms of Art. 22 GDPR do not apply either. In the court's opinion, the decisions were thus purely human decisions; the factual binding effect of the machine suggestions in combination with the internal guidelines was not considered. Thus, constellations of potential automation bias fall between the cracks: the actions of data subjects are made considerably more

[23] https://ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

[24] Österreichische Datenschutzbehörde, decision 16.8.2020, D213.1020, 2020.0513.605.

[25] (Allhutter et al. 2020) point out, that the system's output is supposed to be a "second opinion" but will become a "first opinion" in practical use due to the short time available to case handlers. See also (Scott et al. 2022).

[26] (Art. 29 Data Protection Working Group 2017).

[27] (Scott et al. 2022).

[28] (Niklas, Sztandar-Sztanderska, and Szymielewicz 2015), S. 28, (Scott et al. 2022).

difficult by the engagement of non-transparent systems within the decision-making process, but they are also not entitled to the protection of the additional rights which are only triggered by fully automated decisions.

So far, regulators and courts have ruled that a formal human-in-the-loop is sufficient to prevent a fully automated decision. This means that the problem of automation bias cannot be solved by merely banning fully automated decisions. This could indicate that a new kind of decision category is needed to address human-machine interactions that produce decision-relevant output.

3 DECISION SUPPORT SYSTEMS IN LAW

As the cases show, data protection law does not adequately address the problem of automation bias. In general, the law has so far distinguished between two forms of automated decision systems (ADM) fully automated ADM and partially automated ADM, i.e. decision-support systems with a human in the loop.[29]

3.1 Automation Bias and the GDPR

This interface between machine suggestions and human final decisions has so far only been dealt with in passing, especially in the context of data protection law.[30] Article 22 GDPR establishes the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects or significantly affects that individual (the data subject). The prerequisite for its application is the processing of personal data within the scope of the GDPR, which also applies to public authorities. The right under Art. 22 (1) GDPR, which is interpreted as a prohibition, does not apply if the decision is necessary for the performance of a contract (Art. 22 (2) a)), if a legal basis has been created that provides sufficient safeguards for the rights, freedoms and interests of the data subject (Art. 22 (2) b)), or if it is based on an explicit consent (Art. 22 (2) c)). None of these exceptions apply to public employment services.

It is therefore crucial to understand a decision based solely on fully automated processing. The exact interpretation is unclear and disputed.[31] Legally, there can be no decision if the person concerned already lacks the capacity to decide. In a purely human-based process, this would be the preparation stage for a decision, for example, with the scanning of documents.

Thus, in the example of the Austrian court, if a human retains substantive decision-making power, it could be argued the machine is merely preparing the information, and there is no fully automated decision from a purely external point of view.

However, substantive decision-making power is not congruent with a decision if it is not exercised. It is only when the decision-making power is exercised by a human being that it has an effect on the result. In my opinion, it is not sufficient to limit oneself to a plausibility check or not to influence the automated process. These difficulties of interpretation are compounded by problems of verifiability; decision-making is an internal process that can at best be presumed on the basis of external evidence.

[29] (Pilniok 2022).

[30] (Martini 2021); (Martini, Ruschemeier, and Hain 2021), (Steinbach 2021).

[31] (Bygrave 2020).

If the agents follow the suggestion, it is difficult to actually prove whether a decision has been consciously made or whether a suggestion has been passively adopted.

The formal understanding of fully automated decision-making systems in Art. 22 GDPR provides a normative incentive to develop decision support systems that are not subject to the requirements of the GDPR. However, decades of experience from various application areas,[32] e.g. aviation safety, medicine, etc., as well as basic psychological research[33] shows that simply assuming a human-in-the-loop approach is by no means a sufficient protection. The law should recognise the psychological and technical factors in play in such decision support systems, and develop legal solutions to minimise the corresponding risks. Standards such as Art. 22 GDPR should be read more as a socio-technical norm,[34] or such norms should be created. Many legal requirements do not reflect the fact that digitalisation is a socio-technical development that only works in the interplay between human action and technology.[35] Digitalisation only exists as a result of human-driven processes, but it is increasingly influencing how those processes play out.

3.2 The forthcoming Artificial Intelligence Act and the automation bias

The draft on the Regulation of Artificial Intelligence[36] (AI-Act) at Union level implements the requirement for human oversight in decisions involving artificial intelligence in Art. 14 (2). High-risk AI systems should therefore "be designed and developed so that they can be effectively supervised by natural persons for the duration of the use of the AI system". Article 14 (3) provides that human oversight should either be built into the system (lit. a) or the need for human oversight be highlighted to the user (lit. b) to allow users to understand and be aware of the capacities of the system, and subsequently to interpret, decide on, or deviate from the suggested information. In particular, users are aware of "the danger of potentially over-relying on the output of a high-risk AI system (automation bias) in particular for high risk AI systems used to provide information or recommendations to be taken by natural persons." Through "human supervision", users should, depending on the circumstances and within a proportionate framework, be able to monitor the AI system (para. 4 lit. a) or, in individual cases, to decide against the output of the AI software (para. 4 lit. d), i.e. to follow or not follow the decision proposal generated by the AI. Article 14 (4) (b) explicitly mentions the automation bias. The measures referred to in paragraph 3 are intended to enable the individuals to whom human oversight is assigned to remain *aware* of the possible tendency to automatically rely or over-rely on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons. The actual business conditions for using AI software will usually be such that although users are 'aware' of the biases of the automation, they will not take active control over the decision-making process. For this, the institutional framework conditions of the decision-making situation must be changed to provide time and incentives for administrators to use their own decision-making power, and scope for deviation from system proposals.

[32] (Parasuraman and Riley 1997).

[33] (Lyell and Coiera 2017a).

[34] (Djeffal 2021).

[35] (Mühlhoff 2020).

[36] Proposal for a Regulation of the European Parliament and of the Council Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM/2021/206 final.

It is welcome that the proposed regulation recognises the problem of automation bias. However, its scope is limited to the specific high-risk systems listed in Article 6 in conjunction with Annex III, the scope of which is still evolving during the legislative process. Currently, access to and usage of essential private services, public services, and benefits are envisaged as high-risk systems.[37] Moreover, thus far Art. 14 AI-Act only refers to technical measures; organisational requirements, which in particular consider the context of the decision, are prima facie not required.

4 A NEED FOR NEW REGULATION?

Automation bias remains inadequately addressed by the current legal instruments. Protective mechanisms are needed, particularly in the area of administrative decisions governing existential issues such as employment services and benefits. Interdisciplinary findings from technical science and psychology should be included in the process, with a distinction made between substantive legal limits and procedural requirements.

