# How the Types of Consequences in Social Scoring Systems Shape People's Perceptions and Behavioral Reactions

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

Jens Grossklags
Technical University of Munich
Munich, Germany
jens.grossklags@in.tum.de

## ABSTRACT

In the context of the rise of algorithmic decision-making (ADM) systems, *social scoring systems* are particularly controversial. They aim to encourage socially desirable behaviors by rewarding people with a good score in various decision-making contexts. In this paper, we report the results of a survey following a social scoring experiment, to predominantly understand the impact of the scoring *outcome* and the *decision importance* on people's perceptions and behavioral intentions within an abstract social scoring system. We find that the outcome was pivotal for creating opinion differences regarding people's perceptions, and behavioral reactions. In contrast, the decision importance did not exert a systematic impact on people's perceptions and behavioral reactions, but exacerbated existing opinion differences in terms of perceived effectiveness. Specifically, the outcome strongly shaped the structural relationship between people's experiences, perceptions, and behavioral reactions, creating a substantial outcome favorability bias for people with a bad outcome. Although people with a bad outcome reported an intention to adapt their behaviors, their intention to engage in desired behaviors could not be attributed to a perceived legitimacy of the system. For those with a good outcome, perceptions of procedural justice and legitimacy were weakened by the privacy-invading character of the social scoring system. Our work shows that the outcome people receive might create a pivotal disparate impact on people's overall attitudes towards social scoring, shape their behavioral reactions, and create divergent behavioral motives, suggesting that very distinct societal dynamics may arise.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy** → **Economics of security and privacy**; **Social aspects of security and privacy**.

## KEYWORDS

social scoring systems, experiment, procedural justice, legitimacy



*FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil*

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658986

**ACM Reference Format:**
Carmen Loefflad and Jens Grossklags. 2024. How the Types of Consequences in Social Scoring Systems Shape People's Perceptions and Behavioral Reactions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3630106.3658986

## 1 INTRODUCTION

Social scoring systems classify people based on their social behaviors or personal characteristics into "good" and "bad", and distribute benefits to people with a good score, while denying access to benefits for people with a bad score. As such, these systems aim at making people engage in specific behaviors, which often serve a society-wide goal [89]. Practices going into the direction of social scoring are increasingly deployed around the globe [1, 11, 67, 88]. However, social scoring systems raise strong controversies due to their opacity, privacy-invading character, and adverse impact on different groups of people [21, 91]. The EU AI Act considers social scoring systems as systems creating an "unacceptable risk", as the sorting of individuals may lead to a "detrimental treatment" of people, and violate rights of equality [30]. In contrast, proponents consider social scoring systems a promising tool for governing society-wide behaviors in an automated manner [23, 36, 62]. A key factor adding to the controversial debate is that social scoring systems often violate contextual integrity. Contextual integrity requires that the behavioral score, derived from data collected within a specific domain, should exclusively inform decision-making within this domain [61].

To understand the disparate impact of social scoring systems on different groups of individuals it is imperative to investigate people's attitudes and behavioral responses towards social scoring systems. In this context, one of the foremost questions is how the way people are treated by a social scoring system, i.e., *the types of consequences* they experience, impacts their perceptions and behavioral reactions. Both the score people receive, *the outcome*, as well as the importance of the decision context in which social scores are used for decision-making, *the decision importance*, determine the types of consequences. In this work, we report on the results of a comprehensive survey following an experimental study. The survey investigates the impact of the types of consequences on people's perceptions and behavioral reactions in a social scoring system that violates contextual integrity.

To study people's *perceptions*, we adopt a procedural justice-driven approach [71], and account for several key perceptions that this theory displays as central. To these count judgments of procedural justice, legitimacy, and effectiveness. As for people's *behavioral reactions*, we assess both people's intention to comply with the

system, as well as their intention to adapt their behaviors. We investigate two guiding research questions. First, we ask how the types of consequences, namely the scoring outcome (good vs. bad) and the decision importance (high-stakes vs. low-stakes), affect people's perceptions and behavioral reactions, using multiple linear regressions (RQ1). We further investigate how the types of consequences affect the structural relationship between people's behavioral reactions and perceptions (RQ2). We include individuals' *experiences* into the structural model, evaluating the perceived favorability of their outcomes (outcome favorability), and the degree to which they feel their privacy is invaded (subjective privacy harms).

We find that the outcome created strong opinion differences, in terms of how people experienced the system, what they thought of the system, as well as how they reacted to the system. While there was no systematic effect of the decision importance on perceptions or behaviors, it exacerbated differences in opinions regarding the perceived effectiveness, and people's feeling of being treated with respect. People with a bad outcome generally reported an intention to change their behaviors. At the same time, and opposed to those with a good outcome, their intention to comply with the system could not be attributed to the perception that the system is legitimate. The structural relationship between people's perceptions, experiences, and intention to comply was strongly shaped by the outcome, but not by the decision importance. Perceptions of procedural justice and legitimacy were biased by people's judgments of the outcome, but this bias was much stronger for people with a bad outcome. For those with a good outcome, perceptions of justice and legitimacy were weakened due to the privacy-invading character of the system.

From a higher level perspective, our work shows that the consequences people receive might create a pivotal disparate impact on people's overall attitudes towards social scoring, shape their behavioral reactions, and possibly shift the motivation behind these reactions from being perceived as legitimate to being perceived as coercive. Our results further suggest that very divergent societal dynamics may arise with the introduction of social scoring systems.

## 2 RELATED WORK

### 2.1 Social Scoring Systems

Social scoring refers to the activity of assessing a character trait from behavioral data. Commonly, people with a good score have access to benefits. People with a bad score, in contrast, are deprived of these benefits, or even receive punishments, depending on the strictness of the social scoring system. As such, incentives are created to engage in behaviors that are desired by the system. Specifically, social scoring systems aim at achieving behavioral changes, e.g. to make people behave more trustworthy [17, 67], or more pro-environmentally [1, 11]. Depending on the level of automation, scoring systems can be understood as automated decision-making (ADM) systems, which algorithmically regulate society-wide behaviors, and which dispose of an ordering function [23, 49, 89]. Practices that resemble social scoring are increasingly deployed in Europe [1, 11, 88], and some scholars argue that they will be an integral part of social regulation efforts in the future [36, 62]. Yet, the EU is placing limitations on scoring practices that classify people based on social behaviors or on predicted character traits. The EU AI Act considers AI-based scoring systems "unacceptable" risk applications, and aims at prohibiting their application in contexts that violate contextual integrity, since such a violation could lead to a "detrimental treatment that is disproportionate or unjustified to the gravity of their social behavior" [30]. Contextual integrity requires that information generated in a certain domain should be used to make decisions only in this domain [61]. However, whether contextual integrity is maintained primarily depends on how narrowly the context is defined [82], and the EU AI Act still needs to provide a clear definition of contextual integrity. To date, most scoring efforts are characterized by a violation of contextual integrity [11, 67, 88]. For example, some systems assess people's pro-environmental behaviors, and distribute low-stakes benefits in cultural institutions [1, 11]. In the Chinese Social Credit System (SCS), in contrast, the violation of contextual integrity is deliberately used for making high-stakes decisions, which have an important economic impact on people's lives [24, 29].

In our prior work, we have investigated the impact of different levels of transparency on people's perceptions of and behavioral reactions to a social scoring system, which *maintains* contextual integrity. We found that transparency in social scoring systems is key for preventing undue harms [50]. Building on this finding, the present work is centered on a transparent social scoring system, which *violates* contextual integrity. The following subsection explains the perceptions and behavioral reactions that are the focus of this study.

### 2.2 Using Procedural Justice Theory to Study Perceptions and Behavioral Reactions to Social Scoring

Procedural justice theory is concerned with analyzing how people's perceptions of a decision-making system vary with the level of procedural justice inherent in the decision-making processes [12]. In this context, the theory allows for evaluating the legitimacy of a decision-making system, as well as people's intention to comply with it. It also helps understand how perceptions of legitimacy and compliance intentions are shaped by perceptions of procedural justice [54, 70, 80]. Therefore, we assess both people's perceived procedural justice, and perceived legitimacy when they are subject to a social scoring system. We further measure perceptions of effectiveness to account for the possibility that compliance is also driven by people's belief that the scoring system is successful in imposing consequences [63].

Procedural justice refers to the "perceived fairness of the processes by which outcomes are reached" [46]. To its components count people's control in and understanding of the scoring processes [71], transparency [46], benevolent motives of a decision-making organ [46], as well as a respectful treatment of decision subjects [80]. Perceived legitimacy refers to the appropriateness of a system that operates in a society [68]; to be viewed as legitimate, the processes of a system should be normatively aligned with people's moral codes [38, 80], and elicit an obligation to obey an authority [76, 78, 80]. In addition, legitimacy implies that people develop trust towards a system [80]. While both trust in ADM systems [18, 20, 47], as well as perceived trustworthiness of ADM systems

[64] has found considerable attention in the human-computer interaction literature, only few studies have centered on the perceived legitimacy of ADM systems [52, 84]. Lastly, the effectiveness of ADM systems has been conceptualized as usefulness of a system [8]. In the context of social scoring systems, perceived effectiveness may also be understood as the perceived ability to effectively incentivize people to engage in desirable behaviors [37].

Assessing a variety of perceptions of decision-making systems is key for understanding people's behavioral reactions to a system [70, 76, 80]. In the context of our study, we assess people's intention to comply, which refers to their willingness to engage in the desired behaviors. We further investigate people's behavior changes, in terms of how strongly they adapt their behaviors after having learned how the scoring system functions. Our multidimensional and procedural justice-driven approach allows us to identify possible pathways that shape behavioral reactions. This approach also helps develop a nuanced understanding of the impact of the types of consequences on perceptions and behavioral reactions. The next section elaborates on the consequences emerging from social scoring systems, and presents variables relating to *individual differences* and *experiences*.

