Social Scoring Systems for Behavioral Regulation: An Experiment on the Role of Transparency in Determining Perceptions and Behaviors

Carmen Loefflad, Mo Chen, Jens Grossklags

Technical University of Munich
carmen.loefflad@tum.de, mo.chen@tum.de, jens.grossklags@in.tum.de

Abstract

Recent developments in artificial intelligence research have advanced the spread of automated decision-making (ADM) systems used for regulating human behaviors. In this context, prior work has focused on the determinants of human trust in and the legitimacy of ADM systems, e.g., when used for decision support. However, studies assessing people's perceptions of ADM systems used for behavioral regulation, as well as the effect on behaviors and the overall impact on human communities are largely absent. In this paper, we experimentally investigate people's behavioral adaptations to, and their perceptions of an institutionalized decision-making system, which resembled a social scoring system. Using social scores as incentives, the system aimed at ensuring mutual fair treatment between members of experimental communities. We explore how the provision of transparency affected people's perceptions, behaviors, as well as the well-being of the communities. While a non-transparent scoring system led to disparate impacts both within as well as across communities, transparency helped people develop trust in each other, create wealth, and enabled them to benefit from the system in a more uniform manner. A transparent system was perceived as more effective, procedurally just, and legitimate, and led people to rely more strongly on the system. However, transparency also made people strongly discipline those with a low score. This suggests that social scoring systems that precisely disclose past behaviors may also impose significant discriminatory consequences on individuals deemed non-compliant.

Introduction

Practices to score private individuals and consumers have been developing continuously within the credit sector over the past decades. Yet, with the rise of AI and Big Data, the scoring of individuals and consumers finds application in an increased range of areas (Citron and Pasquale 2014; Mau 2019). An emergent trend of digital governance is the use of automated decision-making (ADM) systems for regulating society-wide behaviors (Vogl et al. 2020), also referred to as *algorithmic regulation* (Yeung 2018; O'Reilly 2013). Social scoring systems, which use behavioral data to calculate a social score, are an instance of ADM system. The score figures as an automated incentive for encouraging "good" behaviors (Cristianini and Scantamburlo 2020).

Social scoring systems have raised substantial concerns relating to their disparate impact and discriminatory consequences (Citron and Pasquale 2014; Zarsky 2016; Packin and Lev-Aretz 2016). Moreover, social scoring systems are often characterized by a lack of transparency regarding their scoring mechanisms or the behaviors they seek to incentivize (Engelmann et al. 2019; Citron and Pasquale 2014). In the context of transparency, early research has called for providing full transparency in ADM systems (Zarsky 2016). Yet, more recent studies have shown that the provision of transparency is not always beneficial, as it may induce privacy-related harms (Loefflad, Chen, and Grossklags 2023), or impair people's trust in ADM systems (Kizilcec 2016). While some consider social scoring systems a promising tool for algorithmic regulation (O'Reilly 2013), the associated concerns necessitate further research. In particular, from an empirical perspective, the implications of social scoring systems are not yet well understood.

In social scoring systems, the degree of regulation depends on people's participation in and interaction with the system (Cristianini and Scantamburlo 2020). While current research has investigated the factors that determine people's perceptions of fairness (Wang, Harper, and Zhu 2020; Lee and Baykal 2017; Lee and Rich 2021), as well as their trust in ADM systems (Park et al. 2019; Yin, Wortman Vaughan, and Wallach 2019), the challenges associated with human involvement in data-driven scoring systems have attracted only small attention (Watkins et al. 2021). However, these assessments are needed for identifying and mitigating potential sources of harm (Barocas and Selbst 2016). Specifically, there is a lack of empirical research to assess whether ADM systems can serve as legitimate mechanisms for behavioral regulation (Watkins et al. 2021).

In this work, we use an empirical approach to investigate how the provision of transparency affects people's perceptions and behaviors in a social scoring system. Thereby, we connect survey measures with experimental behaviors, which can contribute to a more comprehensive understanding of complex phenomena (Glaeser et al. 2000). Specifically, we focus on a social scoring system in which contextual integrity is maintained, as required by the EU AI Act (European Commission 2021). Contextual integrity implies that the score is used for decision-making only in a context that is related to the context from which the score

was derived (Nissenbaum 2004). We conceptualize social scoring systems as instances of ADM systems, combining characteristics of reputation systems and order institutions. We experimentally replicated a social scoring system in the context of a community game of trust. The system aimed to regulate people's trustworthiness behaviors, and to build trust among unknown members of communities. Using a between-subject design, we ran two treatments, which differed in the level of transparency regarding the scoring mechanism. In a post-experimental survey, we investigated people's perceptions of the system. For reference purposes, we ran a baseline treatment, in which a normative intervention system without scores was applied. As such, we address two research questions. First, how do different levels of transparency in a social scoring system affect participants' trust and trustworthiness behaviors within communities of strangers? Second, does the provision of transparency disparately affect people's perceptions of a social scoring system? To answer these questions, we compare people's trust and trustworthiness behaviors between treatments, and elaborate on the determinants of trust in the scoring interventions. We illustrate the impact of the different scoring systems on a community level, examining the well-being of communities in terms of wealth and inequality. We further compare people's perceptions across treatments and examine whether clusters of communities that differ in wealth and inequality exhibit different perceptions. Moreover, we relate people's perceptions to their experiences, in terms of being disciplined by others based on their social score. We also assess whether discriminating against people with a low score is associated with perceptions.

We find that transparency enhanced both trust and trustworthiness in experimental communities. However, people's trust decisions were more strongly conditioned on the score of the interaction partner in a transparent scoring system, suggesting that transparency increased the discriminatory impact on people with a low score. The increased trust and trustworthiness of communities subject to a transparent system resulted in more wealth, as well as lower levels of inequality among community members. Further, people reported more positive perceptions of procedural justice and legitimacy in a transparent social scoring system. Communities subject to a transparent scoring system showed similar perceptions of legitimacy, procedural justice, and effectiveness, even if they differed in their wealth and inequality. In a non-transparent (opaque) system, in contrast, communities with lower levels of wealth reported less positive perceptions. Further, people's use of the score as a basis for their trust decision was tied to perceptions of justice *only* in the non-transparent system. At the same time, the experience of being disciplined on the basis of a score was associated with less positive perceptions of justice, effectiveness, and legitimacy *only* in the non-transparent system.

From a higher-level perspective, these results suggest that transparency is important to achieve that individuals engage in desired behaviors, develop positive perceptions towards the system, and that perceptions are not biased by personal experiences or the macro-economic development of the communities. However, people with a low score were strongly disciplined. Our results thus further imply that on an individual level, transparency may impose significant harms on low-scoring individuals. Thus, while transparency is key to ensuring that communities are not unduly harmed in the aggregate, the discriminatory impacts they create on an individual level should be considered cause for concern.

Our results provide exploratory empirical insights into the ongoing debate on the disparate impact stemming from social scoring systems (Citron and Pasquale 2014; Barocas and Selbst 2016). Our findings underscore the need to take the concerns associated with these systems seriously, especially given the arguments made for the introduction of social scoring systems, and the benefits that their planners hope to achieve (O'Reilly 2013). More generally, our results also point to the direction that the introduction of social scoring systems is worth questioning.

Background and Related Work

In this section, we portray social scoring systems as an instance of ADM systems combining the characteristics of a reputation system and regulatory order institution. In addition, as our work focuses on social scoring systems for building trust and trustworthy behaviors, we elaborate on the concepts of trust and trustworthiness, as well as on the sources of these concepts.