In administrative law, automation bias can constitute a violation of the principle of official investigation or a failure of discretion. So far, these consequences have not been reflected normatively. Further challenges arise in the context of provability in judicial proceedings; without insight into the internal decision-making process those affected face greater difficulty in proving automation bias. As explained elsewhere, in these cases, evidence can be facilitated.[38] This is justified by a distribution of risk as an extension of the rule of law principle: for the state to reap the fruits of the efficiency from using the system, it should then also have to bear the risk and be able to prove that there is no automation bias risk, e.g. through deviation rates or sufficient processing time.

In certain areas sensitive to fundamental rights, such as criminal law, but also in the case of vital administrative services of general interest, systems should not make detailed proposals for decisions, but should be used only as a tool for establishing the facts. In other areas, procedural mechanisms should be put in place to mitigate the risk of automation bias, such as confirmation requirements, justification requirements by the human decision-maker, or a four-eye principle. Insights from psychology, computer sciences, and administrative sciences should be used to determine what kind of confirmation mechanisms are useful, how user interfaces can be constructed, and which technical and institutional safeguards can be considered. Procedurally, time pressure in mass procedures should be considered as risk factor for automation bias, especially if a very short decision time is accompanied by a predominant acceptance of the system proposal.

In decision-making contexts that are particularly sensitive in terms of fundamental rights, it may be appropriate to impose a burden of proof on a case officer who makes extensive use of decision support systems in order to ensure that they address the content of the proposed decision. In any case, it should not be more time-consuming to deviate from the algorithm-based recommendation than to follow the proposed decision.

[37] Annex III (5): Access to and enjoyment of essential private services and public services and benefits: (a) AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services

[38] (Martini, Ruschemeier, and Hain 2021); (Ruschemeier 2023).

FUNDING INFORMATION AND ACKNOWLEDGEMENTS

This paper is part of the project „Automation Bias als Rechtsproblem" at the ABD-institute of the University of Hagen.

REFERENCES

1. Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. "Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective." *Front. Big Data* 3:5. https://doi.org/10.3389/fdata.2020.00005.
2. Alon-Barkat, Saar, and Madalina Busuioc. 2023. "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): 153–69. https://doi.org/10.1093/jopart/muac007.
3. Art. 29 Data Protection Working Group. 2017. "WP 251 Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679.".
4. Bailey, N. R., and M. W. Scerbo. 2007. "Automation-Induced Complacency for Monitoring Highly Reliable Systems: The Role of Task Complexity, System Experience, and Operator Trust." *Theoretical Issues in Ergonomics Science* 8 (4): 321–48. https://doi.org/10.1080/14639220500535301.
5. Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 205395171562251. https://doi.org/10.1177/2053951715622512.
6. Bygrave, Lee A. 2020. "Art. 22 GDPR." In *The EU General Data Protection Regulation (GDPR)*, edited by Christopher Kuner, Lee A. Bygrave, Christopher Docksey, and Laura Drechsler: Oxford University Press.
7. Der Standard. 2022. ""Zum in-Die-Tonne Treten.": Neue Kritik Am AMS-Algorithmus." https://www.derstandard.de/story/2000135277980/neuerliche-kritik-am-ams-algorithmus-zum-in-die-tonne-treten.
8. Djeffal, Christian. 2021. "Art. 22 DSGVO als sozio-technische Gestaltungsnorm: Eine Neuinterpretation der Regelung von automatisierten Entscheidungen." *DuD* 45 (8): 529–33. https://doi.org/10.1007/s11623-021-1484-4.
9. Ebers, Martin, ed. 2023. Stichwort Kommentar Legal Tech: Recht | Geschäftsmodelle | Technik. Baden-Baden: Nomos.
10. Green, Ben, and Yiling Chen. 2019. "The Principles and Limits of Algorithm-in-the-Loop Decision Making." *Proc. ACM Hum.-Comput. Interact.* 3 (CSCW): 1–24. https://doi.org/10.1145/3359152.
11. Indramani L. Singh, Anju L. Singh, and Proshanto K. Saha. 2007. "Monitoring Performance and Mental Workload in an Automated System." In *Engineering Psychology and Cognitive Ergonomics*, 426–35: Springer Nature. https://www.researchgate.net/publication/221096856_Monitoring_Performance_and_Mental_Workload_in_an_Automated_System.
12. Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* [22e druk]. Psychology/economics. New York: Farrar, Straus and Giroux.
13. Kahneman, Daniel, Andrew Rosenfield, Linnea Gandhi, and Tom Blaser. 2016. "Reducing Noise in Decision Making: Interaction." *Harvard business review* 94 (12): 36–43. https://dialnet.unirioja.es/servlet/articulo?codigo=5747404.
14. Kazim, Tatiana, and Joe Tomlinson. 2023. "Automation Bias and the Principles of Judicial Review." *Judicial Review*, 1–8. https://doi.org/10.1080/10854681.2023.2189405.
15. Kolain, Michael, and Hannah Ruschemeier. 2023. "E-Government." In Ebers 2023, 428–44.
16. Liu, Han-Wei, Ching-Fu Lin, and Yu-Jie Chen. 2019. "Beyond State V Loomis: Artificial Intelligence, Government Algorithmization and Accountability." *Int J Law Info Tech* 27 (2): 122–41. https://doi.org/10.1093/ijlit/eaz001.
17. Lyell, David, and Enrico Coiera. 2017a. "Automation Bias and Verification Complexity: A Systematic Review." *Journal of the American Medical Informatics Association* 24: 423–31.