## 2.3 Factors Shaping Perceptions and Behavioral Reactions

*2.3.1 Types of Consequences.* The impact of a social scoring system on individuals depends on both the score they receive (*outcome*), as well as on the importance of the decision context in which the score is consulted for decision-making (*decision importance*). Some scoring systems are only used for making decisions that are of low importance (*low-stakes decisions*), distributing cultural benefits for people with a good outcome [1]. The Chinese SCS, in contrast, uses scores for making decisions that are of high importance (*high-stakes decisions*), which specifically exclude those with a bad outcome from access to important services and goods [22, 28]. Using social scoring systems for making high-stakes decisions further fuels the ethical debate about those systems, and might increase the "detrimental" disparate impact on people with a bad outcome [30]. In social scoring systems, the outcome people receive [52, 85] as well as the decision importance [8, 16, 43, 58] are likely to have a pivotal impact on people's perceptions. Understanding how the consequences individuals receive impact their perceptions is important, because judgments of fairness, legitimacy, or justice in the decision-making procedures shape individuals' behavioral reactions [54, 80]. In addition, groups of people that react differently to social scoring systems may contribute to shaping inequality between individuals [34]. Further, understanding the mechanisms that drive behavioral reactions can provide valuable insights into the societal dynamics that arise with the introduction of a social scoring system [41]. The central question of this work is to understand how specific kinds of consequences affect people's perceptions and their behavioral reactions. In this context, *individual differences*, in terms of people's computer literacy [8, 18, 64, 85], as well as their general privacy attitudes [8, 18] may contribute to shaping people's perceptions and behavioral reactions.

*2.3.2 Individual Differences.* People's literacy towards digital technologies is referred to as AI literacy [51], or computer literacy [85]. People's illiteracy counts towards the most significant dangers in the context of ADM systems [14]; people's limited understanding of how an ADM system works may exacerbate the disparate impact stemming from the system [35]. The EU AI Act specifically emphasizes the need to educate citizens in AI-related fields [30]. Computer literacy greatly impacts judgements of algorithms or ADM systems [8, 18, 64]. Controlling for people's computer literacy may thus help identify sources of bias in the assessment of people's perceptions of social scoring systems.

As ADM systems largely rely on large-scale data collection [59], people's general privacy-related attitudes may shape how people perceive these systems, which we refer to as *general privacy concerns* [66]. General privacy concerns can be negatively associated with judgments of fairness and usefulness of ADM systems [8, 72], shape the legitimate status of ADM systems [18], or impact people's intention to use these systems [40].

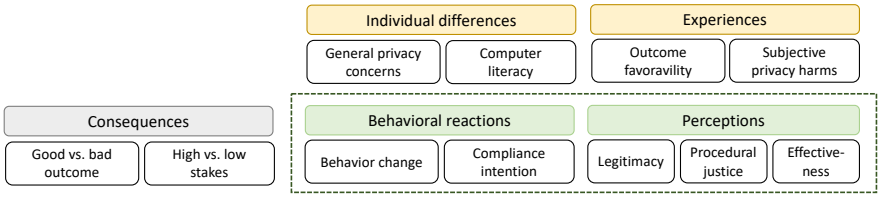
With our second research question, we assess the impact of the types of consequences on the structural relationship between people's perceptions and behaviors, including people's *experiences*, which are presented in the following.

*2.3.3 Experiences.* The large-scale collection of behavioral data, as inherent in social scoring systems [21], raises strong privacy-related concerns [21, 91]. These are also referred to as *subjective privacy harms*. Subjective privacy harms refer to the unwanted state of being observed. They arise when people feel uncomfortable once decision-making systems collect behavioral data, or when people feel that the collection of behavioral data may be disadvantageous to them [15]. Accounting for perceived privacy violations might lead us to better understand people's attitudes towards ADM systems [65], specifically as systems that provide more privacy lead to increased satisfaction among users [2, 53], and may also impact people's behavioral reactions to a system [50].

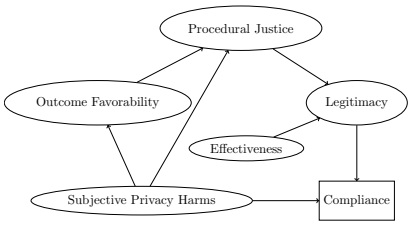
Outcome favorability refers to people's subjective evaluation of the outcome they receive from an ADM system [52, 86]. An important facet of procedural justice theory is to separate people's evaluation of a specific outcome from the evaluation of the procedures leading to this outcome [13]. This separation is important because people who perceive their outcome as favorable possibly consider a decision-making organ as more procedurally just [6], which is also referred to as *outcome favorability bias*. This bias is undesirable, as it undermines an objective assessment of the fairness of a system [85]. For our context, this implies the subjective valence of the consequences that people receive, in terms of their outcome and the decision importance, may significantly shape people's perceptions and behavioral responses.

# 3 RESEARCH QUESTIONS AND HYPOTHESES

In this section, we first establish a set of research hypotheses and questions to answer our first guiding research question of how the types of consequences impact people's perceptions and behavioral reactions. Second, we establish a structural equation model. Structural equation modeling allows for testing hypothesized associations between several latent constructs [42]. In our case, we investigate associations between people's perceptions, experiences, and behavioral reactions. An overview of the research agenda is given in Figure 1a.




**(a) Research overview. First, we investigate how the types of consequences affect behavioral reactions and perceptions, controlling for individual differences. Second, we ask how the types of consequences affect the structural relationship between people's perceptions and behavioral reactions, including people's experiences.**




**(b) Structural model between perceptions, experiences, and compliance, adapted from Loefflad et al. [50]).**

**Figure 1: Research Overview**

## 3.1 Impact of the Types of Consequences on Perceptions and Behavioral Reactions

*3.1.1 Scoring Outcome.* Wang et al. [85] find that people with a good outcome perceive an ADM system significantly more fair than people with a bad outcome. Martin and Waldman [52] find that perceptions of legitimacy are higher for people with a good outcome as opposed to those with a bad outcome . Therefore, we hypothesize:

- **H1**: There is a positive association between receiving a good outcome and perceptions of procedural justice.
- **H2:** There is a positive association between receiving a good outcome and perceptions of legitimacy.

In addition, we address the following research questions:

- **RQ1:** How does the outcome impact perceptions of effectiveness?
- **RQ2:** How does the outcome impact people's behavioral reactions, in terms of their intention to comply, and their behavior change?

*3.1.2 Decision Importance.* Systems making decisions that are of high complexity [16] or high importance [8, 43, 58] to one's life are likely to be judged differently than ADM systems making decisions that are less complex or less important. While some find that applying ADM system to a high-stakes decision can *negatively impact* people's perceived procedural justice of an ADM system [16, 43, 58], judgments of procedural justice may also increase, for example when human decision-making in high-complexity tasks is complemented with algorithmic decision-making [58]. The impact on judgments of fairness or justice when applying ADM systems to high-stakes decisions might further depend on the specific application scenario [8, 10]. Yurrita et al. [90] find that the stakes involved do not impact perceptions of procedural justice. Due to the lack of comprehensive results, we raise the following research question:

- **RQ3:** How does the decision importance impact people's perceptions of procedural justice?

Literature in social psychology suggests that the extent to which the decision-making organ, which uses a specific sanction severity, is perceived as legitimate often depends on whether people consider the deployed procedures of the system just. In addition, people's willingness to comply with a decision-making system used for incentivizing specific behaviors may also depend on their perceptions of procedural justice [57, 83]. An evaluation of the impact of the decision importance on people's perceptions of legitimacy and their intention to comply with a social scoring system thus requires understanding how perceptions of procedural justice are impacted by the consequences people experience, including both the outcome and the decision importance. As outlined, perceptions of procedural justice are likely to be lower for people with a bad outcome [85]. Yet, it might also be plausible that the decision importance and the outcome interact in determining people's perceptions of legitimacy and behavioral reactions [52]. Accounting for potential interaction effects between decision importance and outcome we ask:

- **RQ4:** How does the decision importance impact perceptions of legitimacy, and people's intention to comply?
- **RQ5:** How does the decision importance impact behavioral reactions?

As for the impact of the decision importance on the perceived effectiveness of an ADM system, Araujo et al. [8] find that moving from a low-stakes to a high-stakes decision negatively affects perceptions of *usefulness*. Due to the lack of research regarding the impact of the decision importance on perceived effectiveness, we further ask:

- **RQ6:** How does the decision importance impact perceptions of effectiveness?

*3.1.3 Individual Differences.* Araujo et al. [8] find a negative relationship between privacy concerns and perceptions of fairness, and the perceived usefulness of ADM systems. Chen and Sundar [18] find a negative association between general privacy concerns and people's trust in ADM systems. Further, Jozani et al. [40] find that privacy concerns decrease users' engagement.

Computer literate people tend to perceive an ADM system as more effective [8], more fair [64, 85], and develop higher trust towards ADM systems [64]. As such, it is likely that both computer literacy as well as general privacy concerns contribute to explaining people's perceptions and behavioral reactions in the context of social scoring systems. Therefore, we control for these variables in the regression analyses.

### 3.2 Relationship between Perceptions, Experiences and Compliance Intention

In this section, we establish a research model regarding the relationship between people's perceptions, experiences, and their intention to comply, adapted from our prior work [50]. We briefly re-state the core implications of the literature in social psychology leading to this structural model as explained in detail in [50]. First, decision-making organs are more likely to be viewed as legitimate once they exercise their power in a procedurally just manner [75–77]. In particular, perceptions of procedural justice also shape perceptions of the overall legitimacy of a decision-making system [78]. In turn, in situations in which people perceive a decision organ as legitimate, they are more likely to comply with it [25, 54, 71, 76]. In this regard, high perceptions of legitimacy can lead people to comply with authorities based on a *value-driven motivation*, and not based on instrumental reasoning or coercion [70]. Therefore, we expect direct paths from procedural justice to legitimacy, as well as from perceived legitimacy to the intention to comply. In terms of the effect of perceived effectiveness on judgments of legitimacy, a large strand of literature provides evidence that judgments of procedural justice are more central to perceptions of legitimacy than the perceived effectiveness of a decision-making system [70, 76, 79]. Therefore, we also test for an instrumental source of legitimacy, in which social scoring systems develop a legitimate status through effectively incentivizing desirable behaviors [70].

In terms of the impact of people's experiences, the subjective valence of their outcome, i.e., their outcome favorability, may directly be associated with perceptions of procedural justice [6], which is also referred to as *outcome favorability bias* [86]. Moreover, empirical evidence has suggested that privacy invasion and procedural justice are negatively related to each other [8, 27, 72]. Further, subjective privacy harms might be closely related to individuals' assessment of their outcome [44, 50], and decrease people's satisfaction with it [19]. In addition, subjective privacy harms can decrease compliance behaviors that exceed the expectations of an authority [60]. We, therefore, hypothesize that subjective privacy harms are negatively associated with outcome favorability, perceived procedural justice, as well as with people's intention to comply. These findings lead us to develop a structural model as shown in Figure 1b.