Social Scoring Systems as Algorithmic Regulation Systems

The basic idea of algorithmic regulation with social scoring systems is to govern population-wide behavior in a society using social scores as incentives (Cristianini and Scantamburlo 2020; O'Reilly 2013). Social scores are computed on the basis of behavioral data. Thereby, social scores should give an indication of peoples' traits, e.g., their trustworthiness or prosociality. As such, social scores can also be considered a form of reputation. In Europe, scoring systems are, for example, utilized to assess individuals' pro-environmental behaviors (Boos 2022; Wolfangel 2022). Regulators take these scores as a basis for administering interventions by distributing automated incentives. With this approach, regulators aim at influencing population-wide behaviors, such that a specific goal, which allegedly enhances the well-functioning of a society, is achieved (Yeung 2018; O'Reilly 2013). Traditionally, order institutions, such as courts or the police, have centered on ensuring the well-functioning of society. Nowadays, the availability of large-scale data and sophisticated classification techniques can entirely change the nature of regulatory institutions (O'Reilly 2013; Vogl et al. 2020; Cristianini and Scantamburlo 2020). Hereby, social scoring systems are by some considered a promising form for algorithmic regulation (O'Reilly 2013). Due to their multifaceted nature, scoring systems can thus be understood as an instance of ADM systems combining characteristics of both order systems and reputation systems.

Social scoring systems raise considerable ethical concerns (Citron and Pasquale 2014; Zarsky 2016). The EU AI Act classifies social scoring systems as systems creating an "unacceptable risk", and aims to prohibit social

scoring practices that evaluate individuals' trustworthiness, as this could lead to an unjustified disparity among different groups of people. Moreover, the EU AI Act requires that systems maintain contextual integrity (European Commission 2021). Contextual integrity requires that decision-makers use the score for making decisions only in a domain that corresponds to the domain in which the behavioral score was calculated (Nissenbaum 2004). However, the same-context requirement is difficult to operationalize in reality, and strongly depends on how narrowly the context is defined (Veale and Zuiderveen Borgesius 2021). A further main factor fueling the concerns associated with social scoring systems is the fact that automated decisions are characterized by a lack of transparency. Lacking knowledge about how the system functions could inhibit practitioners from judging the accuracy of a system, or the extent to which disparate impacts could emerge (Citron and Pasquale 2014).

In this work, we empirically investigated how individuals' behaviors in a social scoring system, as well as their perceptions of a social scoring system are impacted by different levels of transparency. Specifically, we replicated a social scoring system in which contextual integrity was maintained. We focused on trustworthiness as an exemplary target behavior of a social scoring system, and analyzed whether social scores can establish trustworthy behaviors, and build trust among members of experimental communities, varying the level of transparency. Using a trust game provides the opportunity to study both people's trustworthiness, as an instance of prosocial behaviors, as well as their trust in each other. Trust is crucial because it is the foundation of social and economic interactions, and likely affected by social scoring systems. In addition, using a trust game allows us to replicate the economic payoff of trust on a community level.

In the following, we first explain the concepts of trust and trustworthiness. Further, considering social scores as a form of reputation, we elaborate on how reputation can help build trust and trustworthiness. Lastly, we contextualize the implications of social scoring mechanisms within the literature that explores the sources of trust and trustworthiness.

Installing Trust and Trustworthiness with Social Scores

As a moral principle, trust refers to individuals' belief that other members of a society would not act to hurt them, even if they had the chance to do so, and even if they had a different set of values (Fukuyama 1995). More broadly speaking, trust refers to a general faith in other people (Fukuyama 1995; Uslaner 1999), reflecting a belief that most individuals are trustworthy. Trustworthiness thus implies that people do not act against the interest of those who have put faith in them (Uslaner 1999).

Trust is commonly associated with economic benefits, as it facilitates transactions with others (Collier 2002). As such, trust also has a strategic dimension, as it is also considered an instrument to extract market and non-market returns (Hardin 1993). In addition, high-trust societies are characterized by an increased welfare compared to low-trust societies, and commonly have lower levels of inequality between members of a society (Knack and Keefer 1997). While current social scoring efforts primarily target small-scale behaviors, they may also be used to address fundamental issues such as trust and trustworthiness (State Council 2014; Engelmann et al. 2021). In this case, it is reasonable to expect that the introduction of a society-wide social scoring system impacts the economic development of a society. In this context, the following section explains the mechanisms by which social scoring systems can impact trust and trustworthiness through their role as reputation systems.

Effects of Reputation Reputation is a powerful mechanism to steer people's behavior in a prosocial direction (Kandori 1992; Camera and Casari 2009; Nowak and Sigmund 1998; Bohnet and Huck 2004), and to build trust among strangers (Zacharia and Maes 2000; Friedman, Khan Jr, and Howe 2000; Aljazzaf, Perry, and Capretz 2010; Duradoni et al. 2018). In situations in which someone is insecure about whether an interaction partner can be trusted, a reputational score that approximates someone's past trustworthiness facilitates the decision of whether to trust. As such, the score is taken as a basis for decision-making, which implies that people with different scores are trusted to different extents. In the context of social scoring systems, this mechanism is considered a form of discrimination (Zarsky 2015). In the long term, and provided that trust is associated with economic benefits, this mechanism makes it in people's rational interest to behave trustworthy (Bohnet and Huck 2004; Engelmann and Fischbacher 2009).

Social scoring systems assess people's past trustworthiness through a behavioral score and provide it to decision-makers as a basis for their choices. The automated nature of social scoring systems thus lies in the calculation of a behavioral score. In this context, the Court of Justice of the European Union has decided that the provision of a credit score already constitutes an automated decision according to the GDPR Article 22(1), as the score plays a central role in the decision-making, for example, by landlords or banks (Court of Justice of the European Union 2023).

As ADM systems are socio-technical artifacts (Kitchin 2017), it is inevitable to consider the broader societal structures into which ADM systems are placed (Araujo et al. 2020). Therefore, we contextualize the effect mechanisms of social scoring systems trying to establish trustworthiness and trust within the broader mechanisms by which a society builds up these concepts.

Sources of Trust and Trustworthiness. Social capital theories portray shared values and norms, which are learned through repeated experiences with others, as central to the development of trust and trustworthiness in a society (Coleman 1990; Putnam, Leonardi, and Nanetti 1993; Brehm and Rahn 1997; Uslaner 2002; Sobel 2002). Thereby, people's involvement in associations is key for developing a common set of values and norms (Putnam, Leonardi, and Nanetti 1993). This account based on social capital illustrates trust and trustworthy behaviors from a moralistic perspective; trustworthy behaviors emerge from a moral conviction (Uslaner 1999), which may also be influenced by the belief that others will behave trustworthy.

An alternate explanation of how societies can build trust and trustworthiness is offered by an institutional account. Here, the presence of a sound institutional infrastructure, establishing mechanisms that enforce contracts (Hardin 1993), and thus address the "risky aspect of trusting strangers" (Steinhardt 2012) can provide a fruitful basis for trust and trustworthiness (Letki 2006; Uslaner 2002). This account illustrates trust from a strategic perspective, as trust is established through a belief that people who are not trustworthy will be detected and punished. Therefore, people will refrain from behaving untrustworthy in the first place (Rothstein 2000). However, only institutions that are perceived as legitimate achieve that people voluntarily engage in desired behaviors (Tyler 2006a; Levi, Sacks, and Tyler 2009).

If social scoring systems are introduced for building trust and trustworthiness, they operate according to the institutional account, as they provide a monetary incentive to behave trustworthy and to trust. Yet, the significance of legitimate mechanisms in shaping specific behaviors, which social psychology literature has emphasized (Tyler 2006b; Levi, Sacks, and Tyler 2009), suggests that if the operating mechanisms of social scoring systems are perceived as illegitimate, the mechanisms that build trust and trustworthiness are unlikely to come into play. Consequently, it becomes less likely that the social scores are considered reliable for decision-making. Thus, to anticipate the effect mechanisms of social scoring systems it is crucial to also assess people's perceptions of a system.