18. Lyell, David, and Enrico Coiera. 2017b. “Automation Bias and Verification Complexity: A Systematic Review.” *J Am Med Inform Assoc* 24 (2): 423–31. https://doi.org/10.1093/jamia/ocw105.
19. Martini, Mario. 2021. “Art. 22 DSGVO.” In *Datenschutz-Grundverordnung: Bundesdatenschutzgesetz*, edited by Boris Paal and Daniel A. Pauly. 3.th ed. München: C. H. Beck.
20. Martini, Mario, Jonas Botta, David Nink, Michael Kolain, and Bertelsmann Stiftung. 2020. “Automatisch erlaubt?”.
21. Martini, Mario, Hannah Ruschemeier, and Jonathan Hain. 2021. “Staatshaftung Für Automatisierte Verwaltungsentscheidungen.” *VerwArch* (1): 1–37.
22. Mühlhoff, Rainer. 2020. “Human-Aided Artificial Intelligence: Or, How to Run Large Computations in Human Brains? Toward a Media Sociology of Machine Learning.” *New Media & Society* 22 (10): 1868–84. https://doi.org/10.1177/1461444819885334.
23. Niklas, Jędrzej, Karolina Sztandar-Sztanderska, and Katarzyna Szymielewicz. 2015. “Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making.” https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf.
24. Nink, David. 2021. Justiz Und Algorithmen: Über Die Schwächen Menschlicher Entscheidungsfindung Und Die Möglichkeiten Neuer Technologien in Der Rechtsprechung. Berlin: Duncker & Humblot.
25. O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin Books. https://images2.penguinrandomhouse.com/cover/9780553418835.
26. Parasuraman, Raja, Robert Molloy, and Indramani L. Singh. 1993. “Performance Consequences of Automation-Induced 'Complacency'.” *The International Journal of Aviation Psychology* 3 (1): 1–23. https://doi.org/10.1207/s15327108ijap0301_1.
27. Parasuraman, Raja, and Victor Riley. 1997. “Humans and Automation: Use, Misuse, Disuse, Abuse.” *Human Factors* 39: 230–53.
28. Pasquale, Frank. 2015. The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge: Harvard University Press.
29. Pilniok, Arne. 2022. “Administratives Entscheiden Mit Künstlicher Intelligenz: Anwendungsfelder, Rechtsfragen Und Regelungsbedarfe.” *JZ* 77 (21): 1021–31. https://doi.org/10.1628/jz-2022-0337.
30. Rovira, Ericka, Kathleen McGarry, and Raja Parasuraman. 2007. “Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task.” *Human Factors* 49 (1): 76–87. https://doi.org/10.1518/001872007779598082.
31. Ruschemeier, Hannah. 2022. “Privacy als Paradox?” In *Künstliche Intelligenz, Demokratie Und Privatheit*, edited by Michael Friedewald and Alexander Roßnagel, 211–38. Baden-Baden: Nomos. https://www.nomos-elibrary.de/10.5771/9783748913344-211/privacy-als-paradox-rechtliche-implikationen-verhaltenspsychologischer-erkenntnisse?page=3. Accessed October 06, 2022.
32. Ruschemeier, Hannah. 2023. “Haftung Des Staates.” in Ebers 2023, 583–93.
33. Ruschemeier, Hannah. 2023 i.E. “"Künstliche Intelligenz" in Der Verwaltung Im Mehrebenensystem.” In *Herausforderungen Für Das Verwaltungsrecht*, edited by Hermann Hill and Veith Mehde, 111–32. Berlin: Duncker & Humblot.
34. Scott, Kristen M., Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. “Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective.” In *FAccT 2022: Conference on Fairness, Accountability, and Transparency*, edited by Association for Computing Machinery, 2138–48.
35. Snow, Thea. 2021. “From Satisficing to Artificing: The Evolution of Administrative Decision-Making in the Age of the Algorithm.” *Data & Policy* 3:e3. https://doi.org/10.1017/dap.2020.25.
36. Steinbach, Kathrin. 2021. *Regulierung Algorithmenbasierter Entscheidungen: Grundrechtliche Argumentation Im Kontext Von Artikel 22 DSGVO*. Internetrecht und digitale Gesellschaft 28. Berlin: Duncker & Humblot. https://elibrary.duncker-humblot.com/book/55403/regulierung-algorithmenbasierter-entscheidungen.
37. Wachter, Sandra. 2019. “Affinity Profiling and Discrimination by Association in Online Behavioural Advertising.” *Berkeley Technology Law Journal*, 1–74. https://doi.org/10.2139/ssrn.3388639.

38. Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2021. “Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI.” *Computer Law & Security Review* 41:105567. https://doi.org/10.2139/ssrn.3547922.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

ALGORITHMIC MANAGEMENT IN THE FOOD DELIVERY SECTOR – A CONTESTED TERRAIN?

EVIDENCE FROM A FIRM-LEVEL CASE-STUDY ON ALGORITHMIC MANAGEMENT AND CO-DETERMINATION

Wotschack, Philip
Weizenbaum Institute for the Networked Society & WZB Berlin Social Science Center
philip.wotschack@weizenbaum-institut.de

Hellbach, Leon
Weizenbaum Institute for the Networked Society & WZB Berlin Social Science Center
leon.hellbach@wzb.eu

Butollo, Florian
Weizenbaum Institute for the Networked Society & WZB Berlin Social Science Center
florian.butollo@weizenbaum-institut.de

Ziour, Jordi
Weizenbaum Institute for the Networked Society & WZB Berlin Social Science Center
jordi.ziour@wzb.eu

KEYWORDS

algorithmic management; co-determination; regulation; platform work; food delivery; precarity

DOI: 10.34669/wi.cp/5.7

ABSTRACT

Forms of algorithmic management (AM) play an essential role in organizing food delivery work by deploying AI-based systems for coordinating driver routes. Given the risks of precarity and threats posed by AM that are typically related to (migrant) platform work, the question arises to what extent structures of co-determination are able to positively shape this type of work and the technologies involved. Based on an intense case-study in a large food delivery company, this paper is guided by three questions: (1) How is algorithm-based management and control used by the company? (2) How is it perceived by the couriers, also in relation to other aspects of their work? (3) What are the works council's priorities, strategies, and achievements regarding co-determination practices? Contrary to the prevalent perception in the literature on the subject of AM, our analysis shows that human agency is still pivotal when algorithm-based systems are used to manage work processes. While data- and AM-related issues do not represent a central area of conflict, we find that co-determination rights in this domain can translate into a powerful bargaining resource of the works council with regard to the companies' digital business model. Our study also shows that algorithmic management poses problems of non-transparency and information asymmetry, which calls for new forms and procedures of co-determination.