A wide body of research has shown that people's perceptions of authorities are shaped by their evaluations of justice of the procedures authorities use to exercise power [79–81]. In our prior work, for example, we have shown that a transparent social scoring system was perceived as significantly more procedurally just than a non-transparent system. The difference in perceived procedural justice strongly shaped the relationship between people's experiences, perceptions, and their intention to comply [50]. In the present work, perceptions of procedural justice may strongly depend on the types of consequences. Therefore, it is likely that the structural model as depicted in Figure 1b differs with the outcome people receive or with the decision importance. Therefore, we raise the following research question:

- **RQ7:** How do the types of consequences affect the structural model between people's experience, perceptions, and intention to comply?

## 4 METHOD

### 4.1 Social Scoring Experiment

The experiment was conducted in the labs of two German universities in June 2023. The experiment consisted of two phases. In the first phase, the *score generation phase*, a social score, mimicking people's trustworthiness, was generated in a repeated trust game. In the second phase, the social score was applied to a hypothetical decision-making scenario, which was either a high-stakes or a low-stakes situation. Both decision-making scenarios differed from the context in which the score was sourced, and thus violated contextual integrity.

*4.1.1 Score Generation.* At the beginning of the experiment, participants were randomly matched to anonymous communities consisting of four people. In each community, the participants interacted with each other for 18 rounds. The number of rounds was unknown to the participants to avoid a behavior change towards the end. In each round, the four community members were randomly matched to pairs of two and played a trust game [9]. A trust game is a sequential game between a first and a second mover. In each round, the roles of the first and second mover were assigned randomly. Both players received a monetary endowment of 10 monetary units (MUs). The first mover decided how much to trust the second mover, i.e., what fraction of the endowment to send to the second mover. The second mover received twice the amount sent by the first mover. Then, the second mover decided how much to send back. The back-sending behaviors give an indication about the second-mover's trustworthiness.

Similar to [50], a social scoring system was introduced to the communities, to make people behave trustworthy. The scoring system assigned a behavioral score based on people's trustworthiness behaviors. Everyone started with a score of 1000. If people did not behave trustworthy the score was reduced. As in [50], trustworthiness was defined as sending back at least the amount that equalized the payoffs between the first and the second mover. The more untrustworthy people behaved, the more points were subtracted. The score was updated after each round.

In contrast to our prior work [50], the score could not be requested by the first movers before deciding how much to send. To each participant, the score was only privately known. As such, there was *no monetary incentive* during the score generation phase to behave trustworthy. People only knew that after the score generation phase, there would be a *hypothetical* scenario, in which their score would matter. The scoring system considered a score above 985 a good score, and a score of 985 and below a bad score. The participants were hinted towards the definition of trustworthiness at the beginning of the experiment; we provided several interaction examples between a first and second mover, and the score that would result from the interaction (Figure 4 in the Appendix). Further, people were told that a good score would indicate that they valued fairness and equality, and that they were socially engaged. The scoring system did not upgrade scores; there was thus no recovery from being assigned a bad score. As such, the score generation phase assigned an *outcome* to each participant, which was either good or bad, and which endogenously emerged from people's behaviors.

*4.1.2 Decision Importance.* After having completed the experiment, participants were presented with a hypothetical scenario, which was adapted to their score, i.e. their outcome. In the scenario, a scoring system was newly introduced to their city, and the score was used to determine access to services and goods. In the low-stakes scenario, good scores could be used for cultural benefits. In the high-stakes scenario, good scores could be used in a university application process to increase the chances of admission. Those with a bad score did not receive benefits (see Appendix A.1). We introduced a manipulation check asking people to assess the relative importance of the scenario in a person's life.

## 4.2 Experimental Surveys

In a pre-survey, participants' computer literacy, as well as their general privacy concerns were measured, to derive factors regarding their *individual differences*. Computer literacy was measured over six items [45, 69, 73, 85], and general privacy harms were measured over four items [18, 26] (Table 3 in the Appendix). The post-survey measured people's perceptions, experiences, and intention to comply as in Loefflad et al. [50] (Table 2 in the Appendix).

*4.2.1 Perceptions and Behavioral Reactions.* Perceived procedural justice was measured over nine items [3, 46, 80], referring to the perceived benevolence of the system, their control in the process, to the general transparency, clarity, and fairness of the scoring process, as well as the extent to which people felt respected. Perceived legitimacy was measured over five items, relating to whether the scoring system was aligned with people's moral values, to people's trust in the system, as well as to their perceived obligation to obey [38]. Perceived effectiveness was measured as usefulness [16], and as the ability to influence key decision makers [37]. People's intention to comply was measured in a hypothetical scenario at the end of the survey. They should suppose that the experiment was played again, under the same conditions. People should indicate the amount they would send back as a second mover, if they had received 20 MUs from the first mover in the first round of the experiment. We refer to the indicated amount as their intention to comply. From the amount indicated as intention to comply, we calculated the intended return ratio, i.e. the amount sent back divided by the amount received. Subtracting the intended return ratio from the return ratio as the first time as second mover in the experiment yields a measure for behavioral change. An attention check was included.

*4.2.2 Experiences.* People's outcome favorability captured first, to which extent people liked how they were treated, second, to which degree they were satisfied with their score, and third, to which extent people perceived the introduction of a social scoring system as favorable. Subjective privacy harms referred first, to whether the collection of behavioral data made participants feel uncomfortable, and second, whether they felt that the collection of behavioral data was disadvantageous to them [15]. Third, we asked people to what extent they felt controlled by the scoring system [50].

## 4.3 Statistical Analyses

*4.3.1 Regression.* We conducted five multiple linear regression analyses for **H1** and **H2**, and **RQ1-RQ6**. Each regression had a specific perception or behavioral reaction as the dependent variable. The independent variables were the types of consequences, in terms of outcome and decision importance, and individual differences, in terms of computer literacy and general privacy concerns. For each dependent variable, we conducted another regression, including the interaction term between outcome and decision importance. Further, we conducted pairwise t-tests to better understand the results.

*4.3.2 Structural Equation Modeling.* Following Loefflad et al. [50], we established a structural equation model between people's experiences, perceptions, and their intention to comply (Figure 1b). We used a multi-group approach to determine whether the types of consequences had an impact on the structural model (**RQ7**). In multi-group structural equation modeling, the structural model is estimated separately for each group. This allows us to analyze whether the relationship between the variables differs depending on the types of consequences. Hereby, we first construct a measurement model. All manifest variables that showed large cross-loadings, high residual error correlations, or low loadings on the latent variables were not retained [42]. Subsequently, we estimated the path model as in [50].

# 5 RESULTS

## 5.1 Descriptive Statistics

390 participants participated in the experiment. We included only those who passed the attention check ($n$=375). 194 participants were in a low-stakes, and 181 in a high-stakes scenario. In the high-stakes scenario, 150 participants had a good, and 44 had a bad outcome. In the low-stakes scenario, 133 participants had a good, and 48 had a bad outcome. Participants' earnings amounted to 19.74 Euros, on average ($SD$=4.53). They completed the experiment in 33.77 minutes, on average ($SD$=6.48). 43.5% of the participants were male, 54.9% were female, 1.0% were diverse, and 1.0% preferred not to reveal their gender. 64.8% of the participants were aged between 17 and 25, 24.0% between 25 and 30, 6.1% between 30 and 40, and 3.2% were above 40 years old. 1.8% did not reveal their age. As for their ethnic background, 61.3% were White or Caucasian, 19.2% were Asian-Pacific, 6.1% were Hispanic, and 1% were Black or African American. 9.1% did not reveal their ethnic background, and 3.6% indicated to be multiracial.

## 5.2 Impact of the Types of Consequences on Perceptions and Behavioral Reactions

In this section, we report on the results of the regression analyses, investigating how the types of consequences, namely the outcome, the decision importance, and their interaction, impact people's perceptions and behavioral reactions.

*5.2.1 Outcome.* A good outcome was positively associated with perceptions of effectiveness ($\beta$=0.98, $p$<0.001, **RQ1**), procedural justice ($\beta$=0.35, $p$<0.01, **H1**), legitimacy ($\beta$=0.56, $p$<0.001, **H2**), as well as with people's intention to comply ($\beta$=1.37, $p$<0.01, **RQ2**) (Table 1). Further, a positive outcome led to a substantial decrease in behavior change ($\beta$=-0.22, $p$<0.001, **RQ2**). Additional two-sample *t-tests* revealed that people with a good outcome perceived the scoring system as significantly more procedurally just ($M$=3.70, $SD$=0.66),

than people with a bad outcome ($M$=3.38, $SD$=0.61, $p$<0.001). Perceived legitimacy was stronger for people with a good outcome ($M$=3.34, $SD$=0.86) than for people with a bad outcome ($M$=2.90, $SD$=0.90, $p$<0.001). Those with a good outcome reported higher perceptions of effectiveness ($M$=3.53, $SD$=0.95) compared to those with a bad outcome ($M$=2.83, $SD$=1.00, $p$<0.001). Their intention to comply was also significantly higher ($M$=14.48, $SD$=2.09) compared to people with a bad outcome ($M$=12.67, $SD$=3.98, $p$<0.001). In contrast, people with a good outcome showed a significantly lower change in behavior ($M$=0.02, $SD$=0.16) than people with a bad outcome ($M$=0.21, $SD$=0.27, $p$<0.001). We also explored the impact of the outcome on people's experiences. People with a good outcome reported a significantly higher outcome favorability ($M$=4.17, $SD$=0.69) than those with a bad outcome ($M$=2.60, $SD$=0.90, $p$<0.001). Those with a bad outcome reported significantly higher subjective privacy harms ($M$=3.60, $SD$=0.98) than those with a good outcome ($M$=3.16, $SD$=0.96, $p$<0.001). These results are depicted in Figure 2b.

*5.2.2 Decision Importance.* A two-sample *t-test* revealed that people considered the high-stakes scenario significantly more important to a person's life ($M$=4.06, $SD$=0.89) than the low-stakes scenario ($M$=3.37, $SD$=1.01, $p$<0.001, $t$=6.93). The decision importance was thus successfully manipulated. We found no main effect of the decision importance on perceived effectiveness, legitimacy, procedural justice, and the intention to comply. However, moving from a high-stakes to a low-stakes scenario led to a decrease in behavior change ($\beta$=-0.07, $p$<0.1). We further found no interaction effects between the decision importance and the outcome in determining perceptions of legitimacy, procedural justice, or the intention to comply. Thus, the decision importance had no effect on people's perceptions of procedural justice (**RQ3**), nor on their perceptions of legitimacy and intention to comply (**RQ4**). For exploratory purposes, we conducted a regression analysis for each of the sub-components of procedural justice. We found a significant interaction effect between the decision importance and the outcome on people's feeling of being treated with dignity and respect ($\beta$=-0.66, $p$<0.05); that is, when changing from a low-stakes to a high-stakes decision the feeling of being treated with dignity and respect increased *only for people with a good outcome* (Figure 2a). In addition to the significant main effect of the decision importance on people's behavior change, we also found an interaction effect between the decision importance and the outcome ($\beta$=0.082, $p$<0.1). Moving from a low-stakes to a high-stakes scenario led to behavioral changes, but only for those with a bad outcome (**RQ5**). We further found an interaction effect between the decision importance and the outcome on perceived effectiveness, i.e., the positive effect of having a good outcome on perceived effectiveness was weaker in a low-stakes situation compared to a high-stakes situation ($\beta$=-0.523, $p$<0.05) (**RQ6**).