In AI Ethics, the perceived fairness of ADM systems constitutes a major concern (Hagendorff 2020). While there is consensus that these systems should serve the common good (Whittlestone et al. 2019), interpretations of fairness and justice vary widely (Birhane et al. 2022). Early studies often assess fairness in ADM systems in a one-dimensional way (Wang, Harper, and Zhu 2020). In this paper, we follow studies recognizing the multifaceted nature of fairness (Loefflad, Chen, and Grossklags 2023; Loefflad and Grossklags 2024; Juijn et al. 2023; Yurrita et al. 2023) and assess perceptions of legitimacy, procedural justice, and effectiveness. Generally, perceptions are likely affected by different system-level properties; one of the most widely discussed system-level factors of ADM systems is their level of transparency (Ehsan et al. 2021; de Fine Licht et al. 2014; Rader, Cotter, and Cho 2018; Schmidt, Biessmann, and Teubner 2020).

Transparency and Contextual Integrity

The EU AI Act aims to prohibit social scoring practices that violate contextual integrity (European Commission 2021). In our prior research, we have shown that a violation of contextual integrity in social scoring systems leads to strong opinion differences between people with good scores and people with bad scores (Loefflad and Grossklags 2024). In this work, in contrast, we investigate a social scoring system in which contextual integrity is *maintained*, but focus on the implications of providing transparency. The EU AI Act imposes increased transparency requirements on systems that impose a high risk, which should "enable users to interpret the system's output and use it appropriately" (European Commission 2021). However, the provision of transparency has ambiguous implications; early work has imperatively called for providing transparency in ADM systems, to help increase the interpretability of decisions (Rader, Cotter, and Cho 2018), or identify arbitrary evaluations, wrong characterizations, and biases (Citron and Pasquale 2014). In addition, there has been a general belief that transparent systems are more fair and trustworthy (Zarsky 2016). However, the transparency ideal has been challenged (Schmidt, Biessmann, and Teubner 2020; Ananny and Crawford 2018); the effects of transparency on perceptions of fairness or trust are contingent on personal characteristics (Eslami et al. 2019), expectations (Kizilcec 2016), or people's control within the decision-making processes (Lee et al. 2019). Transparency may even prove harmful due to its complex interaction with privacy. It is plausible that increased information about the underlying behavior of a score constitutes a violation of privacy (Chen, Engelmann, and Grossklags 2023; Chen and Grossklags 2022; Ananny and Crawford 2018) and imposes privacy harms on decision subjects (Loefflad, Chen, and Grossklags 2023). More studies are thus needed to carefully assess the impacts of transparency in social scoring systems on individual behaviors, perceptions, as well as on society-level implications (Watkins et al. 2021; Chen et al. 2023). We address this gap with an experimental investigation of the implications of transparency in social scoring systems that try to establish trust and trustworthiness.

Method

The experiment consisted of three parts: a pre-survey, a community game of trust, and a post-survey. All parts of the experiment were coded in oTree (Chen, Schonger, and Wickens 2016), which is a standard software for economic experiments. Our institution does not require ethics approval for experiments and questionnaire-based studies. However, we followed standard practices for ethical research, e.g., presenting detailed study procedures, obtaining consent, allowing participants to leave the study at any time, and collecting anonymized data. The sessions were conducted in the ExperimenTUM lab at the Technical University of Munich in August and October 2022. Each session was subject to one of three treatments.

156 people participated in the experiment, most of whom were university students. The analysis contains data from 148 participants who passed the attention check (47 in the baseline, 51 in the transparent, and 50 in the non-transparent treatment). 59.5% of the participants were aged between 17 and 24, 34.5% were aged between 25 and 30, 4.1% were older than 30, and 2.0% did not state their age. 48.0% of the participants were female. 57.4% were White-Caucasian, 27.0% Asian-Pacific, and 4.7% Hispanic. 4.7% stated to be multiracial, and 6.1% did not reveal their ethnicity. An overview of the sessions is given in Table 2 in the Appendix (Loefflad, Chen, and Grossklags 2024).

The present paper uses data from our prior work (Loefflad, Chen, and Grossklags 2023). In this prior work, we have established a structural model to assess the perceived legitimacy of social scoring systems, and to illustrate how perceptions of legitimacy are shaped by the level of transparency (Loefflad, Chen, and Grossklags 2023). The present

paper focuses on the behavioral aspects during the experimental phase. It further investigates macro-economic impacts of trust, and relates people's behaviors in the experimental phase to their perceptions.

Pre-survey

We presented participants with two hypothetical decision scenarios, a standard trust game, and a dictator game, to collect people's general trust in others, and their kindness, respectively. By accounting for people's kindness and their general trust, we establish a measure for people's personal social values, which are important in determining trust and trustworthiness behaviors (Coleman 1990; Putnam, Leonardi, and Nanetti 1993; Uslaner 2002; Sobel 2002).

In a basic trust game with two participants (Berg, Dickhaut, and McCabe 1995), a "first mover" decides what share of a given budget (i.e., the "endowment") to send to a "second mover". The amount sent by the first mover is multiplied by a known factor when it is received by the second mover. The second mover then decides on an amount to return to the initial sender. The amount sent figures as a measure of *trust*; the return ratio, i.e. the fraction of the amount received that is returned, constitutes a measure of *trustworthiness*. Notably, due to the multiplication factor, trusting created wealth; the first mover decides how much wealth is created, the second mover decides about the distribution of wealth. In the trust game with equal endowments of 10 monetary units (MUs) and a multiplication factor of two, we asked participants how much they would send to their interaction partner in the role of the first mover. As such, we derived a measure of people's general trust in others (Glaeser et al. 2000).

A dictator game is typically played between two persons, who receive each a monetary endowment. Person A receives more than person B, which is known to both. Person A decides how much to give to person B. This amount indicates people's kindness. We asked participants how much they would give to an unknown person B, if they were person A and had received 10 MUs; person B had received 0 MUs.

Community Game of Trust

In the second part, each participant was randomly assigned to a community of strangers, consisting of four people. The members interacted with each other in a repeated trust game (as explained in the previous section) over several rounds. In each round, the four members of each group were randomly matched to pairs of two and played a trust game (Berg, Dickhaut, and McCabe 1995). In each interaction, the roles of the first and the second mover were determined randomly. Both players received an initial endowment of 10 MUs. We equipped participants with equal endowments because many people are inequity-averse (Fehr and Schmidt 1999; Bolton and Ockenfels 2000) and equal endowments avoid situations in which preferences for equality undermine trustworthy behavior (Xiao and Bicchieri 2010; Aksoy et al. 2018). Once the interactions in one round ended, a die was rolled to determine whether the experiment continued to another round. The continuation chance was 0.95, resulting in an expected number of rounds of 20. The experiment was conducted in an indefinitely repeated frame to prevent a change of behaviors in later rounds.

Treatments Each session was subject to one of three treatments. In each treatment, an institutional intervention system was introduced, with the purpose of incentivizing people to trustworthy behaviors, which we framed as "fair" behaviors. Fairness meant that the payoffs between the first and second movers, resulting from one interaction, were equal. We ran two interventions using scoring systems, which differed in the level of transparency regarding the scoring mechanism. The scoring interventions operated according to the institutional account, in which a scoring system scored participants based on their behaviors. The scoring system made it possible for the trustors to request the score of their interaction partners. The first movers could thus form an expectation about the trustworthiness behaviors of the trustees based on the behavioral score, and condition their trust choice accordingly. As such, a monetary incentive was provided to behave trustworthy and to trust.

As outlined, the social-capital-based account treats people's values and norms as central to the development of trust and trustworthiness, which people learn through their involvement in associations (Putnam, Leonardi, and Nanetti 1993). We replicated this account in a baseline intervention. In the baseline intervention, normative messages, which reminded people in a community to treat each other fairly were displayed. As such, we mimicked participants' involvement in a common association that values norms of fairness.

In the instructions, participants in all treatments learned that their behaviors were monitored by an institution with the purpose to ensure fairness (see Table 6 in the Appendix (Loefflad, Chen, and Grossklags 2024)).