1 INTRODUCTION

The term algorithmic management (AM) refers to the use of algorithm-based systems and tools in an organization's management of its work force, labor processes, and work performance (see Meijerink & Bondarouk, 2023; Wood, 2021). Often it is based on artificial intelligence (AI) systems that automate decision-making and technology-based control (Kellogg, Valentine & Christin, 2020). AM plays an essential role in organizing food delivery work. Taking into account customer demand with restaurant and driver availabilities, the sequence of distributions is calculated and assigned to the couriers by an app on their mobile phones in order to optimize their routes. This process entails constant tracking of the couriers along their routes. One stream in the scientific literature and public debate on platform work emphasizes the control potential of algorithm-based management systems, often referring to the food delivery sector as a typical example (Veen, Barratt & Goods, 2020; Woodcock, 2020). In this view, workers are not only exposed to precarious working conditions but also to algorithm-based forms of monitoring and control. Given the risks of precarity and threats that AM systems typically pose in platform work, the question arises to what extent structures of co-determination are able to alter the negative nature of this type of work. Based on an intense case-study in a large food delivery company, this paper is guided by three questions: (1) How is algorithm-based management and control used by the company? (2) How is it perceived by the couriers, also in relation to other aspects of their work? (3) What are the works council's priorities, strategies, and achievements regarding co-determination practices? The paper closes by discussing emerging demands regarding the regulation and co-determination of AM. The study is part of the European research project INCODING funded by the European Commission (https://incoding-project.eu). The project conducts firm-level case studies in two sectors in four European countries. It precisely focuses on new challenges for worker representation and regulation in the context of algorithm-based control.

2 BACKGROUND AND RESEARCH QUESTIONS

Work at digital platform companies is a typical field of employment for migrant workers, especially in the area of food delivery. Such platforms are known for precarious working conditions in terms of low-skilled tasks, temporary contracts, low pay and (unreliably) flexible working hours. At the same time, they attract migrant workers (often refugees) due to their easy accessibility through low formal requirements, low language barriers, and short recruitment procedures (van Doorn, Ferrari & Graham, 2022).

While many articles deal with platform work in terms of precarious and migrant work, others have focused primarily on the functioning and impact of AM in their firm-level case studies. Besides (location-based forms of) platform work, such as food delivery or other driving services (see the overview by Lücking, 2019), prominent fields of research regarding AM are logistics (Butollo et al., 2018; Staab & Geschke, 2019), manufacturing (Evers, Krzywdzinski & Pfeiffer, 2019), and HR work (Spielkamp & Gießler, 2020). In the German food delivery sector, AM occurs in the form of "app-based management" (Ivanova et al., 2018) and is often discussed as an example of high external performance control in the sense of Kellogg, Valentine & Christin (2020). The smartphone is the focal point of algorithmic management in location-based platform work. It not only ensures the mobility of platform workers but also enables the extensive collection of data that can be evaluated – in particular positional data via GPS.

The study by Ivanova et al. (2018) on the management of food delivery platform work via smart phone applications provided evidence that tracking movement generates an enormous amount of data, which enables comprehensive control of work processes. Automatically evaluating this data serves to optimize the processes and to monitor the work performance of the "riders," as couriers are called internally. The assignment of work orders is based on data evaluation. Automated decision-making occurs through algorithms, which often creates the impression of technical rationality and objectivity. The app can also be used to generate additional incentives for motivation and performance improvement through push messages. By offering minor choices, the app can foster the impression of autonomy and set incentives to increase individual productivity gains ("digital nudging") (Lücking, 2019).

Data on work performance is sometimes used to initiate competition among workers, but it is also used for hierarchical purposes by dividing couriers into different groups. Lucrative shifts or orders are only displayed to "best performers." A central element of the algorithmic control by the app is information asymmetry: The couriers remain unaware of the exact extent and purpose of the service. They know neither how the summary metrics used to monitor their performance are calculated nor how the metrics enter into decisions on the shifts or orders offered to them (Schreyer & Schrape, 2018).

According to given data protection regulations in Germany, employees must agree to the processing of their personal data individually and voluntarily unless such processing is legitimized by relevant company agreements under data protection law (Wedde, 2020). Problems arise when there exists neither individual consent nor works councils willing and capable of negotiating appropriate company agreements. Consequently, it can be expected that companies using these systems operate in a legal gray area. Often the use of these systems is indeed illegal.

In spring 2021, the data protection officer of the state of Baden-Württemberg raised some concerns regarding the "Scoober" app, an algorithm-based app used by a large food-delivery platform (see Tagesschau from May 21st, 2021): The data that the app collects and stores about couriers is documented in several data reports, showing that it is possible to track down with high precision when a driver is assigned an order, picks it up, and delivers it. The data protection officer concluded that this "is a very close-meshed monitoring of the employment relationship." The exact location of the couriers is passed on at intervals of 15 to 20 seconds. According to the data protection officer, this leads to so-called tracking, i.e., "constant monitoring of work performance," which he believes is "clearly illegal." The app also sends personal data to third parties, such as Google. The food delivery company denied the allegation and argued that the courier app complied with the applicable data protection regulations since time and location data are essential for the delivery service to function properly. The company also stated that the data collected was not used for unauthorized performance or behavior control and that the couriers were informed on how and for what purpose the data is used. The lawsuit is still ongoing. It demonstrates the difficulties and possible limitations when legal regulations regarding data protection are applied.

While the food delivery sector is often regarded as an example of strong algorithm-based control and standardization of low-skilled work, case-studies in the manufacturing or logistics sector draw a more ambiguous picture. On the one hand, algorithm-based work governance at industrial workplaces is also criticized for its potential to gather data on worker productivity and hence the ability to closely monitor activities (Falkenberg, 2018). Particularly, in assembly work and logistics, algorithm-based assistance systems are applied to guide workers through the assembly process or in the selection of parts. On the other hand, studies show that these systems can indeed be deployed with very different

concepts of work: Algorithm-based assistance systems can provide flexible, situational information to employees, or they can be used to improve the transparency of work processes, optimize individual work performance and work organization, and increase the quality of tasks and enhance skills (Klippert, 2020).

The literature on AM also highlights that structures of co-determination can be a crucial factor in this ambiguous field. Several studies show the importance of co-determination regarding both the introduction of new (digital) technology and issues of performance regulation to recognize aspects of a human-oriented design of assistance systems (Albrecht & Görlitz, 2021; Evers, Krzywdzinski & Pfeiffer, 2019; Krzywdzinski, Gerst & Butollo, 2023). A notable result is the relatively high acceptance of digital assistance systems, even in highly standardized processes. There are few conflicts, also due to the strong role of works councils in securing data protection criteria and preventing performance monitoring and behavioral control. Moreover, there is evidence that the acceptance of algorithm-based assistance systems (such as smart wearables) by workers relates to issues of transparency and co-determination. Employees tend to accept such systems if they retain control over the data and data usage and if this has a clear benefit for their work – especially in terms of reducing workload (Evers, Krzywdzinski & Pfeiffer, 2019).

Given the outlined risks of algorithm-based control in the food delivery sector, which is characterized by both a lack of co-determination and a high level of precarious labor conditions, the question arises to what extent algorithmic control is exercised and how structures of co-determination can make a difference here.