*5.2.3 Individual Differences.* Computer literacy was positively associated with perceptions of procedural justice ($\beta$ = 0.115, $p$<0.01), and legitimacy ($\beta$ = 0.093, $p$<0.1), but not with perceived effectiveness, or behavioral reactions (Table 1). General privacy concerns were neither associated with people's perceptions nor with behavioral reactions.

## 5.3 Relationship between Perceptions, Experiences, and Compliance Intention

As perceptions were mainly influenced by the outcome, and not by the decision importance, we modified **RQ7**, and asked how the outcome affects the structural model between people's experiences, perceptions, and intention to comply. We thus specified a multi-group SEM with the outcome as the grouping variable ($n_{good}$=283, $n_{bad}$=92).

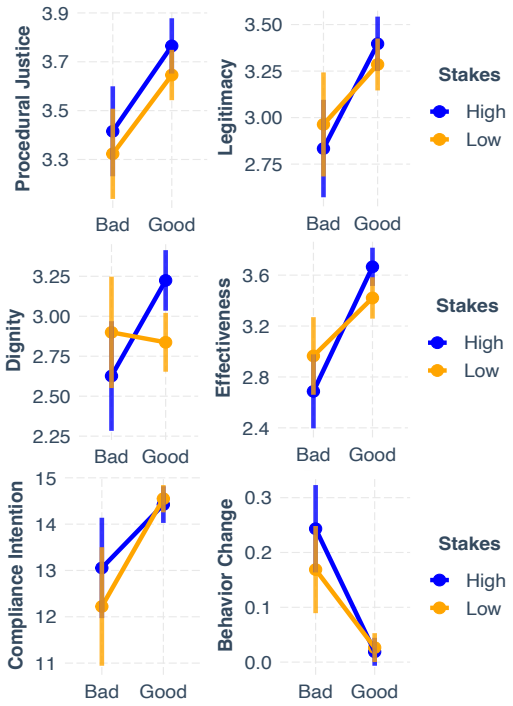
*5.3.1 Measurement Model.* In the measurement model, perceived procedural justice (PJ) was measured over items relating to the dignity and respect participants felt they were given, as well as over the benevolence of the system (b1, b2, b3, d1, see Table 2 in the Appendix). Outcome favorability (Fav) was measured over two items (f1, f2), and legitimacy (Leg) was measured over four items (ob2, na1, na2, tr1). Subjective privacy harms (SPH) were measured over two items (sph1, sph2). The intention to comply (Comp) was included as a manifest variable. Weak invariance was given, suggesting that the factor loadings of the items on the latent variables were equal across groups. This is a prerequisite for drawing meaningful comparisons in a multi-group SEM [87]. Due to the ambiguous role of perceived effectiveness in determining perceptions of legitimacy [70, 76, 79], we compared the model fit of a measurement model with the variable effectiveness to the model fit without it [48]. As the model without effectiveness had a better fit with the data, we did not include this variable.

*5.3.2 Path Model.* The direct path SPH →Fav was invariant across outcomes. The paths PJ→Leg, Fav→PJ, SPH→Leg, and SPH→PJ differed across outcomes. Removing the path SPH→Comp did not deteriorate the model fit. Procedural justice was strongly and directly associated with perceptions of legitimacy, for both outcomes ($\beta$=0.81, $p$<0.001 for bad; $\beta$=0.92, $p$<0.001 for good). Further, procedural justice was positively and indirectly associated with the intention to comply, but only for those with a good outcome ($\beta$=0.15, $p$<0.05). Similarly, perceived legitimacy was positively and directly associated with the intention to comply, but only for those with a good outcome ($\beta$=0.17, $p$<0.05). For both outcomes, outcome favorability was directly and positively associated with perceived procedural justice and indirectly and positively associated with perceived legitimacy. Yet, the effects were stronger for people with a bad outcome (direct effect $\beta$=0.59, $p$<0.001; indirect effect $\beta$=0.54, $p$<0.001), compared to people with a good outcome (direct effect $\beta$=0.28, $p$<0.01; indirect effect $\beta$=0.26, $p$<0.01). For both outcomes, subjective privacy harms had a significant negative effect on outcome favorability ($\beta$=-0.47, $p$<0.001 bad outcome; $\beta$=-0.53, $p$<0.001 good outcome). Further, subjective privacy harms were directly and negatively associated with perceived procedural justice only for those with a good outcome ($\beta$=-0.64, $p$<0.001). Yet, for both outcomes, there was a significant indirect effect on procedural justice ($\beta$=-0.28, $p$<0.01 bad outcome; $\beta$=-0.13, $p$<0.01 good outcome). The total negative effect of subjective privacy harms on perceived procedural justice was significant for both outcomes, but stronger for those with a good ($\beta$=-0.77, $p$<0.001), compared to those with a bad outcome ($\beta$=-0.53, $p$<0.001). Further, subjective privacy harms directly and negatively impacted perceived legitimacy only for those with a good outcome ($\beta$=-0.17, $p$<0.01). However, for both outcomes,

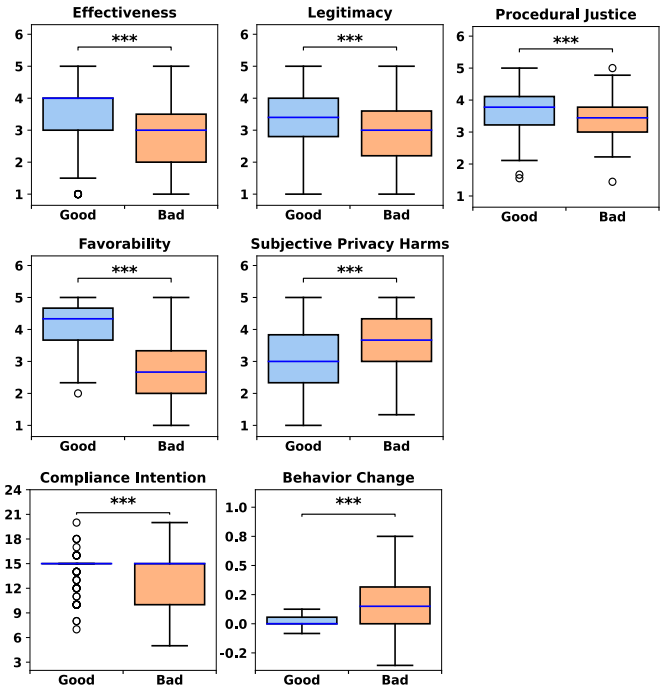
**Table 1: Impact of the types of consequences and individual differences perceptions and behavioral reactions.**

| | Effectiveness | | Procedural Justice | | Legitimacy | | Behavior Change | | Compliance Intention | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Constant | $2.716^{***}$ | $2.529^{***}$ | $3.154^{***}$ | $3.144^{***}$ | $2.896^{***}$ | $2.810^{***}$ | $0.209^{***}$ | $0.239^{***}$ | $12.251^{***}$ | $12.593^{***}$ |
| | (0.309) | (0.318) | (0.205) | (0.213) | (0.278) | (0.288) | (0.061) | (0.063) | (0.861) | (0.890) |
| Good outcome | $0.722^{***}$ | $0.978^{***}$ | $0.336^{***}$ | $0.350^{**}$ | $0.445^{***}$ | $0.563^{***}$ | $-0.184^{***}$ | $-0.224^{***}$ | $1.839^{***}$ | $1.368^{**}$ |
| | (0.116) | (0.162) | (0.077) | (0.108) | (0.105) | (0.147) | (0.023) | (0.032) | (0.325) | (0.454) |
| Low-stakes | −0.116 | 0.278 | $-0.113^{+}$ | −0.092 | −0.052 | 0.130 | −0.013 | $-0.074^{+}$ | −0.109 | −0.832 |
| | (0.100) | (0.201) | (0.066) | (0.134) | (0.090) | (0.182) | (0.020) | (0.040) | (0.279) | (0.562) |
| General privacy concerns | −0.011 | −0.016 | −0.017 | −0.017 | −0.064 | −0.067 | −0.005 | −0.004 | −0.033 | −0.023 |
| | (0.064) | (0.064) | (0.043) | (0.043) | (0.058) | (0.058) | (0.013) | (0.013) | (0.179) | (0.179) |
| Computer literacy | 0.069 | 0.075 | $0.115^{**}$ | $0.115^{**}$ | 0.090 | $0.093^{+}$ | 0.007 | 0.006 | 0.200 | 0.189 |
| | (0.063) | (0.062) | (0.042) | (0.042) | (0.056) | (0.056) | (0.012) | (0.012) | (0.174) | (0.174) |
| Good outcome x Low-stakes | | $-0.523^{*}$ | | −0.027 | | −0.241 | | $0.082^{+}$ | | 0.959 |
| | | (0.231) | | (0.155) | | (0.209) | | (0.046) | | (0.647) |
| Observations | 375 | 375 | 375 | 375 | 375 | 375 | 375 | 375 | 375 | 375 |
| $R^2$ | 0.097 | 0.109 | 0.069 | 0.069 | 0.054 | 0.058 | 0.151 | 0.158 | 0.082 | 0.087 |
| Adjusted $R^2$ | 0.087 | 0.097 | 0.059 | 0.057 | 0.044 | 0.045 | 0.142 | 0.147 | 0.072 | 0.075 |

*Note:* $^{+}$p<0.1;$^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001




**(a) Interaction effects between outcome and decision importance (stakes) on people's perceptions and behavioral reactions.**




**(b) Two-sample *t-tests* of perceptions, experiences, and behavioral reactions between people with a good and bad outcome. $^{***}$: $p$<0.001**

**Figure 2: Illustration of the impact of the types of consequences on perceptions, experiences, and behavioral reactions.**

there was a strong indirect negative association between subjective privacy harms and perceived legitimacy, which was stronger for those with a good ($\beta$=-0.71, $p$<0.001), compared to those with a bad outcome ($\beta$=-0.49, $p$<0.001). The structural models are displayed in Figure 3. Overviews of the effects and correlations are given in Tables 4 and 5 (Appendix). We evaluated the prospective power of the model in terms of detecting differences between the local paths Fav→PJ, SPH→PJ, SPH→Leg, and Leg→Comp of the two outcomes. To detect a difference of 0.3 a sample size of 50 observations per group is required, yielding a power of 95% [39, 55].