Behavioral Evaluation and Scores In the scoring interventions, a scoring system evaluated the back-sending behaviors of the second movers. We categorized the second movers' back-sending behaviors along three thresholds. First, people generally are inequity-averse, and find it fair to send back an amount that equalizes the payoffs between the first and second mover ($x_{equalizing}$) (Xiao and Bicchieri 2010; Aksoy et al. 2018). Due to the equal endowments, the back-sending behaviors always have the order $x_{equalizing} \geq x_{half} \geq x_{third}$. Using these thresholds, the back-sending behaviors were classified into four categories. Any amount that equaled at least $x_{equalizing}$ was considered very trustworthy, and any amount between $x_{equalizing}$ and x_{half} as trustworthy. Any amount returned between x_{half} and x_{third} was considered untrustworthy. Any amount returned less than x_{third} was considered very untrustworthy (Table 3; Appendix (Loefflad, Chen, and Grossklags 2024)).

Over the course of the experiment, each participant's trustworthiness as a second mover was aggregated to a reputational signal. The signal was framed as *standing*. The standing ranged from A to E, where A was the highest, and E was the lowest standing. The standing was based on a numerical score between 1000 to 0. Everyone started with a numerical score of 1000 and a standing of A. The standing could be lost in case the fairness principle was not respected. In this case, a point penalty was introduced, re-

flecting the severity of misbehavior, using the classification of behaviors into four categories (Table 3 in the Appendix (Loefflad, Chen, and Grossklags 2024)). For example, receiving 20 MUs and sending back 0 MUs led to a rounded point penalty of 38 points; four points for each MU retained up to one-third of the received amount, two points for each MU retained between one-third and one-half of the received amount, and one point for each MU retained between one-half of the received amount and the payoff-equalizing amount. After each round, the resulting numerical scores and standings were updated. The score ranges were set after simulating the evolution of scores under this point penalty mechanism, using trust game data from (Berg, Dickhaut, and McCabe 1995) (Figure 5; Appendix (Loefflad, Chen, and Grossklags 2024)). Participants received hints about the definition of fairness (Table 6; Appendix (Loefflad, Chen, and Grossklags 2024)). We included comprehension questions to ensure that participants were familiarized with the experimental procedures and the implications of their decisions. In case of wrong answers, participants were educated about the experimental setup again.

Transparency and Institutional Intervention. In the scoring treatments, participants knew that their behavior as a second mover was evaluated. In the non-transparent treatment, people were only informed about their standing, not about the underlying score. In the transparent treatment, both the standing, the score, as well as the score ranges (Table 4 in the Appendix (Loefflad, Chen, and Grossklags 2024)) were known to the participants. They were *not precisely* educated about the point penalty mechanism, but hinted towards it, as in Loefflad and Grossklags (2024). In the scoring treatments, the first movers had the possibility to request the standing, but not the score of their interaction partners before deciding whether to trust.

In the scoring treatments, an intervention happened whenever the standing of the second mover was lowered. In this case, the trustee received the information that their standing had been lowered, and was reminded to treat others in a fair manner (Table 7 in the Appendix (Loefflad, Chen, and Grossklags 2024)). In the baseline setting, the evaluation mechanism was *hypothetically* calculated; an intervention happened whenever the *hypothetical* standing was lowered. In this case, the affected second mover was reminded to treat others fairly.

As such, the experiment provides a framework for studying individual-level trust interactions between unknown members of a community, in which an institutional scoring system is implemented to facilitate the development of trustworthy behaviors and trust. Moreover, due to the repeated horizon, our framework also allows us to study the development of trust and trustworthiness in the communities over time, and to assess the macro-economic development of the communities, in terms of achieved wealth and inequality levels between community members. Running a baseline mechanism next to the scoring treatments allows us first, to test whether trust and trustworthiness arise from repeated experiences absent any material incentive, and second, to contrast the development of trust and trustworthiness of communities with a scoring system to that of communities without a scoring system, experiencing a much less intrusive intervention.

Post-experimental Survey

After the community game of trust, we administered a comprehensive survey to understand people's perceptions of the institutional intervention systems. The survey measured factors that are used to explain people's acceptance of traditional regulatory institutions (Mazerolle et al. 2013; Tyler and Jackson 2013; Tyler and Fagan 2008; Brockner 2002; Alessandro et al. 2021) and ADM systems (Lee 2018; Lee et al. 2019; Wang, Harper, and Zhu 2020; Martin and Waldman 2022; Yu and Li 2022; Castelo, Bos, and Lehmann 2019). To these factors count perceptions of procedural justice, legitimacy, outcome favorability, and effectiveness.

The legitimacy scale was measured via the extent to which the fairness goal of the institutions was aligned with participants' own viewpoints (2 questions), participants' perceived obligation to behave according to the postulated goal (1 question), and their perceived trust in the institution (1 question) (α=0.81) (Jackson et al. 2023). Perceived procedural justice addressed the respect people felt they were given, the clarity of the scoring mechanism, people's control in the scoring process, and the generally perceived fairness of the system (Lee et al. 2019; Tyler and Jackson 2013) (α=0.9). Perceived effectiveness relates to participants' perception of how well the institution succeeded in establishing the fairness goal, and to the perceived quality and usefulness of decisions (α=0.85). Perceived outcome favorability (4 questions) measured the extent to which participants liked how they were treated by others, to which participants were satisfied with their respective experimental outcomes, and to which the presence of the intervention system was perceived as favorable (α=0.77). Two questions measured participants' trust perceptions; one question asked how well people could trust others, and another question asked how strongly people felt they were trusted by others. An attention check was added in the last third of the survey. All questions were measured on a 5-item Likert scale. An overview of the questions is given in Table 5 in the Appendix (Loefflad, Chen, and Grossklags 2024).

Results

Trust and Trustworthiness Behaviors

In this section, we first report on the differences in behaviors between the treatments. Subsequently, we assess how the behavioral score determined people's trust decisions.

Between-Treatment Differences Using one-sided Mann-Whitney *U* tests, we find that people's trustworthiness was significantly higher in the transparent scoring treatment (*M*=.67, *SD*=.13) as compared to the non-transparent scoring treatment (*M*=.59, *SD*=.23, $p<.001$). Compared to the baseline intervention (*M*=0.52, *SD*=0.28), participants' trustworthiness was significantly higher in both the transparent ($p<.001$) as well as the non-transparent scoring system ($p<.001$). People's trust levels were significantly higher in the transparent treatment (*M*=7.92, *SD*=3.16) compared to

the baseline intervention (M=6.41, SD=4.24, $p<.001$) and the non-transparent scoring treatment (M=6.41, SD=3.89, $p<.001$). However, no significant difference in trust between the non-transparent and baseline treatments was found (see Figure 3; Appendix (Loefflad, Chen, and Grossklags 2024)). The increased trustworthiness in the transparent treatment is reflected in the distribution of final standings (Figure 4; Appendix (Loefflad, Chen, and Grossklags 2024)). In the transparent intervention, the majority of participants remained very trustworthy and thus had a standing A at the end of the experiment. The distribution of final standings is similar in the baseline and the non-transparent treatments. In the baseline and the non-transparent treatment, a considerable fraction of participants were classified in the lowest classes D and E (39% in the baseline, and 33% in the non-transparent system). Compared to the transparent scoring system, only a few participants remained very trustworthy, with a standing of A (36% in the baseline, and 31% in the non-transparent system). Lastly, people requested the score significantly more frequently in the opaque scoring system (M=.71, SD=.30) than in the transparent scoring system (M=.60, SD=.37) ($p<.001$).

As indicated by the intra-class correlation coefficients (ICC), participants' overall trusting behavior over time was more stable in the baseline treatment ($ICC^{trust}_{base} = 66\%$) as compared to the scoring treatments ($ICC^{trust}_{trans} = 48\%$, $ICC^{trust}_{opaque} = 49\%$). Reversely, people's trustworthiness (tw) behaviors over time were more homogeneous in the scoring treatments ($ICC^{tw}_{trans} = 60\%$, $ICC^{tw}_{opaque} = 59\%$) compared to the baseline treatment ($ICC^{tw}_{base} = 56\%$). Participants' trusting behaviors towards partners with the same standing were substantially more homogeneous in the transparent intervention ($ICC^{standing}_{trans} = 64\%$), compared to the non-transparent intervention ($ICC^{standing}_{opaque} = 37\%$).