3 EVIDENCE FROM THE CASE STUDY

The following results are based on an intense firm-level case study from 2022 in a large food delivery company. In contrast to other parts of the platform economy, the company issues fixed-term and permanent contracts to their couriers. After long periods of labor disputes, structures of co-determination have been introduced. This specific organizational setting gives us the opportunity to study the role of co-determination in the food delivery sector, which has not been covered in academic literature before. We have conducted interviews with managers, members of the work council, couriers, and external experts. In this section we outline the main results with regard to (1) the company's aims and use of algorithmic management, (2) the experience and evaluation of AM practices by the couriers and the works council, and (3) the works council's priorities, strategies and achievements regarding co-determination practices.

3.1 Management objectives regarding the use of AM

In the observed company, algorithm-based management takes place via an app that couriers need to install on their cell phones. It assigns jobs to couriers, navigates them to the destination and transmits information about pickup and arrival times to customers. This means the company continually tracks the location, speed, response time, delivery time, and route of the couriers. But, according to the company officials and couriers we interviewed, this information is not used to discipline couriers and achieve performance gains, at least not in an automated way. The management emphasized that individual performance characteristics are neither generated nor used for individual performance control. The works council is skeptical in this respect and fears that such information might be used for regular performance reviews.

Overall, our study provides evidence that AM is mainly used by the company for functional reasons, i.e., for optimizing the sequence and allocation of orders. Humans could clearly not oversee and efficiently manage such large numbers of couriers and orders in the delivery area. According to a typology by Nies (2021), this type of technology use represents "process-oriented rationalization," in contrast to rationalization strategies focusing on individual performance control. This orientation fulfills the function of maximizing efficiency by processing data fast, keeping routes short, and enlarging the geographical scope of deliveries. Nevertheless, it does not mainly aim at individual work performance since couriers are not expected to finish more than around two deliveries per hour and the maximum distance of orders cannot exceed a given number of kilometers.

How does algorithmic management relate to control issues in our case study? It is evident that couriers are instructed and directed and that their performance is recorded (e.g., start of work, speed, distance, and number of orders). The number of orders also feeds into a bonus system, which rewards couriers when achieving certain numbers of orders per month. But no direct disciplining occurs if couriers are too slow. The technically possible control potential is clearly not exhausted here. We do not find evidence for automated forms of performance control, trying to push couriers or punishing them if late on arrival. The app does indicate couriers who get behind schedule by highlighting the arrival time in red, but it does not execute any automated forms of sanctions. The main variable for the company's productivity, regarding the delivery process, is the efficient coordination of tasks – not the individual work performance.

3.2 Experience and evaluation of AM practices by couriers

Tracking and performance recording are widely accepted by the couriers we interviewed, who consider it as "part of the job." We also find evidence that some couriers even prefer to work with the app over constantly being monitored by a human superior. The app is partly experienced as a liberation from direct, personal management control. Interaction with private apps or tracking of private information are more likely to be discussed as hazards. Hence, there is often the desire for a company cell phone. At the same time, the works council and some riders with a critical stance have strong concerns regarding data protection issues. They emphasize the risk that the company might collect and process information that is not obligatory for the mere execution of the work process. Issues of algorithmic control and data acquisition are seen as a crucial point for negotiations between the works council and the management. Interestingly, the works council applies a kind of double-edged strategy here. On the one hand, it strives for more transparency and co-determination regarding the development and functioning of the app. On the other hand, they can use their information and approval rights (granted by the Works Council Constitution Act) to enforce non-AM related claims. In this respect, blocking and delaying software adaptations by not consenting to its implementation represents a strong means to pressure companies that apply digital business models.

Surprisingly, basic flaws of the app are a major topic amongst couriers. Bad navigation and poorly calculated arrival times are seen as an obstacle to good work performance. Moreover, the lack of transparency of the app was seen as a major shortcoming. Couriers are unsure what information is tracked and who might possibly see it and use it for performance assessments. As stated above, our research does not provide evidence of such malpractice at the company surveyed. Still, the insecurity about whether this is done does unsettle couriers and thus results in indirect disciplining. As one rider comments:

> So, there's this fear that it'll kind of backfire on me. That there is something like a digital profile of me. And if I somehow make mistakes or become rebellious, then I only get very thankless orders, so to speak. I already had the feeling that a few colleagues were very reserved when it came to criticism or confrontation. (Courier)

Feelings of insecurity in this regard may be even more significant amongst vulnerable groups like migrant workers, who represent a large proportion of the workforce.

In line with the existing literature emphasizing information asymmetries due to the black-box character of algorithmic systems, it is difficult for the works council to understand and evaluate the functions of the app regarding their effects on couriers. The works council criticizes that the management only reluctantly provides insights on these matters. As a consequence, the works council and individual couriers have developed reverse engineering strategies to grasp the functioning of the app, i.e., using their own Python programming skills and documentation to assess the algorithm of the app.

3.3 Works council priorities, strategies, and achievements

Regarding the labor policy background, the company is characterized by a very active, dedicated general works council, which uses all options to improve the couriers' working conditions (including appealing to the labor court). However, the focus is not mainly on control questions relating to the algorithm, but on other topics. This includes the definition of the delivery area (which the company wants to be as large as possible) or the destination of the last delivery (as close to the riders' home as possible). Work cell phones, work equipment (first of all, the bikes), pay and working time issues, as well as a fair distribution of shifts, are major issues forming the companies' main contested terrain. The works council has been successfully engaged in all these issues. The app and related control issues, in contrast, rather remain secondary. The works council is primarily concerned with access to the functional parameters, understanding how the app is processing this information and how it affects the work of the couriers. The works council recognizes the need to engage with the app, but reports difficulties in doing so:

> I have an idea of what I do as a works council member – co-determination rights. But the problem is when it comes to the question of what I should deal with precisely. I'm poking around in the dark. (Works council member)

Therefore, the works council can only assess the consequences of AM to a limited extent. Thus, the scope for co-determination is restricted, and there remain uncertainties about the effects of possible changes in the AM-system. This is illustrated by the attempts to co-determine the length of tracking intervals:

> The thing is, we have no idea about what the impact of, for example, extending the tracking intervals will be. That's always the problem. And we are not told that either. If I have a minute now […], could it be that the orders will become totally stupid for the couriers? Because they aren't tracked as often anymore. And then they get worse jobs? Maybe they'll get better as a result, but those are the scenarios that we can't answer. (Works council member)

4 Conclusions and Discussion – Lessons Learned for the Regulation and Co-Determination of AM

Regarding the regulation of AM, Germany is, on the one hand, characterized by an overall lack of formalized regulations with an explicit focus on AM issues. On the other hand, a relatively large number of established legal regulations, sectoral and company agreements, and union and works council activities are already indirectly governing the field of AI and AM application. They address issues of data protection, platform work, co-determination, or discrimination. However, in many respects, the existing national regulations do not cover specific issues that arise in the course of AM, as shown by our case study (see also Krzywdzinski, Gerst & Butollo, 2023; Molina et al., 2023).

