Automated Decision-Making Systems for Behavioral Regulation: Understanding Perceptions and Behavioral Reactions

Carmen Loefflad

Technical University of Munich, Munich, Germany
carmen.loefflad@tum.de

Abstract

Automated decision-making systems for regulating human behaviors are increasingly deployed around the globe. This project investigates people's perceptions of and behavioral reactions to automated decision-making systems used for behavioral regulation; first, in experimentally replicated social scoring systems, and second, in survey-based studies on the Chinese Social Credit System (SCS). The objective of this project is to understand the overall legitimacy of these systems, the sources of harm stemming from these systems, and to identify future areas of research necessary to determine whether these systems can be ethically justified.

Introduction

A growing body of research investigates how humans perceive algorithmic decision-making (ADM) systems. Prior work mainly studies ADM systems used for *decision support*, e.g., when used in loan approval (Yurrita et al. 2023) or healthcare scenarios (Araujo et al. 2020; Lee and Rich 2021), in legal contexts (Araujo et al. 2020), for online recommendations (Araujo et al. 2020), or management decisions (Binns et al. 2018; Lee 2018). Situated in this broader research strand, my research focuses on ADM systems used for *regulating behaviors* (Yeung 2018). I empirically investigate how people perceive and how they react to such systems from a behavioral viewpoint.

In this context, I adopt an approach that is rooted in procedural justice theory (Tyler 2006), which allows to study perceptions and behavioral reactions from a multi-dimensional perspective. I apply this approach to investigate ADM systems for behavioral regulation from two angles. First, focusing on social scoring systems as an instance of ADM systems for behavioral regulation (Cristianini and Scantamburlo 2020), I experimentally investigate perceptions of and behavioral reactions to social scoring systems using a combination of experimental, scenario-based, and survey-based approaches. Second, I investigate perceptions and behaviors in a real-world social scoring system, namely the Chinese Social Credit System (SCS) (Chen and Grossklags 2022).

Research Framework

ADM Systems for Behavioral Regulation

In ADM systems for behavioral regulation, the decision of the system is a behavioral assessment, often in the form of a quantified *score* (Mau 2019). The score is issued to decision-makers in various fields to determine access to services or goods. As such, ADM systems for behavioral regulation rely on a quantified assessment of one's behaviors, which incorporates behavioral adaptions in a feedback loop (Cristianini and Scantamburlo 2020). The behavioral aspect differentiates ADM systems used for behavioral regulation from those used for decision support. Scholars refer to this practice also as algorithmic regulation (Yeung 2018; O'Reilly 2013). Algorithmic regulation should help achieve a broader society-wide goal, e.g., establishing a more trustworthy society (State Council 2014), in which people behave in a prosocial fashion (Yeung 2018). However, there are significant ethical concerns associated with ADM systems used for regulating behaviors, relating to their opacity, arbitrariness (Citron and Pasquale 2014), and potential disparate impact on different groups of people (European Commission 2021; Loefflad, Chen, and Grossklags 2024a). A main factor adding to this debate is the fact that these systems often violate contextual integrity. Contextual integrity requires that the score is used to make decisions only in domains that are related to the domain from which the score was sourced (Nissenbaum 2004). My research investigates the extent to which disparate impacts and the conceptually identified concerns arise (Zarsky 2016; Citron and Pasquale 2014). To this end, I focus on both perceptual and behavioral implications for those who are directly subjected to the decision-making of ADM systems for behavioral regulation.

Understanding Perceptions and Behavioral Reactions

To investigate people's perceptions and behavioral reactions to ADM systems used for behavioral regulation, I combine experimental methods with scenario-based and survey-based approaches. For the behavioral aspect, I study the prosociality of individuals subject to ADM systems used for behavioral regulation. For the perceptual aspect, I adopt an approach that is rooted in procedural justice theory, and investigate *perceptions of legitimacy, procedural justice, and*

effectiveness. I also account for people's *experiences*, in terms of how favorable they perceive their outcome (*outcome favorability*) (Brockner 2002; Brockner et al. 2003), as well as the degree of privacy invasion, which I refer to as *subjective privacy harms* (Calo 2011). I have established a research model to assess how perceptual and experiential factors shape the overall legitimacy of and people's behavioral reactions toward such systems. This research model also allows to examine how the relationship between perceptions, experiences, and behavioral aspects is impacted by external factors, for example, transparency (Loefflad, Chen, and Grossklags 2023).