Explaining Trust. We conducted multilevel regression analyses clustered at the individual level to explain people's trust under different levels of transparency. We aimed at identifying the importance of the interaction partners' standing for people's trust decisions, relative to other factors that may explain trust, to which count repeated experiences with others (Camera and Casari 2009; Bohnet and Huck 2004), and socio-economic variables (Johnson and Mislin 2011). For *experiences of trustworthiness*, two independent variables (IV) were included, indicating whether a very trustworthy, or a very untrustworthy experience (Table 3; Appendix (Loefflad, Chen, and Grossklags 2024)) had been made in previous rounds. Two further IV approximated the continuum of previous experiences of trustworthiness: the previously experienced return ratio, and the average of all previously experienced return ratios. We included IV relating to *experiences of trust*: the average trust received in all previous rounds as second mover, the trust received in the previous round as second mover, and the amount sent by the first mover. Gender, age, ethnicity, and the round number were included as controls.

Table 1 shows that there is some explanatory effect from *experiences of trustworthiness* in the scoring treatments. In a non-transparent system, experiencing very untrustworthy

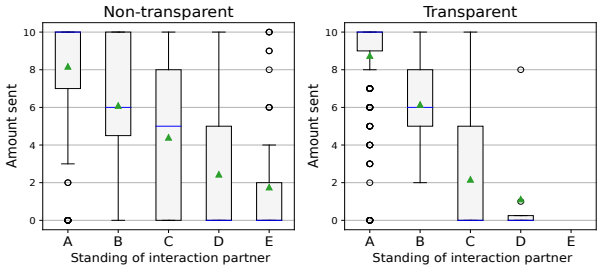

Figure 1: Trust (amount sent) based on the standing of the interaction partner.

behaviors led to a deterioration of trust, both directly after the experience had been made (β=-0.98, $p<.05$), and also in later rounds following the negative experience (β=-1.31, $p<.01$). In a transparent system, in contrast, a very trustworthy experience led to an increase in trust (β=1.01, $p<.01$). However, in both scoring treatments, the standing of the interaction partner was the most significant predictor of people's trust, even if experiences and socio-economic characteristics were controlled for. In addition, in the transparent scoring mechanism, the reduction in trust towards those with a standing below A was much stronger than in cases in which the mechanism was not transparent. Thus, people discriminated more strongly in the transparent system (see Figure 1). Moreover, once people had a signal about others' past trustworthiness, there was no systematic effect from people's socio-demographic characteristics on their trust toward others.

In the baseline intervention, in contrast, socio-economic indicators, such as being aged between 30-40 or being of Asian-Pacific or Hispanic origin were important determinants of people's trust in their interaction partners. In addition, repeated experiences of untrustworthiness from others most significantly contributed to decreasing trust in the baseline scenario (β=-1.64, $p<.001$). Lastly, in all treatments, people's kindness, approximated by the dictator decision in the pre-survey, contributed to explaining trust.

Development of Communities

Considering the experimental communities as exemplary societies, we assess the impact of social scoring systems from a society-wide perspective. To examine the impact of the interventions on the experimental communities, we distinguished the communities based on the inequalities of earnings, and the wealth the communities generated over the course of the experiment. The measure of wealth was given over the cumulative earnings of all people in a community over all rounds. This measure also indicates the developed trust between community members. The measure of inequality was given as the standard deviation of the earnings of community members. The inequality characterizes the fairness of the distribution of wealth among community members. We clustered the experimental groups according to their wealth and inequality, using k-means clustering, and the elbow method for identifying the adequate number of clusters (Cui et al. 2020). We identified five (baseline), three

	Non-transparent	Transparent	Baseline
Intercept	9.31^{***} (0.54)	7.90^{***} (0.68)	6.73^{***} (0.83)
Standing of interaction partner B	-2.02^{***} (0.30)	-2.80^{***} (0.62)	
Standing of interaction partner C	-3.51^{***} (0.34)	-6.71^{***} (0.48)	
Standing of interaction partner D	-4.68^{***} (0.55)	-6.44^{***} (0.78)	
Standing of interaction partner E	-6.57^{***} (0.36)		
Very untrustworthy experience, any round	-1.31^{**} (0.42)	0.030 (1.09)	-1.27^{**} (0.47)
Very trustworthy experience, any round	0.55 (0.37)	0.23 (0.47)	0.76 (0.49)
Very untrustworthy experience, previous round	-0.98^{*} (0.39)	−0.94 (0.97)	-1.64^{***} (0.37)
Very trustworthy experience, previous round	−0.04 (0.29)	1.01^{**} (0.37)	0.46 (0.41)
Previously experienced trustworthiness	−0.56 (0.58)	−0.14 (1.49)	−0.29 (0.52)
Previously experienced trust	-0.08^{*} (0.03)	−0.09 (0.05)	0.04 (0.03)
Average experienced trust	0.18^{**} (0.06)	0.16 (0.09)	−0.04 (0.05)
Average experienced trustworthiness	0.26 (1.13)	−1.35 (1.97)	0.31 (0.92)
Initial trust	0.13 (0.08)	0.04 (0.09)	−0.08 (0.15)
Dictator	0.22 (0.11)	0.49^{**} (0.16)	0.48^{*} (0.23)
Round number	0.04^{*} (0.02)	−0.03 (0.03)	−0.02 (0.02)
Male	−0.97 (0.53)	−0.02 (0.71)	1.49 (0.82)
Gender unrevealed	−2.68 (2.09)		
Age 25-30	−0.91 (0.59)	0.92 (0.60)	1.47 (0.87)
Age 30-40	0.38 (1.12)		-6.94^{**} (2.67)
Age >40	0.04 (1.59)	2.93 (2.14)	
Asian-Pacific	−0.51 (0.60)	0.31 (0.73)	-2.35^{*} (0.94)
Hispanic	−1.24 (1.08)	1.13 (1.38)	-5.57^{**} (1.96)
Multiracial	−1.31 (1.15)	−0.83 (1.28)	−3.08 (1.89)
Ethnicity unrevealed	−0.83 (1.41)	2.37^{*} (1.13)	−0.62 (2.81)
R^2 fixed effects	0.59	0.57	0.42
R^2 total effects	0.73	0.74	0.73
Observations	449	246	472
Log Likelihood	−984.44	−524.79	−1,115.29
AIC	2,022.88	1,097.58	2,272.57
BIC	2,133.77	1,181.71	2,359.87
Num.groups:participant.code	46	42	47
Var: participant.code	2.02	2.19	5.77
Var: Residual	3.95	3.22	5.16
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 1: Trust behaviors. Very untrustworthy and trustworthy experiences are defined in Table 3 (Appendix; (Loefflad, Chen, and Grossklags 2024)). In the scoring treatments, the data includes decisions from trustors, who have made a score request.

(transparent), and four (non-transparent) clusters.

The upper row in Figure 2 shows the community-wise boxplots of the average earnings of the community members. The level of earnings increases from baseline to transparent; per round, people earned on average 13.14 MUs in the baseline, 13.42 MUs in the non-transparent, and 14.95 MUs in the transparent treatment. The average range of people's earnings within a community amounted to 3.97 MUs in the baseline intervention, 1.84 MUs in the non-transparent, and 1.26 MUs in the transparent intervention; this suggests that the transparent scoring intervention generally led to a smaller inequality between group members. The bottom row indicates, for each experimental community, the cumulative generated wealth over the course of the experiment; in the scoring treatments, the lines are colored according to the cluster they belong to. Here, we distinguish the top and the bottom cluster, as well as all clusters in between. The cumulative earnings of different experimental communities substantially drift apart in both the baseline and non-transparent treatment; a much less pronounced drift is visible in the transparent case. Thus, communities that were subject to a transparent scoring system were much more similar regarding the creation of wealth than those subject to a non-transparent scoring treatment or a symbolic intervention mechanism. At the same time, the inequality between members of these communities was substantially lower. As such, transparency led to a more equitable distribution of benefits within communities, and a more homogeneous development across communities, in terms of inequality and wealth.