Previous studies have identified challenges for policy-makers and the regulation of AM in the following three areas of the German workplace, which are confirmed by our findings: (1) Transparency issues: Employers often do not provide sufficient information on the methods used in AI applications. (2) Control issues: According to the existing data protection regulations, employers may collect and process individual data when this information is used to fulfill the specific work purpose. Since this regulation leaves room for interpretation, misuse by companies can occur. (3) Co-determination issues: Processual forms of co-determination gain importance (Krzywdzinski, Gerst & Butollo, 2023), because governance and monitoring of AI and AM are becoming permanent tasks in the context of systems that are frequently updated. Rights of co-determination are less effective as soon as such systems have been introduced, amplifying the importance that employees, works councils, and HR managers possess the appropriate skills and information to draw the right conclusions, anticipating possible long-term effects and unintended consequences.

Our findings complement existing research in this field by shedding light on the role and interplay of management objectives, experiences of couriers, strategies of works councils, and co-determination issues regarding AM. A question of crucial importance is to what extent and in which way AM has become a new contested terrain of labor policies in the food delivery sector. Overall, our findings do not support the idea of strong labor conflicts regarding issues of AM in the German regulatory context. Problems and conflicts rather arise from the couriers' general precarious work and employment conditions.

Our study did not find evidence for algorithm-based performance control at the individual level, as suggested by the respective literature in the field of AM (Kellogg, Valentine & Christin, 2020) and the platform economy (Schreyer & Schrape, 2018). The given potential of a rigid, algorithmically driven control system, as it is provided by the collection of vast amounts of data and technological possibilities (as demonstrated in other cases), has not been realized in practice in this case. Moreover, we find close linkages and interactions between (automated) algorithm-based order assignments and human readjustments by couriers and operators. In this respect, the term algorithmic management might be misleading and should be used more carefully in the scientific debate, since it tends to suggest and emphasize the (AI-based) substitution of management functions.

Concerning the strategies of the works council and couriers, labor policies are first of all concerned with traditional issues in terms of pay, working hours, work equipment, or safety issues. Despite the works council's engagement and (fixed- and long-term) employment contracts, classic elements of precarious employment in the low-wage sector tend to persist, such as low pay, the lack of provision of core work equipment, bad and often dangerous working conditions, and insecure employment prospects due to high market fluctuations. Still, for many workers, especially migrants, who often are

particularly reliant on initial labor market access, this form of work offers low threshold job opportunities.

When trying to tackle issues of AM, the works council faces difficulties to obtain the necessary information on the parameters feeding into the AM system, to understand their functioning and interaction, and to evaluate the effects of possible changes and alternative usages – despite rather rich co-determination rights and recent reforms (Work Council Modernization Act) in the German context. This raises the crucial question to what extent employee representations are able and need to be enabled to co-determine AI- or AM-based systems themselves, as often suggested in the current debate, underlining the need for more processual rights. An alternative approach to co-determination might put more emphasis on regulating the effects of AM-based systems to prevent negative outcomes regarding staffing, work hours, workload, or safety. Such an approach would rely on classical fields and instruments of employee representation.

Eventually, we find evidence that given regulations touching issues of data protection and technology can provide powerful means to works councils to achieve goals in other areas of action. In the digital platform economy, both efficient day-to-day business and quick innovation depend greatly on the collection and processing of data as well as on the fast and continuous development of (globally used) software. Putting pressure on the collection or processing of data can therefore quickly threaten companies' core business interests and amplify their cooperativeness in bargaining processes. In this respect, existing co-determination rights regarding issues of AM can provide a new, powerful bargaining resource to employee representation in AM-based business models. To the best of our knowledge, this fact has not yet received much attention in previous research. It underlines the need to study bargaining processes, power resources, and negotiation strategies in the area of AM more carefully. Future research in this field should take a broader perspective on AM-related policies in organizations, also considering issues and conflicts in other, "traditional" areas of action. Moreover, it seems useful to build on insights from bargaining and power resource theories to extend our understanding of the role AM-related issues and conflicts in organizational labor policies.

ACKNOWLEDGEMENTS

This study is part of the European research project "Democracy at Work through Transparent and Inclusive Algorithmic Management" (INCODING) funded by the European Commission under EaSI (https://incoding-project.eu).

REFERENCES

1. Albrecht, Thorben; Görlitz, Julia. 2021. *Künstliche Intelligenz als Handlungsfeld für Gewerkschaften*. Edited by denk-doch-mal.de. https://denk-doch-mal.de/thorben-albrecht-julia-goerlitz-kuenstliche-intelligenz-als-handlungsfeld-fuer-gewerkschaften/, 11/24/2021.

2. Butollo, Florian; Engel, Thomas; Füchtenkötter, Manfred; Koepp, Robert; Ottaiano, Mario. 2018. *Wie stabil ist der digitale Taylorismus? Störungsbehebung, Prozessverbesserungen und Beschäftigungssystem bei einem Unternehmen des Online-Versandhandels*. In: *AIS-Studien* 11 (2): 143–159.

3. Evers, Maren; Krzywdzinski, Martin; Pfeiffer, Sabine. 2019. *Wearable Computing im Betrieb gestalten*. In: *Arbeit* 28 (1): 3–27. DOI: 10.1515/arbeit-2019-0002.

4. Falkenberg, Jonathan. 2018. *Mobile Kontrolleure. Eine arbeitssoziologische Analyse digitaler Assistenzsysteme in der Logistik 4.0*. In: Anemari Karačić & Hartmut Hirsch-Kreinsen (eds.): Logistikarbeit in der digitalen Wertschöpfung. Perspektiven und Herausforderungen für Arbeit durch technologische Erneuerungen. Tagungsband zur gleichnamigen Veranstaltung am 5. Oktober 2017: FGW – Forschungsinstitut für gesellschaftliche Weiterentwicklung e.V.: 37–56.