## 6 DISCUSSION AND CONCLUDING REMARKS

In this work, we investigated the impact of the types of consequences in social scoring systems, in terms of the outcome and the decision importance, on people's perceptions and behavioral reactions. Our results show that the outcome people received was pivotal for determining their experiences with the system, what they thought of the system, as well as how they reacted to a social scoring system. In addition, the outcome strongly shaped the impact of people's experiences, in terms of outcome favorability and subjective privacy harms, on their perceptions and behavioral

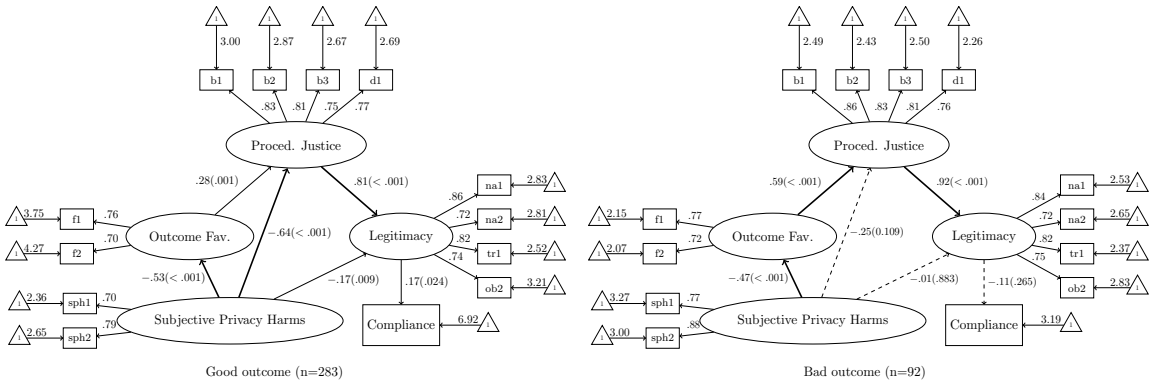



**Figure 3: Path estimates (*p-values*) of the scoring model. Dashed lines indicate insignificant effects, normal lines significant effects (0.001<*p*<0.1), thick lines significant effects with *p*<0.001. Manifest Variables (MVs) are indicated by rectangles, latent variables (LVs) by ellipses, intercepts by triangles. Paths from LVs to MVs indicate factor loadings. Robust CFI: 0.97, Robust RMSEA: 0.05, SRMR: 0.04.**

reactions. Specifically, the impact of the outcome favorability on the perceived procedural justice, legitimacy, as well as on the intention to comply was more significantly and more strongly pronounced for those with a bad outcome. The divergent impact of outcome favorability between those with a good and those with a bad outcome is likely due to the much weaker perceived procedural justice of those with a bad outcome [13].

Subjective privacy harms were associated with a decrease in people's outcome favorability, for both good and bad outcomes, which follows the established relationship between privacy concerns and satisfaction [2, 19, 53]. However, subjective privacy harms were negatively associated with perceived procedural justice and legitimacy only for people with a good outcome; this may be because people with a good outcome are forced to reveal behavioral information to have access to benefits. Those with a bad score, in contrast, do not necessarily share their score. Following identity theory, people who value privacy generally seek to control what others get to know, and thus regulate the public picture of themselves conveyed to others [4, 74]. A transparent score gives a precise indication of past behaviors, which may have caused a lack of control over which information is shared, and thus induced the negative association between subjective privacy harms and perceived procedural justice and legitimacy. Nonetheless, people with a good outcome felt their privacy was significantly less invaded compared to people with a bad outcome. This may point to a rationalization of privacy-related harms; the benefits associated with the use of ADM systems can lead people to judge the costs of privacy invasion less negatively, and, therefore, induce them to trade away privacy concerns [31].

In terms of the impact of the decision importance on people's perceptions and behavioral reactions, we found that applying social scores to a high-stakes compared to a low-stakes decision did neither affect people's perceptions of procedural justice, legitimacy, nor their intention to comply. This result stands in stark contrast to previous research; commonly, perceptions of fairness or justice [8, 43, 58] and perceptions of legitimacy [52] of ADM systems sharply decrease when moving from low-stakes to high-stakes decision scenarios. Yet, disentangling the concept of procedural justice, we found that people with a good outcome felt significantly more respected in a high-stakes compared to a low-stakes scenario. Moreover, applying social scores to high-stakes decisions increased the perceived effectiveness of social scoring systems, but only for those with a good outcome. This underscores the importance of studying perceptions from a multi-dimensional angle, as a one-dimensional assessment, e.g., of fairness, may prevent researchers from developing a nuanced understanding of the consequences of different aspects of social scoring systems [90].

People with a bad outcome showed a strong intention to adapt their behavior after learning the consequences attached to their score, even in a low-stakes decision context. This behavioral mechanism can be interpreted from several angles. Rationally, the non-accessibility of rewards may have constituted a perceived loss, and thus balanced the gains from non-compliance to such an extent that non-compliance becomes unprofitable [5]. An alternative interpretation is based on a relational authority follower model, which assesses the relationship between sanctions and compliance, considering both the moral message the authority communicates through the sanction, as well as individuals' perception of the sanction [56, 57]. Specifically, the model postulates that a sanction can convey a moral disapproval of non-compliance, but only if the sanction is perceived as retributive, and not as a fine through which non-compliance can be compensated. The mechanism that sanctions can increase compliance through the communication of moral disapproval only applies when the deployed procedures of the authority are considered just [56]. In the context of our experiment, this model suggests that those with a bad outcome considered the consequences, even those in a low-stakes context, as retributive. It consequently seems that even low-stakes implications communicated that non-compliance is morally wrong. However, those with a bad outcome reported low perceptions of procedural justice and legitimacy. Further, our SEM showed that opposed to those with a good outcome, the intention to comply for those with a bad outcome was neither associated with perceived legitimacy nor with perceived procedural justice. For those with a bad outcome, the motivation for compliance was thus unrelated to legitimate viewpoints.

In this context, literature in social psychology has established that an authority can use distinct power structures to make people comply. Besides legitimate mechanisms, compliance can also be achieved through coercion. Coercion is typically evoked by close monitoring and punishment of citizens [7, 33]. Those with a bad outcome thus may have voiced the intention to comply based on a subtle feeling of being coerced, even in the hypothetical scenario of our experiment. Following the slippery slope framework [41], the kind of power an authority applies has considerable implications, as it impacts cognitions of the authority, changes the underlying motivation for compliance, and shapes the overall climate emerging in a society subject to the decision-making of an authority [33, 41]. In this context, it is worth noting that voluntary compliance is substantially furthered by a respectful treatment of individuals [32]. However, those with a bad outcome felt significantly less respected than those with a good outcome. This exacerbates the concern that social scoring systems may create substantial controversies.

To develop a more comprehensive understanding of the dynamics evolving in a social scoring system, several aspects require future investigation. The behavior change of those with a bad outcome implies that people would behave differently in the future. This may affect people's scores, and, consequently, their perceptions of the system. Future efforts should study the dynamic effects of social scoring on people's perceptions and behavioral reactions, mimicking their "feedback-loop" character [23]. Moreover, the observation that higher computer literacy was associated with more positive perceptions of procedural justice should be taken notice of. Future research should explore the potential risk that social scoring systems create adverse impacts on individuals who differ in their education or computer literacy.

Contextualizing our findings within our prior research on social scoring, people's perceptions of procedural justice, as well as of legitimacy were considerably lower in a setting that violates contextual integrity compared to a setting that maintains contextual integrity [50], notably even for those with a good outcome. Our work thus underscores the careful approach of the EU, which aims to impose limitations on scoring practices that violate contextual integrity [30]. However, most real-world practices related to social scoring clearly violate contextual integrity. Moreover, the EU has yet to offer a concrete definition of the contextual integrity concept. Future work needs to assess different degrees of contextual integrity, and investigate whether the maintaining of this concept would suffice for mitigating the opposing dynamics that the scoring outcome creates. Understanding these aspects is imperative to shape the legal boundaries of social scoring practices more narrowly.

## ACKNOWLEDGMENTS

We wish to thank the anonymous reviewers for their valuable feedback. We further thank Felix Fischer, Mo Chen, Emmanuel Syrmoudis and the facilitators of the ExperimenTUM lab and the MELESSA 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