Perceptions

In this section, we compare people's perceptions across treatments. We further assess differences in perceptions across clusters of communities. We relate people's perceptions to their decision-making based on others' scores, as well as to their experiences of being disciplined based on their own scores.

Between-Treatment Differences Using one-sided Mann-Whitney *U* tests, we find that the institutional scoring system was considered significantly more legitimate once it acted

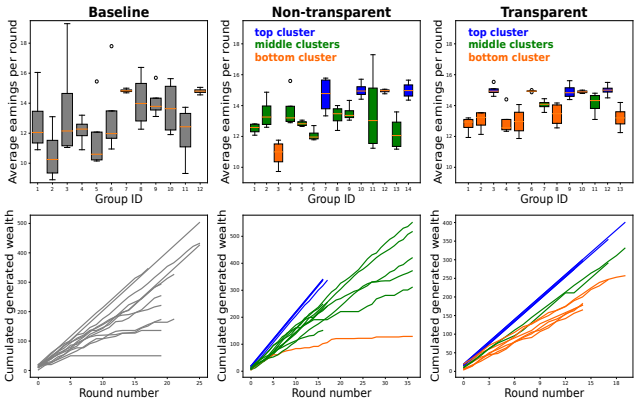

Figure 2: Upper row: Boxplots of average earnings per round within the communities. Lower row: Cumulative generated wealth over the course of the experiment. The colors indicate clusters of communities that are similar in terms of wealth and inequality of earnings.

in a transparent manner ($p<.05$). Those who knew how the mechanism worked, also considered the institutional working mechanisms more effective ($p<.05$), and more procedurally just ($p<.05$). In addition, under a transparent intervention, people reported a weakly greater perceived mutual trust, in terms of the ability to trust others ($p<.1$), as well as the feeling that they were trusted by others ($p<.1$). As compared to the baseline treatment, people considered the scoring treatments more effective ($p<.001$), and also reported greater satisfaction with the scoring treatments in general ($p<.01$). Perceptions of mutual trust were significantly lower in the baseline, as compared to the scoring treatments ($p<.001$) (see Figure 3 in the Appendix (Loefflad, Chen, and Grossklags 2024)).

Perceptions and Community Differences As outlined, the heterogeneity between communities regarding the developed trust and inequality was more pronounced in the non-transparent than in the transparent scoring system. To investigate potential sources of this variation between clusters of communities *within* the respective scoring treatments, we compare the perceptions of people belonging to the top cluster to the perceptions of people belonging to the bottom cluster. In the non-transparent mechanism, participants in the most wealthy cluster differed from those in the least wealthy cluster in their perception that the system was legitimate ($p=.067$), the viewpoint that the deployed mechanisms were procedurally just ($p=.022$), and the perception that the scoring system was effective ($p=.014$). In the transparent system, there was only a minor difference between the most and least wealthy cluster of communities. Here, the most and least wealthy clusters differed in that participants in the less wealthy cluster considered the institutional scoring system less effective ($p=.04$). Most importantly, participants from these two different transparent clusters reported similar viewpoints regarding both the procedural justice of the deployed scoring mechanism, as well as the perceived legitimacy of the scoring system.

Perceptions and Reputational Discrimination Drawing on the definition of discrimination as trusting people of different standings differently (Zarsky 2015), people discriminated in both scoring interventions, by responding with a reduction in trust towards those with a standing below A (Table 1). This suggests that people whose standing had been downgraded faced some sort of discriminatory experience, which they were very likely aware of. Therefore, we first investigate whether experiences of discrimination affected people's perceptions of a system. Reversely, these results suggest that most individuals contributed to the disciplining mechanism, by reducing trust. Therefore, we further assess whether disciplinary actions were associated with specific perceptions of the system.

We computed a measure of *experienced discrimination* for all individuals who have had a standing of A as a trustee at least once, *and* who had been a trustee with a standing below A at least once. In both situations, a score request by the trustor must have been made. The difference between the average MUs received in a standing A and the average MUs received in a standing below A constitutes the experienced discrimination. We further computed a monetary measure of people's *discrimination against others*. The measure was computed only for those individuals who, as trustors making a score request, faced at least one partner with standing A, and at least one partner with a standing below A. The difference in average trust issued towards those with a standing A and that issued towards those with a standing below A constitutes the discrimination measure (Lane 2016).

In the non-transparent system, those who were discriminated against more strongly reported more negative perceptions of legitimacy (β= -.38, p=.06) and procedural justice (β=-.37, p=.078). They also felt that others trusted them less (β=-.732, $p<.001$), and considered the institution less effective (β=-.58, p=.003). In contrast, we did not observe a relationship between perceptions of the system and experiences of discrimination in the transparent treatment. This suggests that perceptions of the scoring system, notably its legitimacy and procedural justice were not affected by experiences of discrimination.

In the non-transparent treatment, people who perceived the scoring mechanism as procedurally just showed higher levels of discrimination against others (β=.45, $p<.01$). Yet, there were no correlations between the contribution to the disciplining mechanism and perceptions of the institution when the scoring mechanism was transparent. Combined with the observation that trust behaviors towards people of the same standing were very volatile, this suggests that discriminatory actions in the non-transparent system were contingent on high perceptions of procedural justice.

Discussion and Conclusion

Technological advancements have made the possibility of algorithmically regulating society through public-sector ADM systems increasingly plausible (Yeung 2018; O'Reilly 2013). Social scoring systems are an instance of ADM systems, using behavioral scores to deliberately alter people's behaviors (Cristianini and Scantamburlo 2020). In this context, the EU AI Act forbids social scoring practices that vio-

late contextual integrity, and imposes increased transparency requirements (European Commission 2021). This suggests that a substantial source of harm stemming from social scoring systems may be eliminated by providing contextual integrity. In this work, we, therefore, studied the behavioral, perceptual and society-level implications of a social scoring system used for behavioral regulation. We focused on a system in which contextual integrity was *maintained*, and investigated the implications of *different levels of transparency*. We focused on people's trustworthiness, as well as their trust toward each other, and assessed their perceived legitimacy, effectiveness, procedural justice, and outcome favorability.

The intended goal of the scoring system to increase trust and trustworthiness was more firmly realized once the scoring mechanism was transparent, leading to a considerably higher trust of people in each other. Additionally, once people had a clear explanation of the behavior underlying a specific social score, their trust towards people with the same score was very uniform. In this case, reductions of trust towards those with a lower social score were substantially more pronounced, compared to cases in which the mechanism was opaque. The provision of transparency thus led to strong disciplining actions, probably because information about the scoring mechanism helped people judge others' behaviors more confidently and, therefore, made them discipline others more firmly. Reversely, once the scoring mechanism was opaque, trust behaviors towards people of the same standing were very volatile. This may be due to the fact that people do not feel comfortable with acting upon decisions made by ADM systems, if they do not exactly know how these decisions were made (Rader, Cotter, and Cho 2018).

Communities subject to a transparent system created more wealth compared to communities subject to a non-transparent system. In addition, in a transparent scoring system, the between-community differences in wealth and inequality were less firmly pronounced than in a non-transparent system. Considering the experimental communities as exemplary societies, our work thus illustrated how differences in the design of socio-technical systems impact societies in the aggregate.

Providing transparency made people perceive the scoring system as significantly more effective, procedurally just, and legitimate. Our results thus confirm that transparency enhances positive perceptions of ADM systems (Tiarks 2021; de Fine Licht and de Fine Licht 2020). At the same time, our results lead to questions about the current design of less transparent scoring systems, which are often described as black boxes (Citron and Pasquale 2014). In a non-transparent system, people's tendency to discipline others was correlated with their viewpoints; those who considered the system more procedurally just also disciplined others to a higher extent. In addition, in a non-transparent scoring system, experiences of being disciplined disparately impacted perceptions of procedural justice and legitimacy; those who were disciplined by others to a higher extent also held more negative viewpoints of the system. No such associations were found in the transparent case.