5. Ivanova, Mirela; Bronowicka, Joanna; Kocher, Eva; Degner, Anne. 2018. *Foodora and Deliveroo: The App as a Boss? Control and autonomy in app-based management - the case of food delivery riders*. Düsseldorf: Hans-Böckler-Stiftung (Working Paper Forschungsförderung, 107).

6. Kellogg, Katherine C.; Valentine, Melissa A.; Christin, Angéle. 2020. *Algorithms at Work: The New Contested Terrain of Control*. In: *ANNALS* 14 (1): 366–410. DOI: 10.5465/annals.2018.0174.

7. Klippert, Jürgen. 2020. *Gute Arbeit mit MES. Mensch-Organisation-Technik bei Manufacturing Execution Systems*. Frankfurt am Main: Ressort Zukunft der Arbeit der IG-Metall.

8. Krzywdzinski, Martin; Gerst, Detlef; Butollo, Florian. 2023. *Promoting human-centred AI in the workplace. Trade unions and their strategies for regulating the use of AI in Germany*. In: *Transfer: European Review of Labour and Research* 29 (1): 53-70. DOI: 10.1177/10242589221142273.

9. Lücking, Stefan. 2019. *Arbeiten in der Plattformökonomie. Über digitale Tagelöhner, algorithmisches Management und die Folgen für die Arbeitswelt*: Hans-Böckler-Stiftung (Forschungsförderung, Report Nr. 5).

10. Meijerink, Jeroen; Bondarouk, Tanya. 2023. *The duality of algorithmic management: Toward a research agenda on HRM algorithms, autonomy and value creation*. In: *Human Resource Management Review* 33 (1): 1–14. DOI: 10.1016/j.hrmr.2021.100876.

11. Molina, Oscar; Butollo, Florian; Makó, Csaba; Godino, Alejandro; Holtgrewe, Ursula; Illsoe, Anna et al. 2023. *It takes two to code: a comparative analysis of collective bargaining and artificial intelligence*. In: *Transfer: European Review of Labour and Research* 29 (1): 87–104. DOI: 10.1177/10242589231156515.

12. Nies, Sarah. 2021. *Eine Frage der Kontrolle? Betriebliche Strategien der Digitalisierung und die Autonomie von Beschäftigen in der Produktion*. In: *Berliner Journal für Soziologie* 31 (3-4): 475–504. DOI: 10.1007/s11609-021-00452-8.

13. Schreyer, Jasmin; Schrape, Jan-Felix. 2018. *Plattformökonomie und Erwerbsarbeit. Auswirkungen algorithmischer Arbeitskoordination - das Beispiel Foodora*. Düsseldorf: Hans-Böckler-Stiftung (Working Paper Forschungsförderung, 087).

14. Spielkamp, Matthias; Gießler, Sebastian. 2020. *Automatisiertes Personalmanagement und Mitbestimmung. KI-basierte Systeme für das Personalmanagement – was ist fair, was ist erlaubt?* Düsseldorf: Hans-Böckler-Stiftung (Working Paper Forschungsförderung, 191).

15. Staab, Philipp; Geschke, Sascha Christopher. 2019. *Ratings als arbeitspolitisches Konfliktfeld. Das Beispiel Zalando*. Düsseldorf: Hans-Böckler-Stiftung (Study der Hans-Böckler-Stiftung, 429).

16. Tagesschau. 2021. *Tagesschau. Sendung vom 21.05.2021, 20:00 Uhr*. Edited by Tagesschau. https://www.tagesschau.de/multimedia/sendung/ts-42969.html, 5/21/2021.

17. van Doorn, Niels; Ferrari, Fabian; Graham, Mark. 2022. *Migration and Migrant Labour in the Gig Economy: An Intervention*. In: *Work, Employment and Society*. DOI: 10.1177/09500170221096581.

18. Veen, Alex; Barratt, Tom; Goods, Caleb. 2020. *Platform-Capital's 'App-etite' for Control: A Labour Process Analysis of Food-Delivery Work in Australia*. In: *Work, Employment and Society* 34 (3): 388–406. DOI: 10.1177/0950017019836911.

19. Wedde, Peter. 2020. *Arbeitsrechtliche Aspekte und Beschäftigtendatenschutz*. Edited by Algorithm Watch. https://algorithmwatch.org/de/auto-hr/gutachten-arbeitsrecht-datenschutz-wedde/, 2/28/2020.

20. Wood, Alex J. 2021. *Algorithmic Management: Consequences for Work Organisation and Working Conditions*. Seville: European Commission (JRC Working Papers Series on Labour, Education and Technology).

21. Woodcock, Jamie. 2020. *The algorithmic panopticon at Deliveroo: Measurement, precarity, and the illusion of control*. In: *Ephemera* 20 (3): 67–95.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

GLOBAL LABOR BEHIND THE DIGITAL INTERFACE

ALIENATION AND AGENCY OF MTURK WORKERS

Kassem, Sarrah
University of Tübingen
Tübingen, Germany
sarrah.kassem@uni-tuebingen.de

KEYWORDS

MTurk; digital labor; digital platforms

ABSTRACT

The workers of Amazon Mechanical Turk (MTurk), who labor for the digital labor platform, are situated in their own material realities that are stretched across the globe. While they labor microtasks that produce data later used for machine learning algorithms to train AI, they are located in their own local and national contexts. MTurk is a fascinating case to examine, given that it poses one of the platforms within Amazon's growing ecosystem where migrant labor is integral to its functioning. The question arises how does the specific organization of MTurk, being both web-based and a gig platform, alienate workers and how does this relate to the ways by which workers organize? Essentially, it is argued that racialized realities and divisions of labor are present on the digital shopfloor level and relate to the ways by which workers are estranged from their work and fellow humans, while these realities also relate to the ways by which they organize in alternative forms.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

THE IMPACT OF HATE SPEECH ABOUT REFUGEES ON POLITICAL ATTITUDES

EVIDENCE FROM AN ONLINE EXPERIMENT ON SEARCH ENGINES

Pradel, Franziska
Technical University of Munich,
Department of Governance,
School of Social Sciences and Technology,
Munich, Germany
franziska.pradel@tum.de

KEYWORDS

search engines; hate speech; trust; backfire effect; polarization; radicalization; attitudes towards refugees; algorithmic bias

ABSTRACT

Based on an online experiment, I assess the effects of hate speech compared to positive and neutral speech about refugees in search engines on trust and policy preferences. The experiment varies the tone of the suggestions (control, positive, neutral, negative) and the source of the suggestions (search engine and politician). The study provides insights into the polarizing potential of hate speech among individuals self-identifying at the left and right margins of the political spectrum. There are fundamentally different effects of positively biased information, in which persons with such group identities are much closer in their attitudes than persons exposed to other refugee-related information. Furthermore, search engines are perceived as politicized when they are politically biased, and the general trust in the source and its content erodes and is similar to the level of a typical politicized source (i.e., a politician). These findings are particularly alarming because the study shows that people with a right-wing political ideology are almost three times more likely to click on hate speech suggestions than those with a left-wing political ideology. Thus, especially strong political group identity plays a crucial role in how politically biased information influences political attitudes and how individuals engage with it online.