Published Articles on Social Scoring Systems

In my experimental works, I replicated people's experiences with a social scoring system, which aimed to regulate people's *trustworthiness*, as an instance of prosocial behaviors. In this context, I investigated how a *system-level property*, namely the provision of transparency, affects people's perceptions of and behavioral reactions to a social scoring system, in which contextual integrity is *maintained*. I found that the determinants of legitimacy were strongly shaped by transparency; specifically, in a non-transparent system, perceptions of legitimacy were biased by people's outcome favorability (Loefflad, Chen, and Grossklags 2023). From a behavioral angle, transparency was important to ensure that individuals engage in the desired behaviors, and that communities equally benefit from the introduction of a social scoring system, by developing trust and a less pronounced disparate impact *in the aggregate*. However, transparency also increased the discriminatory actions against people with a low score (Loefflad, Chen, and Grossklags 2024b).

Following these findings, I investigated how an *individual-level property*, namely the types of consequences, in terms of the outcome (good vs. bad outcome) and the decision importance (high vs. low importance) shape perceptions and behavioral reactions in a social scoring system in which contextual integrity is *violated*. I found that the outcome created pivotal opinion differences between people with a good and people with a bad outcome, which were partly exacerbated by the decision importance. Moreover, the outcome changed the underlying motivation to comply with the system, suggesting that the types of consequences might induce very opposing dynamics (Loefflad and Grossklags 2024). My results also suggest that revealing a social score to receive benefits constitutes a strong privacy invasion (Loefflad, Chen, and Grossklags 2024b), specifically once the score is transparent and gives precise indications about the underlying behaviors (Loefflad, Chen, and Grossklags 2023).

Research on the Chinese SCS

The Chinese SCS is a multifaceted system, consisting of several branches at the national, provincial and local level (Chen, Engelmann, and Grossklags 2023). In my research, I focus on the local government-run SCS (Loefflad 2024). These local SCS evaluate citizens according to a broad range of aspects, including their prosocial behaviors. In China, most cities have regulations regarding a local SCS, but not all cities have launched a personal credit scoring system. Systems that issue a behavioral score provide benefits for people with a good score, and punishments for people with a bad score (Li and Kostka 2022; Engelmann et al. 2019; Chen, Engelmann, and Grossklags 2022; Engelmann et al. 2021). While data collection affects all individuals and organizations, the active utilization of local SCS systems is voluntary. As such, groups of people can be distinguished based on their involvement in local government-run SCS.

Following the perceptual framework (Loefflad, Chen, and Grossklags 2023), a large-scale survey was conducted with citizens in different cities in China. I investigated citizens' perceptions and experiences of the local SCS. For comparative purposes, I retrieved the *normative expectations* of groups of people who are not actively involved in local SCS. In addition, the survey shed light on the behavioral implications of local SCS; focusing on people's *compliance* and *engagement* as instances of prosociality, I assessed whether groups of people with different degrees of involvement in local SCS also differ in behaviors.

My preliminary results indicate that the local SCS cannot live up to what is normatively expected, specifically in terms of procedural justice and effectiveness. From a behavioral viewpoint, using local SCS systems is strongly associated with compliance and engagement. However, while the presence of a local SCS in a city raises the bottom line of prosocial behaviors, it also imposes an upper limit on prosocial behaviors, which might be due to a motivation crowding-out effect (Frey 2012).