Clusters of communities subject to a transparent scoring system, which differed in generated wealth and inequality, did not hold different viewpoints regarding the legitimacy, effectiveness, or procedural justice of the system. These clusters differed from each other in that people in the clusters that achieved the highest wealth had a stronger general faith in others. In line with previous studies (Wang, Harper, and Zhu 2020; Kizilcec 2016; Lee and Rich 2021), this finding may indicate that perceptions of social scoring systems are also shaped by cultural and social characteristics, such as their general trust, education, or ethnic background. Moreover, in a transparent system, communities reported similar perceptions of legitimacy of the scoring system, even if they differed in the developed wealth and inequality. Once the mechanism was opaque, in contrast, those clusters that perceived the scoring system as legitimate, effective, and procedurally just also achieved higher wealth and a lower inequality level between community members. Clusters that did not share these viewpoints, in contrast, were characterized by higher inequality levels, and decreased wealth.

While our work showed that transparency in social scoring systems, as required by the EU AI Act (European Commission 2021), is crucial for eliminating disparate impacts in the aggregate, it also suggests that transparency may well impose harm on the individual level. Transparency made people strongly discipline non-compliant individuals, by reducing trust towards those with a low standing, and, therefore, their experimental earnings. If social scoring systems are applied to real-world decision contexts, considerable discriminatory consequences could thus arise. To better understand the harms imposed by social scoring systems on an individual level, future work needs to elaborate more closely on the effects of transparency considering differences in socio-demographic or cultural aspects, particularly because these factors can moderate the impact of system-level properties on perceptions, e.g., of fairness, of ADM systems (Lee and Baykal 2017; Lee and Rich 2021; Wang, Harper, and Zhu 2020). Further, the role of people's general involvement in AI-related technologies, for example, their prior experiences with, general trust in, or knowledge of ADM systems should be considered (Ehsan et al. 2024; Silva, Chen, and Zhu 2022; Choung, David, and Ross 2023).

In the specific case of social scoring, it is also important to understand the reactions of people with different social scores, even if contextual integrity is maintained (Loefflad and Grossklags 2024). Social scoring systems, even if lawful, need to ensure that disciplining effects do not lead to discriminatory outcomes. Experiences of discrimination can nurture mistrust in others (Rothstein and Stolle 2008), and towards the ADM system (Schmidt, Biessmann, and Teubner 2020). Such investigations are imperative behind the unclear legal status of social scoring systems that maintain contextual integrity. Our results suggest that the harms associated with social scoring systems are substantial, and that tighter regulation is, therefore, needed. Finally, it is important to note that understanding the broader implications of social scoring gains importance as social scoring systems for behavioral regulation are increasingly deployed (Wolfangel 2022; Stadt Wien 2021; Boos 2022).

Acknowledgments

We wish to thank the anonymous reviewers for their valuable feedback. We further thank Felix Fischer, Emmanuel Syrmoudis and the facilitators of the ExperimenTUM lab for their support. We are grateful for funding support from the Bavarian Research Institute for Digital Transformation (bidt) and the Institute for Ethics in Artificial Intelligence (IEAI). Responsibility for the contents of this publication rests with the authors.

References

Aksoy, B.; Harwell, H.; Kovaliukaite, A.; and Eckel, C. 2018. Measuring Trust: A Reinvestigation. *Southern Economic Journal*, 84(4): 992–1000.

Alessandro, M.; Cardinale Lagomarsino, B.; Scartascini, C.; Streb, J.; and Torrealday, J. 2021. Transparency and Trust in Government. Evidence from a Survey Experiment. *World Development*, 138.

Aljazzaf, Z.; Perry, M.; and Capretz, M. 2010. Online Trust: Definition and Principles. In *Proceedings – 5th International Multi-Conference on Computing in the Global Information Technology, ICCGI 2010*, 163–168. Piscataway, NJ, USA: IEEE Computer Society.

Ananny, M.; and Crawford, K. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability. *New Media & Society*, 20(3): 973–989.

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.

Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact. *California Law Review*, 104(3): 671–732.

Berg, J.; Dickhaut, J.; and McCabe, K. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1): 122–142.

Birhane, A.; Ruane, E.; Laurent, T.; S. Brown, M.; Flowers, J.; Ventresque, A.; and L. Dancy, C. 2022. The Forgotten Margins of AI Ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 948–958. New York, NY, USA: ACM.

Bohnet, I.; and Huck, S. 2004. Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change. *American Economic Review*, 94(2): 362–366.

Bolton, G. E.; and Ockenfels, A. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90(1): 166–193.

Boos, D. 2022. Bologna Introduces Social Credit App to Promote "Virtuous Behavior". https://europeanconservative.com/articles/news/bologna-introduces-social-credit-app-to-promote-virtuous-behavior/. Last accessed on April 16, 2024.

Brehm, J.; and Rahn, W. 1997. Individual-level Evidence for the Causes and Consequences of Social Capital. *American Journal of Political Science*, 41(3): 999–1023.

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.

Camera, G.; and Casari, M. 2009. Cooperation among Strangers under the Shadow of the Future. *American Economic Review*, 99(3): 979–1005.

Castelo, N.; Bos, M. W.; and Lehmann, D. R. 2019. Task-dependent Algorithm Aversion. *Journal of Marketing Research*, 56(5): 809–825.

Chen, D. L.; Schonger, M.; and Wickens, C. 2016. oTree – An Open-Source Platform for Laboratory, Online, and Field Experiments. *Journal of Behavioral and Experimental Finance*, 9: 88–97.

Chen, M.; Bogner, K.; Becheva, J.; and Grossklags, J. 2023. On the Transparency of the Credit Reporting System in China. *Humanities and Social Sciences Communications*, 10(1): 1–10.

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.

Choung, H.; David, P.; and Ross, A. 2023. Trust in AI and its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*, 39(9): 1727–1739.

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

Coleman, J. S. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Collier, P. 2002. Social Capital and Poverty: A Microeconomic Perspective. In Grootaert, C.; and van Bastelaer, T., eds., *The Role of Social Capital in Development: An Empirical Assessment*, 19–41. Cambridge, UK: Cambridge University Press.

Court of Justice of the European Union. 2023. The General Data Protection Regulation (GDPR) Opposes Two Data Processing Practices by Credit Information Agencies. https://curia.europa.eu/jcms/upload/docs/application/pdf/2023-12/cp230186en.pdf. Last accessed on July 26, 2024.

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

Cui, M.; et al. 2020. Introduction to the k-means Clustering Algorithm based on the Elbow Method. *Accounting, Auditing and Finance*, 1(1): 5–8.

de Fine Licht, J.; Naurin, D.; Esaiasson, P.; and Gilljam, M. 2014. When Does Transparency Generate Legitimacy? Experimenting on a Context-bound Relationship. *Governance*, 27(1): 111–134.

de Fine Licht, K.; and de Fine Licht, J. 2020. Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations Are Key When Trying to Produce Perceived Legitimacy. *AI & SOCIETY*, 35(4): 917–926.

Duradoni, M.; Paolucci, M.; Bagnoli, F.; and Guazzini, A. 2018. Fairness and Trust in Virtual Environments: The Effects of Reputation. *Future Internet*, 10(6): 1–15.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–19. New York, NY, USA: ACM.

Ehsan, U.; Passi, S.; Liao, Q. V.; Chan, L.; Lee, I.-H.; Muller, M.; and Riedl, M. O. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–32. New York, NY, USA: ACM.

Engelmann, D.; and Fischbacher, U. 2009. Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game. *Games and Economic Behavior*, 67(2): 399–407.

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. New York, NY, USA: ACM.

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. New York, NY, USA: ACM.