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

ALGORITHMS OF WAR

AFFECTIVE AFFORDANCES OF RECONTEXTUALISED WAR ON TIKTOK

Primig, Florian
Freie Universität Berlin
Berlin, Germany
florian.primig@fu-berlin.de

KEYWORDS

TikTok; war; affordances

ABSTRACT

We live in an "era of becoming a witness" (Givoni, 2011, p. 165), that is under the digitalized conditions of our contemporary knowledge society, the boundaries of what we can and ought to know have widened. Importantly, perceiving and performing oneself as a witness or knower online is bound to the socio-technical conditions of platforms, their affordances and logics (Szulc, 2018; van Dijck & Poell, 2013). Memetic performance (Zulli & Zulli, 2022) or remixing and recontextualization (Primig et al., 2023) is the creative heartbeat of TikTok. Russia's full-scale invasion of Ukraine is reimagined under this paradigm that transgresses "arrested war" (Hoskins & O'Loughlin, 2015) and recenters users' creative contributions and the power of platforms. Paying attention to the subtle power of platforms exercised through creators' algorithmically facilitated creative performances reveals discoursive practices of recontextualization within emerging affective networks of sound and images (Primig, Szabó & Lacasa, 2023, p. 7) where every topic, idea or belief can easily be hemmed into ever new contexts of trending online self-performance. Distant suffering is thus brought closer, but platform affordances also take on a stronger role in the interpretive struggles of war as they (co-)determine users' affective repertoires with trend templates and algorithmic curation: A development worth studying further.

REFERENCES

Givoni, M. (2011). Witnessing/testimony. Mafte'akh, 2(1), 147–170. https://cris.bgu.ac.il/en/publications/witnessingtestimony

Hoskins, A., & O'Loughlin, B. (2015). Arrested war: the third phase of mediatization. Information, Communication & Society, 18(11), 1320–1338. https://doi.org/10.1080/1369118X.2015.1068350

Primig, F., Szabó, H. D., & Lacasa, P. (2023). Remixing war: An analysis of the reimagination of the Russian–Ukraine war on TikTok. Frontiers in Political Science, 5, Article 1085149. https://doi.org/10.3389/fpos.2023.1085149

Szulc, L. (2018). Profiles, identities, data: Making abundant and anchored selves in a platform society. Communication Theory, 29(3), 169–188. https://doi.org/10.1093/ct/qty031

van Dijck, J., & Poell, T. (2013). Understanding social media logic. Media and Communication, 1(1), 2–14. https://doi.org/10.17645/mac.v1i1.70

Zulli, D., & Zulli, D. J. (2022). Extending the Internet meme: Conceptualizing technological mimesis and imitation publics on the TikTok platform. New Media & Society, 24(8), 1872–1890. https://doi.org/10.1177/1461444820983603

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

DIGIDAN

CAN WE STILL DISTINGUISH BETWEEN HUMANS AND MACHINES?

Strasser, Anna
DenkWerkstatt Berlin
Berlin, Germany
annakatharinastrasser@gmail.com

KEYWORDS

digital replica; LLMs; indistinguishability

ABSTRACT:

Large language models (LLMs) are a booming field of AI research. LLMs are able to produce grammatically correct linguistic outputs with fluency, often similar to that of a human. Many of their outputs can hardly be distinguished from linguistic outputs created by humans, even though some demonstrate a lack of common sense. Neural networks have proven to outperform humans in games and practical domains based on pattern recognition. Now, we might stand at a road junction where artificial entities might enter the realm of human communication. I will invite the audience to experience how difficult it is to distinguish between human- and machine-generated text. To this end, I will present the DigiDan installation, a recorded Zoom call with three interactors, namely Anna Strasser (researcher), Daniel Dennett (origin), and DigiDan (digital replica). Whereby Dennett and DigiDan will just be visible by their initials and act only via chat. Then I will report on the making and evaluation of DigiDan – an LLM fine-tuned with the corpus of Daniel Dennett (Strasser et al., 2023; Schwitzgebel et al., 2023) and give an overview of ethical consequences, touching issues of the difficulty to recognize human authorship, copyright, new forms of plagiarism, counterfeits of individuals, and the spread of misinformation and toxic language.

REFERENCES

Schwitzgebel, E., Schwitzgebel, D., Strasser, A. (2023). Creating a Large Language Model of a Philosopher. Mind & Language, 1-22. https://doi.org/10.1111/mila.12466

Strasser, A., Crosby, M., Schwitzgebel, E. (2023). How far can we get in creating a digital replica of a philosopher? In R. Hakli, P. Mäkelä, J. Seibt (eds.), *Social Robots in Social Institutions*. Proceedings of Robophilosophy 2022. Series Frontiers of AI and Its Applications, vol. 366, 371–380. IOS Press, Amsterdam. https://doi.org/10.3233/FAIA220637

Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

AI AS YOUR CREATIVE PARTNER

THE POWER OF PROMPT ENGINEERING

Winters, Thomas
Department of Computer Science; Leuven.AI,
KU Leuven
Leuven, Belgium
thomas.winters@kuleuven.be

KEYWORDS

large language models; prompt engineering; GPT; text generation; humor generation

ABSTRACT

These days, we are surrounded by creative text generators like GPT that seem to be able to generate any text we want. But how do we ensure that AI truly aids us in overcoming our unique creative challenges? In this talk, we uncover the remarkable potential of "prompt engineering" – the art of enhancing our communication with AI. We dive into the world of autoregressive text generators, learn their inner mechanisms and which training phases they went through to get to the current state-of-the-art text generators. These insights help understand why certain prompt engineering techniques (such as few-shot prompting, role-prompting and chain-of-thought prompting) are able to outperform simpler prompting methods. We show how even some of AI's classic hard problems, such as humor generation, become even more within reach thanks to these large language models and their prompt engineering techniques.