In addition, following our prior work (Loefflad, Chen, and Grossklags 2023), I investigated the relationship between perceptions, experiences, and prosocial behaviors, focusing on users of local SCS only. My results reveal two notable differences compared to the scoring experiments conducted in a Western context. First, the strongly negative association between subjective privacy harms and perceptions, as revealed in my experimental works, could not be replicated among Chinese citizens. Second, perceptions of legitimacy of decision-making systems are commonly strongly enhanced by procedures that are perceived as just (Tyler 2006; Tyler and Lind 1992). While my experimental studies confirmed this association (Loefflad, Chen, and Grossklags 2023; Loefflad and Grossklags 2024), Chinese citizens' perceptions of legitimacy of local SCS were only weakly due to perceptions of procedural justice. Instead, these perceptions are primarily driven by instrumental considerations, such as the perceived effectiveness of the local SCS and the extent to which individuals receive favorable outcomes.

Future Work

My results offer evidence against the engrossing arguments made for the introduction of ADM systems for behavioral regulation (O'Reilly 2013). In this context, two strands of research need to be investigated in the future. First, it is important to identify further sources of disparate impact stemming from regulatory ADM systems. For example, similar to (Schoeffer, Kuehl, and Machowski 2022; Wang, Harper, and Zhu 2020), I found that people's AI literacy is positively

associated with procedural justice (Loefflad and Grossklags 2024). From a policy perspective, the disparate impacts arising from differences in AI literacy constitute a main concern (European Commission 2021), specifically as differences in perceptions of justice further shape the perceptual and behavioral dynamics emerging from ADM systems used for behavioral regulation (Loefflad, Chen, and Grossklags 2023; Loefflad and Grossklags 2024). A second strand of research may investigate how the identified disparate impacts can be mitigated. In this context, future work should elaborate on whether the opposing opinion differences that the outcome creates (Loefflad and Grossklags 2024) could be mitigated by adhering to the concept of contextual integrity.

Lastly, my work suggests that people's socio-cultural background shapes their perceptions of an ADM system used for behavioral regulation, for example, by altering the determinants of perceived legitimacy, or people's sensitivity to the privacy invasion of such systems. Future work should, therefore, identify the disparate impact that might arise due to differences in socio-cultural characteristics. This is specifically important once such systems are utilized in societies that are characterized by a strong cultural diversity. The questions of which additional disparate impacts may arise, as well as whether the identified concerns can be mitigated is key to informing policymakers on how to set the legal boundaries of ADM systems used for regulating behaviors.

Acknowledgments

I wish to thank Jens Grossklags, Mo Chen, Emmanuel Syrmoudis and Felix Fischer for their support. I also thank the study participants, as well as the facilitators of the MELESSA lab and the ExperimenTUM lab. I am grateful for funding support from the Bavarian Research Institute for Digital Transformation (bidt) and the Institute for Ethics in Artificial Intelligence (IEAI).

References

Araujo, T.; Helberger, N.; Kruikemeier, S.; and de Vreese, C. 2020. In AI We Trust? Perceptions about Automated Decision-making by Artificial Intelligence. *AI & SOCIETY*, 35: 611–623.

Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; and Shadbolt, N. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–14. ACM.

Brockner, J. 2002. Making Sense of Procedural Fairness: How High Procedural Fairness Can Reduce or Heighten the Influence of Outcome Favorability. *The Academy of Management Review*, 27(1): 58–76.

Brockner, J.; Heuer, L.; Magner, N.; Folger, R.; Umphress, E.; van den Bos, K.; Vermunt, R.; Magner, M.; and Siegel, P. 2003. High Procedural Fairness Heightens the Effect of Outcome Favorability on Self-evaluations: An Attributional Analysis. *Organizational Behavior and Human Decision Processes*, 91(1): 51–68.

Calo, R. M. 2011. The Boundaries of Privacy Harm. *Indiana Law Journal*, 86(3): 1132–1162.

Chen, M.; Engelmann, S.; and Grossklags, J. 2022. Ordinary People as Moral Heroes and Foes: Digital Role Model Narratives Propagate Social Norms in China's Social Credit System. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 181–191. ACM. ISBN 9781450392471.

Chen, M.; Engelmann, S.; and Grossklags, J. 2023. Social Credit System and Privacy. In Trepte, S.; and Masur, P. K., eds., *The Routledge Handbook of Privacy and Social Media*, 227–236. New York, NY: Routledge.