[1] 2021. Kultur Token - Klima schonen und Kultur genießen. https://digitales.wien.gv.at/projekt/kultur-token/. Last acessed on April 16, 2024.
[2] Mousa Ahmed Albashrawi and Luvai Motiwalla. 2019. Privacy and Personalization in Continued Usage Intention of Mobile Banking: An Integrative Perspective. *Information Systems Frontiers* 21 (2019), 1031–1043. https://doi.org/10.1007/s10796-017-9814-7
[3] Martin Alessandro, Bruno Cardinale Lagomarsino, Carlos Scartascini, Jorge Streb, and Jerónimo Torrealday. 2021. Transparency and Trust in Government. Evidence from a Survey Experiment. *World Development* 138, Article 105223 (2021), 18 pages. https://doi.org/10.1016/j.worlddev.2020.105223
[4] Bradley J. Alge. 2001. Effects of Computer Surveillance on Perceptions of Privacy and Procedural Justice. *Journal of Applied Psychology* 86, 4 (2001), 797–804. https://doi.org/10.1037/0021-9010.86.4.797
[5] Michael G. Allingham and Agnar Sandmo. 1972. Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics* 1, 3 (1972), 323–338. https://doi.org/10.1016/0047-2727(72)90010-2
[6] Maureen L. Ambrose and Carol T. Kulik. 1989. The Influence of Social Comparisons on Perceptions of Procedural Fairness. *Journal of Business and Psychology* 4, 1 (1989), 129–138. http://www.jstor.org/stable/25092220
[7] James Andreoni, Brian Erard, and Jonathan Feinstein. 1998. Tax Compliance. *Journal of Economic Literature* 36, 2 (1998), 818–860. https://www.jstor.org/stable/2565123
[8] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes de Vreese. 2020. In AI We Trust? Perceptions about Automated Decision-making by Artificial Intelligence. *AI & SOCIETY* 35 (2020), 611–623. https://doi.org/10.1007/s00146-019-00931-w
[9] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior* 10, 1 (1995), 122–142. https://doi.org/10.1006/game.1995.1027
[10] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 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)*. ACM, 1–14. https://doi.org/10.1145/3173574.3173951
[11] David Boos. 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.
[12] Joel Brockner. 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 (2002), 58–76. http://www.jstor.org/stable/4134369
[13] Joel Brockner and Batia Mishan Wiesenfeld. 1996. An Integrative Framework for Explaining Reactions to Decisions: Interactive Effects of Outcomes and Procedures. *Psychological Bulletin* 120, 2 (1996), 189–208. https://doi.org/10.1037/0033-2909.120.2.189
[14] Jenna Burrell. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3, 1, Article 2053951715622512 (2016), 12 pages. https://doi.org/10.1177/2053951715622512
[15] Ryan M. Calo. 2011. The Boundaries of Privacy Harm. *Indiana Law Journal* 86, 3 (2011), 1132–1162. https://www.repository.law.indiana.edu/ilj/vol86/iss3/8
[16] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825. https://doi.org/10.1177/0022243719851788
[17] Mo Chen, Severin Engelmann, and Jens Grossklags. 2023. Social Credit System and Privacy. In *The Routledge Handbook of Privacy and Social Media*, Sabine Trepte and Philipp K. Masur (Eds.). Routledge, New York, NY, 227–236. https://doi.org/10.4324/9781003244677-26
[18] Tsai-Wei Chen and S. Shyam Sundar. 2018. This App Would Like to Use Your Current Location to Better Serve You: Importance of User Assent and System Transparency in Personalized Mobile Services. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 1–13. https://doi.org/10.1145/3173574.3174111
[19] Yang Cheng and Hua Jiang. 2020. How Do AI-driven Chatbots Impact User Experience? Examining Gratifications, Perceived Privacy Risk, Satisfaction, Loyalty, and Continued Use. *Journal of Broadcasting & Electronic Media* 64, 4 (2020), 592–614. https://doi.org/10.1080/08838151.2020.1834296
[20] Hyesun Choung, Prabu David, and Arun Ross. 2023. Trust in AI and its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543
[21] Danielle Citron and Frank Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89, 1 (2014), 1–33. https://digitalcommons.law.uw.edu/wlr/vol89/iss1/2
[22] Rogier Creemers. 2018. *China's Social Credit System: An Evolving Practice of Control.* SSRN Working Paper 3175792. https://doi.org/10.2139/ssrn.3175792

[23] Nello Cristianini and Teresa Scantamburlo. 2020. On Social Machines for Algorithmic Regulation. *AI & Society* 35 (2020), 645–662. https://doi.org/10.1007/s00146-019-00917-8

[24] Xin Dai. 2018. *Toward a Reputation State: The Social Credit System Project of China.* SSRN Working Paper 3193577. https://doi.org/10.2139/ssrn.3193577

[25] Eric S. Dickson, Sanford C. Gordon, and Gregory A. Huber. 2022. Identifying Legitimacy: Experimental Evidence on Compliance with Authority. *Science Advances* 8, 7 (2022). https://doi.org/10.1126/sciadv.abj7377

[26] Tamara Dinev and Paul Hart. 2006. An Extended Privacy Calculus Model for E-commerce Transactions. *Information Systems Research* 17, 1 (2006), 61–80. http://www.jstor.org/stable/23015781

[27] Erik R. Eddy, Dianna L. Stone, and Eugene E. Stone-Romero. 1999. The Effects of Information Management Policies on Reactions to Human Resource Information Systems: An Integration of Privacy and Procedural Justice Perspectives. *Personnel Psychology* 52, 2 (1999), 335–358. https://doi.org/10.1111/j.1744-6570.1999.tb00164.x

[28] Severin Engelmann, Mo Chen, Lorenz Dang, and Jens Grossklags. 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)*. ACM, 78–88. https://doi.org/10.1145/3461702.3462535

[29] Severin Engelmann, Mo Chen, Felix Fischer, Ching-Yu Kao, and Jens Grossklags. 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)*. ACM, 69–78. https://doi.org/10.1145/3287560.3287585

[30] 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. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[31] Nathanael J. Fast and Arthur S. Jago. 2020. Privacy Matters… or Does It? Algorithms, Rationalization, and the Erosion of Concern for Privacy. *Current Opinion in Psychology* 31 (2020), 44–48. https://doi.org/10.1016/j.copsyc.2019.07.011

[32] Lars P. Feld and Bruno S. Frey. 2018. Illegal, Immoral, Fattening or What?: How Deterrence and Responsive Regulation Shape Tax Morale. In *Size, Causes and Consequences of the Underground Economy*, Christopher Bajada and Friedrich Schneider (Eds.). Routledge, New York, NY, 15–37.

[33] Katharina Gangl, Eva Hofmann, and Erich Kirchler. 2015. Tax Authorities' Interaction with Taxpayers: A Conception of Compliance in Social Dilemmas by Power and Trust. *New Ideas in Psychology* 37 (2015), 13–23. https://doi.org/10.1016/j.newideapsych.2014.12.001

[34] Frederic Gerdon, Ruben L. Bach, Christoph Kern, and Frauke Kreuter. 2022. Social Impacts of Algorithmic Decision-Making: A Research Agenda for the Social Sciences. *Big Data & Society* 9, 1, Article 20539517221089305 (2022), 13 pages. https://doi.org/10.1177/20539517221089305

[35] Anne-Britt Gran, Peter Booth, and Taina Bucher. 2021. To Be or Not to Be Algorithm Aware: A Question of a New Digital Divide? *Information, Communication & Society* 24, 12 (2021), 1779–1796. https://doi.org/10.1080/1369118X.2020.1736124

[36] Anna Hornik, Georg Klose, Thomas Stehnken, Florian Spalthoff, Holger Glockner, Christian Grünwald, Daniel Bonin, and Julian Sachs. 2020. Zukunft von Wertvorstellungen der Menschen in Unserem Land. Bundesministerium für Bildung und Forschung. https://www.informatik.fb2.frankfurt-university.de/~jschaefer/assets/BMBFForesightWertestudieLangfassung.pdf

[37] Matthew Hornsey, Leda Blackwood, Winnifred Louis, Kelly Fielding, Ken Mavor, Thomas Morton, Anne O'Brien, Karl-Erik Paasonen, Joanne Smith, and Katherine White. 2006. Why Do People Engage in Collective Action? Revisiting the Role of Perceived Effectiveness. *Journal of Applied Social Psychology* 36 (2006), 1701–1722. https://doi.org/10.1111/j.0021-9029.2006.00077.x

[38] Jonathan Jackson, Ben Bradford, Chris Giacomantonio, and Rebecca Mugford. 2023. Developing Core National Indicators of Public Attitudes towards the Police in Canada. *Policing and Society* 33, 3 (2023), 276–295. https://doi.org/10.1080/10439463.2022.2102757

[39] Lisa Jobst, Martina Bader, and Morten Moshagen. 2023. A Tutorial on Assessing Statistical Power and Determining Sample Size for Structural Equation Models. *Psychological Methods* 28, 1 (2023), 207–221. https://doi.org/10.1037/met0000423

[40] Mohsen Jozani, Emmanuel Ayaburi, Myung Ko, and Kim-Kwang Raymond Choo. 2020. Privacy Concerns and Benefits of Engagement with Social Media-Enabled Apps: A Privacy Calculus Perspective. *Computers in Human Behavior* 107, Article 106260 (2020), 15 pages. https://doi.org/10.1016/j.chb.2020.106260

[41] Erich Kirchler, Erik Hoelzl, and Ingrid Wahl. 2008. Enforced versus Voluntary Tax Compliance: The "Slippery Slope" Framework. *Journal of Economic Psychology* 29, 2 (2008), 210–225. https://doi.org/10.1016/j.joep.2007.05.004

[42] Rex B. Kline. 2016. *Principles and Practice of Structural Equation Modeling (4th ed.).* The Guilford Press, New York, NY.

[43] Markus Langer, Cornelius J. König, and Maria Papathanasiou. 2019. Highly Automated Job Interviews: Acceptance under the Influence of Stakes. *International Journal of Selection and Assessment* 27 (2019), 217–234. https://doi.org/10.1111/ijsa.12246

[44] Ashworth Laurence and Clinton Free. 2006. Marketing Dataveillance and Digital Privacy: Using Theories of Justice to Understand Consumers' Online Privacy Concerns. *Journal of Business Ethics* 67 (2006), 107–123. https://doi.org/10.1007/s10551-006-9007-7

[45] Min Kyung Lee and Su Baykal. 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)*. ACM, 1035–1048. https://doi.org/10.1145/2998181.2998230

[46] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 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 (2019), 1–26. https://doi.org/10.1145/3359284

[47] Min Kyung Lee and Katherine Rich. 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)*. ACM, 1–14. https://doi.org/10.1145/3411764.3445570

[48] Roy Levy and Gregory R. Hancock. 2007. A Framework of Statistical Tests For Comparing Mean and Covariance Structure Models. *Multivariate Behavioral Research* 42, 1 (2007), 33–66. https://doi.org/10.1080/00273170701329112

[49] Haili Li and Genia Kostka. 2022. Accepting but not Engaging with it: Digital Participation in Local Government-run Social Credit Systems in China. *Policy & Internet* 14, 4 (2022), 845–874. https://doi.org/10.1002/poi3.316

[50] Carmen Loefflad, Mo Chen, and Jens Grossklags. 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)*. AIS, Article 13. https://aisel.aisnet.org/icis2023/iot_smartcity/iot_smartcity/13

[51] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–16. https://doi.org/10.1145/3313831.3376727

[52] Kirsten Martin and Ari Waldman. 2022. Are Algorithmic Decisions Legitimate? The Effect of Process and Outcomes on Perceptions of Legitimacy of AI Decisions. *Journal of Business Ethics* 183, 3 (2022), 653–670. https://doi.org/10.1007/s10551-021-05032-7

[53] Juan-Gabriel Martínez-Navalón, Vera Gelashvili, and Alba Gómez-Ortega. 2021. Evaluation of User Satisfaction and Trust of Review Platforms: Analysis of the Impact of Privacy and E-WOM in the Case of TripAdvisor. *Frontiers in Psychology* 12, Article 750527 (2021), 12 pages. https://doi.org/10.3389/fpsyg.2021.750527

[54] Lorraine Mazerolle, Emma Antrobus, Sarah Bennett, and Tom R. Tyler. 2013. Shaping Citizen Perceptions of Police Legitimacy: A Randomized Field Trial of Procedural Justice. *Criminology* 51, 1 (2013), 33–63. https://doi.org/10.1111/j.1745-9125.2012.00289.x

[55] Morten Moshagen. 2022. semPower: Power Analyses for SEM (R package Version 1.2.0). https://cloud.r-project.org/web/packages/semPower/semPower.pdf.