Eslami, M.; Vaccaro, K.; Lee, M. K.; Elazari Bar On, A.; Gilbert, E.; and Karahalios, K. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. New York, NY, USA: 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.

Fehr, E.; and Schmidt, K. M. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3): 817–868.

Friedman, B.; Khan Jr, P. H.; and Howe, D. C. 2000. Trust Online. *Communications of the ACM*, 43(12): 34–40.

Fukuyama, F. 1995. *Trust: The Social Virtues and the Creation of Prosperity*. New York, NY, USA: Free Press.

Glaeser, E. L.; Laibson, D.; Scheinkman, J.; and Soutter, C. L. 2000. Measuring Trust. *The Quarterly Journal of Economics*, 115(3): 811–846.

Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines*, 30: 99–120.

Hardin, R. 1993. The Street-Level Epistemology of Trust. *Politics & Society*, 21(4): 505–529.

Jackson, J.; Bradford, B.; Giacomantonio, C.; and Mugford, R. 2023. Developing Core National Indicators of Public Attitudes towards the Police in Canada. *Policing and Society*, 33(3): 276–295.

Johnson, N.; and Mislin, A. 2011. Trust Games: A Meta-Analysis. *Journal of Economic Psychology*, 32(5): 865–889.

Juijn, G.; Stoimenova, N.; Reis, J.; and Nguyen, D. 2023. Perceived Algorithmic Fairness Using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 775–785. New York, NY, USA: ACM.

Kandori, M. 1992. Social Norms and Community Enforcement. *The Review of Economic Studies*, 59(1): 63–80.

Kitchin, R. 2017. Thinking Critically about and Researching Algorithms. *Information, Communication & Society*, 20(1): 14–29.

Kizilcec, R. F. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 2390–2395. New York, NY, USA: ACM.

Knack, S.; and Keefer, P. 1997. Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics*, 112(4): 1251–1288.

Lane, T. 2016. Discrimination in the Laboratory: A Meta-Analysis of Economics Experiments. *European Economic Review*, 90: 375–402.

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 Baykal, S. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 1035–1048. New York, NY, USA: ACM.

Lee, M. K.; Jain, A.; Cha, H. J.; Ojha, S.; and Kusbit, D. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–26.

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. New York, NY, USA: ACM.

Letki, N. 2006. Investigating the Roots of Civic Morality: Trust, Social Capital, and Institutional Performance. *Political Behavior*, 28(4): 305–325.

Levi, M.; Sacks, A.; and Tyler, T. 2009. Conceptualizing Legitimacy, Measuring Legitimating Beliefs. *American Behavioral Scientist*, 53: 354–375.

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. Atlanta, GA, USA: AIS.

Loefflad, C.; Chen, M.; and Grossklags, J. 2024. Social Scoring Sytems for Behavioral Regulation. An Experiment on the Role of Transparency in Determining Perceptions and Behaviors – Appendix. https://osf.io/phjw5. Last accessed on July 26, 2024.

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. New York, NY, USA: ACM.

Martin, K.; and Waldman, A. 2022. Are Algorithmic Decisions Legitimate? The Effect of Process and Outcomes on Perceptions of Legitimacy of AI Decisions. *Journal of Business Ethics*, 183(3): 653–670.

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

Mazerolle, L.; Antrobus, E.; Bennett, S.; and Tyler, T. R. 2013. Shaping Citizen Perceptions of Police Legitimacy: A Randomized Field Trial of Procedural Justice. *Criminology*, 51(1): 33–63.

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

Nowak, M.; and Sigmund, K. 1998. Evolution of Indirect Reciprocity by Image Scoring. *Nature*, 393: 573–577.

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.

Packin, N. G.; and Lev-Aretz, Y. 2016. On Social Credit and the Right to be Unnetworked. *Columbia Business Law Review*, 2016(2): 339–425.

Park, J. S.; Barber, R.; Kirlik, A.; and Karahalios, K. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–15.

Putnam, R. D.; Leonardi, R.; and Nanetti, R. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, NJ, USA: Princeton University Press.

Rader, E.; Cotter, K.; and Cho, J. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–13. New York, NY, USA: ACM.

Rothstein, B. 2000. Trust, Social Dilemmas and Collective Memories. *Journal of Theoretical Politics*, 12(4): 477–501.

Rothstein, B.; and Stolle, D. 2008. The State and Social Capital: An Institutional Theory of Generalized Trust. *Comparative Politics*, 40(4): 441–459.

Schmidt, P.; Biessmann, F.; and Teubner, T. 2020. Transparency and Trust in Artificial Intelligence Systems. *Journal of Decision Systems*, 29(4): 260–278.

Silva, D. E.; Chen, C.; and Zhu, Y. 2022. Facets of Algorithmic Literacy: Information, Experience, and Individual Factors Predict Attitudes toward Algorithmic Systems. *New Media & Society*, 26(5): 2992–3017.

Sobel, J. 2002. Can We Trust Social Capital? *Journal of Economic Literature*, 40(1): 139–154.

Stadt Wien. 2021. Kultur Token – Klima schonen und Kultur genießen. https://digitales.wien.gv.at/projekt/kultur-token/. Last accessed on April 16, 2024.

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

Steinhardt, H. C. 2012. How Is High Trust in China Possible? Comparing the Origins of Generalized Trust in Three Chinese Societies. *Political Studies*, 60(2): 434–454.

Tiarks, E. 2021. The Impact of Algorithms on Legitimacy in Sentencing. *Journal of Law, Technology and Trust*, 2(1): 1–23.

Tyler, T. R. 2006a. Psychological Perspectives on Legitimacy and Legitimation. *Annual Review of Psychology*, 57: 375–400.

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

Tyler, T. R.; and Fagan, J. 2008. Legitimacy and Cooperation: Why Do People Help the Police Fight Crime in Their Communities? *Ohio State Journal of Criminal Law*, 6: 231–276.

Tyler, T. R.; and Jackson, J. 2013. Popular Legitimacy and the Exercise of Legal Authority: Motivating Compliance, Cooperation and Engagement. *Psychology Public Policy and Law*, 20(1): 78–95.

Uslaner, E. 2002. *The Moral Foundations of Trust*. Cambridge, UK: Cambridge University Press.

Uslaner, E. M. 1999. Trust but Verify: Social Capital and Moral Behavior. *Social Science Information*, 38(1): 29–55.

Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act – Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach. *Computer Law Review International*, 22(4): 97–112.

Vogl, T.; Seidelin, C.; Ganesh, B.; and Bright, J. 2020. Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities. *Public Administration Review*, 80(6): 946–961.

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. New York, NY, USA: ACM.

Watkins, E. A.; Moss, E.; Metcalf, J.; Singh, R.; and Elish, M. C. 2021. Governing Algorithmic Systems with Impact Assessments: Six Observations. In *Proceedings of the 2021*

AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, 1010–1022. New York, NY, USA: ACM.

Whittlestone, J.; Nyrup, R.; Alexandrova, A.; and Cave, S. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 195–200. New York, NY, USA: ACM.

Wolfangel, E. 2022. Nein, Bayern bereitet keine Überwachung chinesischer Art vor. https://www.zeit.de/digital/datenschutz/2022-07/oeko-token-bayern-belohnungssystem-social-scoring. Last accessed on April 16, 2024.

Xiao, E.; and Bicchieri, C. 2010. When Equality Trumps Reciprocity. *Journal of Economic Psychology*, 31(3): 456–470.

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

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. New York, NY, USA: ACM.

Yu, L.; and Li, Y. 2022. Artificial Intelligence Decision-Making Transparency and Employees' Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort. *Behavioral Sciences*, 12(5): 1–17.

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. New York, NY, USA: ACM.

Zacharia, G.; and Maes, P. 2000. Trust Management through Reputation Mechanisms. *Applied Artificial Intelligence*, 14(9): 881–907.

Zarsky, T. 2015. Understanding Discrimination in the Scored Society. *Washington Law Review*, 89(4): 1375–1412.

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.