Chen, M.; and Grossklags, J. 2022. Social Control in the Digital Transformation of Society: A Case Study of the Chinese Social Credit System. *Social Sciences*, 11(6): 1–23.

Citron, D.; and Pasquale, F. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89(1): 1–33.

Cristianini, N.; and Scantamburlo, T. 2020. On Social Machines for Algorithmic Regulation. *AI & SOCIETY*, 35: 645–662.

Engelmann, S.; Chen, M.; Dang, L.; and Grossklags, J. 2021. Blacklists and Redlists in the Chinese Social Credit System: Diversity, Flexibility, and Comprehensiveness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 78–88. Association for Computing Machinery.

Engelmann, S.; Chen, M.; Fischer, F.; Kao, C.-Y.; and Grossklags, J. 2019. Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines "Good" and "Bad" Behavior. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency*, FAT* '19, 69–78. ACM.

European Commission. 2021. *A Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021) 206 final)*. European Commission.

Frey, B. S. 2012. Crowding Out and Crowding In of Intrinsic Preferences. In Brousseau, E.; Dedeurwaerdere, T.; and Siebenhuner, B., eds., *Reflexive Governance for Global Public Goods*. The MIT Press.

Lee, M. K. 2018. Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society*, 5(1): 1–16.

Lee, M. K.; and Rich, K. 2021. Who is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–14. ACM.

Li, H.; and Kostka, G. 2022. Accepting but not Engaging with it: Digital Participation in Local Government-run Social Credit Systems in China. *Policy & Internet*, 14(4): 845–874.

Loefflad, C. 2024. *On Human Reactions to Social Scoring Systems*. Ph.D. thesis, Technical University of Munich. Submitted.

Loefflad, C.; Chen, M.; and Grossklags, J. 2023. Factors Influencing Perceived Legitimacy of Social Scoring Systems: Subjective Privacy Harms and the Moderating Role of Transparency. In *Proceedings of the International Conference on Information Systems*, ICIS '23. AIS.

Loefflad, C.; Chen, M.; and Grossklags, J. 2024a. Reputational Discrimination and Fairness in China's Social Credit System. *Digital Governance: Research and Practice*.

Loefflad, C.; Chen, M.; and Grossklags, J. 2024b. Social Scoring Systems for Behavioral Regulation: An Experiment on the Role of Transparency in Determining Perceptions and Behaviors. In *Proceedings of the 7th AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24.

Loefflad, C.; and Grossklags, J. 2024. How the Types of Consequences in Social Scoring Systems Shape People's Perceptions and Behavioral Reactions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM.

Mau, S. 2019. *The Metric Society: On the Quantification of the Social*. Cambridge, UK: Polity Press.

Nissenbaum, H. 2004. Privacy as Contextual Integrity. *Washington Law Review*, 79(1): 119–157.

O'Reilly, T. 2013. Open Data and Algorithmic Regulation. In Goldstein, B.; and Dyson, L., eds., *Beyond Transparency: Open Data and the Future of Civic Innovation*, chapter 22, 289–300. San Francisco, CA, USA: Code for America Press.

Schoeffer, J.; Kuehl, N.; and Machowski, Y. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1616–1628. ACM.

State Council. 2014. *Planning Outline for the Construction of a Social Credit System (2014-2020)*. (in Chinese).

Tyler, T. R. 2006. *Why People Obey the Law*. Princeton, NJ, USA: Princeton University Press.

Tyler, T. R.; and Lind, E. 1992. A Relational Model of Authority in Groups. *Advances in Experimental Social Psychology*, 25: 115–192.

Wang, R.; Harper, F. M.; and Zhu, H. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–14. ACM.

Yeung, K. 2018. Algorithmic Regulation: A Critical Interrogation. *Regulation & Governance*, 12(4): 505–523.

Yurrita, M.; Draws, T.; Balayn, A.; Murray-Rust, D.; Tintarev, N.; and Bozzon, A. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–21. ACM.

Zarsky, T. 2016. The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology, & Human Values*, 41(1): 118–132.