[56] Laetitia B. Mulder. 2009. The Two-Fold Influence of Sanctions on Moral Concerns. In *Psychological Perspectives on Ethical Behavior and Decision Making*, David De Cremer (Ed.). Information Age Publishing, Inc., Charlotte, NC, 169–180.

[57] Laetitia B. Mulder. 2018. When Sanctions Convey Moral Norms. *European Journal of Law and Economics* 46 (2018), 331–342. https://doi.org/10.1007/s10657-016-9532-5

[58] Rosanna Nagtegaal. 2021. The Impact of Using Algorithms for Managerial Decisions on Public Employees' Procedural Justice. *Government Information Quarterly* 38, 1, Article 101536 (2021), 10 pages. https://doi.org/10.1016/j.giq.2020.101536

[59] Sue Newell and Marco Marabelli. 2015. Strategic Opportunities (and Challenges) of Algorithmic Decision-making: A Call for Action on the Long-term Societal Effects of 'Datification'. *The Journal of Strategic Information Systems* 24, 1 (2015), 3–14. https://doi.org/10.1016/j.jsis.2015.02.001

[60] Brian P. Niehoff and Robert H. Moorman. 1993. Justice as a Mediator of the Relationship between Methods of Monitoring and Organizational Citizenship Behavior. *The Academy of Management Journal* 36, 3 (1993), 527–556. http://www.jstor.org/stable/256591

[61] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (2004), 119–157. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10

[62] Tim O'Reilly. 2013. Open Data and Algorithmic Regulation. In *Beyond Transparency: Open Data and the Future of Civic Innovation*, Brett Goldstein and Lauren Dyson (Eds.). Code for America Press, San Francisco, CA, Chapter 22, 289–300. https://beyondtransparency.org/

[63] Robert J. Sampson, Stephen W. Raudenbush, and Felton Earls. 1997. Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. *Science* 277, 5328 (1997), 918–924. https://doi.org/10.1126/science.277.5328.918

[64] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 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). ACM, 1616–1628. https://doi.org/10.1145/3531146.3533218
[65] David E. Silva, Chan Chen, and Ying Zhu. 2022. Facets of Algorithmic Literacy: Information, Experience, and Individual Factors Predict Attitudes toward Algorithmic Systems. *New Media & Society*, Article 14614448221098042 (2022). https://doi.org/10.1177/14614448221098042
[66] Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. 2001. E-privacy in 2nd Generation E-commerce: Privacy Preferences Versus Actual Behavior. In *Proceedings of the 3rd ACM Conference on Electronic Commerce (EC '01)*. ACM, 38–47. https://doi.org/10.1145/501158.501163
[67] State Council. 2014. *Planning Outline for the Construction of a Social Credit System (2014-2020)*. (in Chinese).
[68] Mark C. Suchman. 1995. Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review* 20, 3 (1995), 571–610. http://www.jstor.org/stable/258788
[69] S. Shyam Sundar and Sampada Marathe. 2010. Personalization versus Customization: The Importance of Agency, Privacy, and Power Usage. *Human Communication Research* 36 (2010), 298–322. https://doi.org/10.1111/j.1468-2958.2010.01377.x
[70] Jason Sunshine and Tom R. Tyler. 2003. The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing. *Law & Society Review* 37, 3 (2003), 513–548. http://www.jstor.org/stable/1555077
[71] John W. Thibaut and Laurens Walker. 1975. *Procedural Justice: A Psychological Analysis*. L. Erlbaum Associates, Hillsdale, NJ.
[72] Neil Thurman, Judith Moeller, Natali Helberger, and Damian Trilling. 2019. My Friends, Editors, Algorithms, and I. *Digital Journalism* 7, 4 (2019), 447–469. https://doi.org/10.1080/21670811.2018.1493936
[73] Meng-Jung Tsai, Ching-Yeh Wang, and Po-Fen Hsu. 2019. Developing the Computer Programming Self-Efficacy Scale for Computer Literacy Education. *Journal of Educational Computing Research* 56, 8 (2019), 1345–1360. https://doi.org/10.1177/0735633117746747
[74] John C. Turner and Rina S. Onorato. 1999. Social Identity, Personality, and the Self-concept: A Self-categorizing Perspective. In *The Psychology of the Social Self*, Tom R. Tyler, Roderick M. Kramer, and Oliver P. John (Eds.). Psychology Press, New York, NY, 11–46. https://doi.org/10.4324/9781315805689
[75] Tom R. Tyler. 2006. Psychological Perspectives on Legitimacy and Legitimation. *Annual Review of Psychology* 57 (2006), 375–400. https://doi.org/10.1146/annurev.psych.57.102904.190038
[76] Tom R. Tyler. 2006. *Why People Obey the Law*. Princeton University Press, Princeton, NJ. http://www.jstor.org/stable/j.ctv1j66769
[77] Tom R. Tyler and Steven L. Blader. 2003. The Group Engagement Model: Procedural Justice, Social Identity, and Cooperative Behavior. *Personality and Social Psychology Review* 7, 4 (2003), 349–361. https://doi.org/10.1207/S15327957PSPR0704_07
[78] Tom R. Tyler and Jeffrey Fagan. 2008. Legitimacy and Cooperation: Why Do People Help the Police Fight Crime in Their Communities? *Ohio State Journal of Criminal Law* 6 (2008), 231–276. https://scholarship.law.columbia.edu/faculty_scholarship/414
[79] Tom R. Tyler and Yuen J. Huo. 2002. *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*. Russell Sage Foundation, New York, NY. http://www.jstor.org/stable/10.7758/9781610445429
[80] Tom R. Tyler and Jonathan Jackson. 2013. Popular Legitimacy and the Exercise of Legal Authority: Motivating Compliance, Cooperation and Engagement. *Psychology Public Policy and Law* 20, 1 (2013), 78–95. https://doi.org/10.1037/a0034514
[81] Tom R. Tyler and E. Lind. 1992. A Relational Model of Authority in Groups. *Advances in Experimental Social Psychology* 25 (1992), 115–192. https://doi.org/10.1016/S0065-2601(08)60283-X
[82] Michael Veale and Frederik Zuiderveen Borgesius. 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 (2021), 97–112. https://doi.org/10.9785/cri-2021-220402
[83] Peter Verboon and Marius van Dijke. 2011. When do Severe Sanctions Enhance Compliance? The Role of Procedural Fairness. *Journal of Economic Psychology* 32, 1 (2011), 120–130. https://doi.org/10.1016/j.joep.2010.09.007
[84] Ari Waldman and Kirsten Martin. 2022. Governing Algorithmic Decisions: The Role of Decision Importance and Governance on Perceived Legitimacy of Algorithmic Decisions. *Big Data & Society* 9, 1 (2022), 16 pages. https://doi.org/10.1177/20539517221100449
[85] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 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)*. ACM, 1–14. https://doi.org/10.1145/3313831.3376813
[86] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 4, Article 27 (2022). https://doi.org/10.1145/3519266
[87] Keith F. Widaman. 1987. *Journal of Educational Statistics* 12, 3 (1987), 308–313. http://www.jstor.org/stable/1164692
[88] Eva Wolfangel. 2022. Nein, Bayern bereitet keine Überwachung chinesischer Art vor. *Zeit* (2022). https://www.zeit.de/digital/datenschutz/2022-07/oeko-token-bayern-belohnungssystem-social-scoring Last accessed on April 16, 2024.
[89] Karen Yeung. 2018. Algorithmic Regulation: A Critical Interrogation. *Regulation & Governance* 12, 4 (2018), 505–523. https://doi.org/10.1111/rego.12158
[90] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 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)*. ACM, 1–21. https://doi.org/10.1145/3544548.3581161
[91] Tal Zarsky. 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 (2016), 118–132. http://www.jstor.org/stable/43671285

# A APPENDICES

## A.1 Scenario

This subsection displays the scenarios people read after the experimental phase. First, a short introduction was given. Then, depending on the treatment, a low-stakes or a high-stakes context was presented. The outcome that people read was adapted to the score they reached in the experimental phase.

**Introduction:** The city in which you live has introduced a social scoring system. With the introduction of the system, the city aims at encouraging good behaviors among citizens and promoting social engagement. Under the social scoring system, behavioral scores are calculated to assess how well citizens behave and how engaged they are. The scoring system in the city functions in a similar manner to the scoring system you experienced in the experiment. A high score shows that a person would value fairness and equality, and indicates a person would be socially engaged. A low score shows that a person would not particularly value fairness and equality, and indicates a person would not be particularly socially engaged.

**Low stakes:** Good behavioral scores can be exchanged for benefits in cultural institutions. Several cultural institutions in your city take part in the program. They consider the score in the pricing and offering of tickets. In some cultural institutions people with high scores can enjoy reduced entrance fees. In other cultural institutions, people with high scores also receive priority seats. Showing the score when buying tickets or visiting a cultural event is voluntary.

- **Good outcome:** Based on how you treated other community members in this experiment, you would receive a score of [...] in the social scoring system of your city. This is considered a high score, and shows that you would value fairness and equality, and indicates that you would be socially engaged. Therefore, you can enjoy benefits offered by cultural institutions, such as privileged access to limited tickets, or discounts on regular entrance fees.
- **Bad outcome:** Based on how you treated other community members in this experiment, you would receive a score of [...] in the social scoring system of your city. This is considered a low score, and shows that you would not particularly value fairness and equality, and indicates that you would not be particularly socially engaged. Therefore, you cannot enjoy benefits offered by cultural institutions, such as privileged access to limited tickets, or discounts on regular entrance fees.

**High stakes:** Good behavioral scores can be exchanged for benefits in public institutions. Several public institutions in your city

take part in the program. They consider the score when deciding about access to services and goods. Suppose you have just finished your bachelor's degree. You plan to apply for the restricted master's program at a prestigious university in your city. The university aims at contributing to the city's initiatives of encouraging good behaviors and social engagement among citizens. With the introduction of the scoring system, the university now includes behavioral scores from the scoring system in their admission processes. In their admission process for restricted admission master's programs, the university primarily considers students' final grades from their undergraduate studies. A number of applicants with exceptional grades are directly admitted to the program. For the remaining students, the university takes into account both applicant's final grades, as well as their social score. High scores improve an applicant's bachelor's grade, and increase the chance of being admitted. Submitting a score to the application process is voluntary. You have completed your bachelor's degree with success, but your grades are not sufficient to be considered exceptional. Therefore, you do not receive direct admission.

- **Good outcome:** Based on how you treated other community members in this experiment, you would receive a score of [...] in the social scoring system of your city. This is considered a high score, and shows that you would value fairness and equality, and indicates that you would be socially engaged. Your good score improves your bachelor's grade. Consequently, you are admitted to the master's program.
- **Bad outcome:** Based on how you treated other community members in this experiment, you would receive a score of [...] in the social scoring system of your city. This is considered a low score, and shows that you would not particularly value fairness and equality, and indicates that you would not be particularly socially engaged. Therefore, your bachelor's grade is not improved. As your bachelor's grade alone is not sufficient, you are not admitted to the master's program.

## A.2 Additional Tables and Figures

In this section, the survey questions for the pre-experimental and the post-experimental survey are displayed. In addition, we provide complementary information to the structural equation model, in terms of an overview of the direct and indirect effects, as well as the correlation matrix of the manifest variables. We also present an example of the experimental interface, which was used to educate participants about the experiment.

**Table 2: Survey questions measuring favorability (Fav), subjective privacy harms (SPH), perceived effectiveness (Eff), perceived procedural justice (PJ), and perceived legitimacy (Leg). A manipulation check (Manip) was added.**

| Abbreviation | Question |
|---|---|
| **Fav** | |
| f1 | I generally like how I am treated in the scoring system. |
| f2 | The introduction of the scoring system is favorable to me. |
| f3 | I am satisfied with my score. |
| **SPH** | |
| ctrl1 | I feel controlled by the scoring system. |
| sph1 | I think the collection of behavioral data under the scoring system is disadvantageous to me. |
| sph2 | I feel uncomfortable that the scoring system collects data about my behavior. |
| **Eff** | |
| eff1 | I think the scoring system will successfully influence people to behave in fair manner. |
| eff2 | Having the scoring system in the city is useful to me. |
| **PJ** | |
| b1 | I think the scoring system will lead to an increased well-being in the city. |
| b2 | I think the scoring system will make decisions that are good for everyone in the city. |
| b3 | I think the scoring system will be necessary to help people in the city to become more socially engaged. |
| d1 | I think under the scoring system people will be treated with dignity and respect. |
| c1 | I understand the rules and methods the scoring system uses to decide on my score. |
| t1 | I understand how my behaviors impact my score. |
| t2 | I think the mechanism the scoring system uses to calculate my score is fair. |
| pc1 | I am able to influence the data that the scoring system considers to evaluate my score. |
| pc2 | I am able to influence my score such that I am satisfied with it. |
| **Leg** | |
| ob1 | I feel a duty to comply with the behaviors that the scoring system promotes. |
| ob2 | I think that complying with the behaviors the system propagates benefits everyone in the city. |
| na1 | I generally support how the scoring system acts. |
| na2 | The scoring system has the same sense of right and wrong as I do. |
| tr1 | I can trust the scoring system to make the right decisions. |
| **Manip** | |
| m1 | In a person's life, how important do you consider the decision to be admitted to a master's program? |
| m2 | In a person's life, how important do you consider the decision to receive benefits in cultural institutions? |

**Table 3: Pre-survey questions before the experiment, measuring computer literacy (CL) and general privacy concerns (GPC). Participants should indicate the extent to which they agreed with each statement, on a 5-point Likert scale.**

| Abbreviation | Question |
|---|---|
| **CL** | |
| cl1 | I have good programming skills. |
| cl2 | I have good knowledge of computer algorithms. |
| cl3 | I can make use of programming to solve a problem. |
| cl4 | I understand how my e-mail provider's spam filter works. |
| cl5 | I understand how Amazon recommends products for me to purchase. |
| cl6 | A little bit of intuition is all that is needed to figure out how to use any new technology. |
| **GPC** | |
| gpc1 | I am concerned that personal data I leave online might be misused. |
| gpc2 | I am concerned about providing information to online websites because of what others might do with it. |
| gpc3 | I am concerned about providing information to online websites because it could be used in ways I cannot predict. |
| gpc4 | I am concerned that others can find private information about me online. |

**Table 4: Overview of direct effects (DE) and indirect effects (IE) of the structural equation model.**

| **Bad outcome** | Criterion | DE | IE | SE | CI-low | CI-high | p-value |
|---|---|---|---|---|---|---|---|
| Subj. Privacy Harm ⟶ | Favorability | -0.47*** | | 0.09 | -0.65 | -0.30 | <0.001 |
| | Proced. Justice | -0.25 | | 0.15 | -0.55 | 0.05 | 0.109 |
| | Proced. Justice | | -0.28** | 0.09 | -0.44 | -0.11 | 0.001 |
| | Legitimacy | -0.01 | | 0.08 | -0.17 | 0.15 | 0.883 |
| | Legitimacy | | -0.49*** | 0.11 | -0.71 | -0.27 | <0.001 |
| | Compliance | | 0.03 | 0.03 | -0.03 | 0.08 | 0.332 |
| Favorability ⟶ | Proced. Justice | 0.59*** | | 0.13 | 0.33 | 0.85 | <0.001 |
| | Legitimacy | | 0.54*** | 0.12 | 0.30 | 0.77 | <0.001 |
| | Compliance | | -0.06 | 0.05 | -0.16 | 0.05 | 0.300 |
| Procedural Justice ⟶ | Legitimacy | 0.92*** | | 0.06 | 0.81 | 1.03 | <0.001 |
| | Compliance | | -0.10 | 0.09 | -0.27 | 0.07 | 0.266 |
| Legitimacy ⟶ | Compliance | -0.11 | | 0.09 | -0.29 | 0.07 | 0.265 |
| **Good outcome** | | | | | | | |
| Subj. Privacy Harm ⟶ | Favorability | -0.53*** | | 0.08 | -0.68 | -0.38 | <0.001 |
| | Proced. Justice | -0.64*** | | 0.08 | -0.79 | -0.49 | <0.001 |
| | Proced. Justice | | -0.13** | 0.04 | -0.22 | -0.05 | 0.002 |
| | Legitimacy | -0.17** | | 0.06 | -0.29 | -0.05 | 0.009 |
| | Legitimacy | | -0.71*** | 0.06 | -0.83 | -0.58 | <0.001 |
| | Compliance | | 0.03 | 0.03 | -0.02 | 0.09 | 0.284 |
| Favorability ⟶ | Proced. Justice | 0.28** | | 0.08 | 0.13 | 0.44 | 0.001 |
| | Legitimacy | | 0.26** | 0.08 | 0.11 | 0.41 | 0.001 |
| | Compliance | | 0.04+ | 0.02 | 0.00 | 0.08 | 0.075 |
| Procedural Justice ⟶ | Legitimacy | 0.81*** | | 0.05 | 0.71 | 0.92 | <0.001 |
| | Compliance | | 0.15* | 0.05 | 0.04 | 0.26 | 0.025 |
| Legitimacy ⟶ | Compliance | 0.17** | | 0.06 | 0.05 | 0.28 | 0.024 |

**Table 5: Correlation of manifest variables used in the structural equation model. ***: $p$<0.001, **: $p$<0.01, *: $p$<0.05.**

| | b1 | b2 | b3 | d1 | na1 | na2 | ob2 | tr1 | f1 | f2 | sph1 | sph2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| b1 | | | | | | | | | | | | |
| b2 | 0.71*** | | | | | | | | | | | |
| b3 | 0.64*** | 0.59*** | | | | | | | | | | |
| d1 | 0.66*** | 0.61*** | 0.64*** | | | | | | | | | |
| na1 | 0.68*** | 0.63*** | 0.60*** | 0.62*** | | | | | | | | |
| na2 | 0.55*** | 0.59*** | 0.45*** | 0.49*** | 0.61*** | | | | | | | |
| ob2 | 0.62*** | 0.63*** | 0.56*** | 0.53*** | 0.65*** | 0.54*** | | | | | | |
| tr1 | 0.60*** | 0.67*** | 0.59*** | 0.57*** | 0.70*** | 0.65*** | 0.57*** | | | | | |
| f1 | 0.43*** | 0.38*** | 0.42*** | 0.36*** | 0.48*** | 0.33*** | 0.40*** | 0.44*** | | | | |
| f2 | 0.38*** | 0.35*** | 0.33*** | 0.31*** | 0.41*** | 0.32*** | 0.39*** | 0.37*** | 0.65*** | | | |
| sph1 | -0.44*** | -0.42*** | -0.37*** | -0.37*** | -0.43*** | -0.39*** | -0.48*** | -0.45*** | -0.38*** | -0.47*** | | |
| sph2 | -0.49*** | -0.46*** | -0.49*** | -0.48*** | -0.55*** | -0.37*** | -0.43*** | -0.49*** | -0.31*** | -0.28*** | 0.57*** | |
| comp | 0.07 | 0.06 | 0.06 | 0.07 | 0.09 | 0.12* | 0.10 | 0.07 | 0.14** | 0.10* | -0.18*** | -0.08 |

### Example Table

After each round, you will be shown a summary of your previous interactions. Below is an exemplary table, which provides a summary of past interactions after Round 2. The entries in blue are those that determine your payoff. They will not be highlighted in the experiment.

- Round 1: You are the first mover. Both you and your match partner have an endowment of 10. You decided to send 8. Therefore, your match partner received 16. Your match partner sent back 5. As a result, you earned 7 monetary units (10-8+5 =7) and your match partner earned 21 monetary units (10+16-5=21). As you are the first mover your score is not affected by the interaction. Your score remains 1000.
- Round 2: You are the second mover, and your match partner is the first mover. Both you and your match partner have an endowment of 10. Your match partner sent you 10. Therefore, you received 20. You sent back 12. As a result, your payoff amounts to 18 monetary units (10+20-12=18) and that of your match partner to 12 monetary units (10-10+12=12). Your score after the interaction is 997.

| | **Roles** | | **Endowments** | | **Interactions** | | | **Payoffs** | | **Score** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Round** | **My role** | **Role of match partner** | **Your endowment** | **Endowment of match partner** | **Money sent by first mover** | **Money received by second mover** | **Money returned by second mover** | **Your payoff** | **Payoff of match partner** | **Your score** |
| 1 | **First mover** | Second mover | **10** | 10 | **8** | 16 | **5** | **7** | 21 | 1000 |
| 2 | **Second mover** | First mover | **10** | 10 | 10 | **20** | **12** | **18** | 12 | 997 |

**Figure 4: Examples used to educate participants about the scoring mechanism.**