



Information Systems Research

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies

Kevin Bauer, Andrej Gill

To cite this article:

Kevin Bauer, Andrej Gill (2023) Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies. Information Systems Research

Published online in Articles in Advance 03 May 2023

. <https://doi.org/10.1287/isre.2023.1217>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2023 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies

Kevin Bauer,^{a,*} Andrej Gill^b

^aInformation Systems Department, University of Mannheim, 68161 Mannheim, Germany; ^bEconomics and Management, Johannes Gutenberg-University, 55128 Mainz, Germany

*Corresponding author

Contact: kevin.bauer@uni-mannheim.de,  <https://orcid.org/0000-0001-8172-1261> (KB); gill@uni-mainz.de,  <https://orcid.org/0000-0003-1013-0979> (AG)

Received: July 23, 2021

Revised: August 6, 2022; January 8, 2023

Accepted: February 11, 2023


Published Online in Articles in Advance: May 3, 2023

<https://doi.org/10.1287/isre.2023.1217>

Copyright: © 2023 The Author(s)

Abstract. Predictive algorithmic scores can significantly impact the lives of assessed individuals by shaping decisions of organizations and institutions that affect them, for example, influencing the hiring prospects of job applicants or the release of defendants on bail. To better protect people and provide them the opportunity to appeal their algorithmic assessments, data privacy advocates and regulators increasingly push for disclosing the scores and their use in decision-making processes to scored individuals. Although inherently important, the response of scored individuals to such algorithmic transparency is understudied and therefore demands further research. Inspired by psychological and economic theories of information processing, we aim to fill this gap. We conducted a comprehensive experimental study with five treatment conditions to explore how and why disclosing the use of algorithmic scoring processes to (involuntarily) scored individuals affects their behaviors. Our results provide strong evidence that the disclosure of fundamentally erroneous algorithmic scores evokes self-fulfilling prophecies that endogenously steer the behavior of scored individuals toward their assessment, enabling algorithms to help produce the world they predict. Occurring self-fulfilling prophecies are consistent with an anchoring effect and the exploitation of available moral wiggle room. Because scored individuals interpret others' motives for overriding human expert and algorithmic scores differently, self-fulfilling prophecies occur in part only when disclosing algorithmic scores. Our results emphasize that isolated transparency measures can have considerable side effects with noticeable implications for the development of automation bias, the occurrence of feedback loops, and the design of transparency regulations.

History: Santanam Raghu, Senior Editor; Gordon Burtch, Associate Editor.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as "Information Systems Research. Copyright © 2023 The Author(s). <https://doi.org/10.1287/isre.2023.1217>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc-nd/4.0/>."

Funding: The authors acknowledge financial support from the Leibniz Institute for Financial Research SAFE and the Johannes Gutenberg-University Mainz.

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/isre.2023.1217>.

Keywords: algorithmic transparency • algorithmic scoring • self-fulfilling prophecies • second-order beliefs • algorithm aversion

1. Introduction

Private organizations and public institutions increasingly rely on predictive algorithmic methods to augment their decision-making processes. A report by BCC Research (BCC 2021) estimates the global market for predictive analytics to grow from \$11.1 billion in 2022 to \$23.9 billion by 2027, representing a compound annual growth rate of 16.5% for the five years between 2022 and 2027.

Predictive algorithmic scores already shape millions of people's social and economic well-being and are often unnoticed. Examples include human resource (HR) managers who consult algorithmic scores to assess applicants' future job performance during hiring processes

(Tambe et al. 2019), insurance agents who use machine-learning models to compute prospective customers' risk premiums (Huang and Meng 2019), investors leveraging credit or profit scores to estimate borrowers' default risks (Serrano-Cinca and Gutiérrez-Nieto 2016), and judges who invoke algorithmically generated recidivism scores to make bail decisions (Kleinberg et al. 2018). The promise of algorithmic scoring processes is twofold. On the one hand, they aim to alleviate inefficient information asymmetries so that individuals relying on the score, users of the algorithm, can make better decisions that affect scored individuals: data subjects.¹ On the other hand, algorithms supposedly ensure that human

hunches, subjective perceptions, and moods no longer distort the evaluation of data subjects and associated decisions that concern them.

However, prominent examples such as the Correctional Offender Management Profiling for Alternative Sanctions system (Angwin et al. 2016) and Amazon's machine-learning recruitment tool (Dastin 2018) show that these algorithmic methods can spectacularly fail in making good on their promises. Because scores can have substantial personal consequences, for example, not being hired or not being released on bail, researchers, activists, and regulators increasingly advocate disclosing the outcome, use, and workings of algorithmic scoring processes to the public, especially to scored individuals themselves (Goodman and Flaxman 2017, Cabral 2021). Following Louis Brandeis's notion that "Sunlight is said to be the best disinfectant," such algorithmic transparency aims to enhance data subjects' understanding of the process and effectively enable them to contest scoring results. According to the OECD, more than 60 countries have implemented or proposed more than 700 policy instruments regulating the use of algorithmic assessment and decision-making systems from 2017 onward (OECD AI 2021). The European Union, for instance, put into effect the General Data Protection Regulation in 2018 and proposed the even more comprehensive Artificial Intelligence Act in 2021, regulating the disclosure, storage, and processing of personal data (Parliament and Council of European Union 2016, 2021).

Although the growing number of transparency requirements is generally a welcome trend projected to balance power asymmetries between users of algorithmic systems and (unknowingly or involuntarily) scored individuals, potential downstream consequences of shedding more light on modern algorithmic scoring processes have only recently come to the attention of researchers. A nascent literature focuses on the impact of explanations about why algorithms produce certain outputs on user behaviors, showing that explainability affects users' perceptions, attitudes, and reliance on the system (Rader et al. 2018, Dodge et al. 2019, Poursabzi-Sangdeh et al. 2021, Senoner et al. 2021). Another stream in this literature explores the consequences of disclosing algorithmic prediction performances to users, producing evidence that it improves perceptions of the system (Warshaw et al. 2015, You et al. 2022). However, a surprisingly understudied facet of algorithmic transparency is the reaction of scored individuals to learning that a private organization or public institution used a certain algorithmic score to augment decisions that affect them. Exploring potential downstream ramifications regarding the more vulnerable party, whom the transparency aims to empower, constitutes a pivotal research objective. Our study takes the first step to filling this research gap. More specifically, this study aims to answer two key questions:

(1) Will disclosing the result and the use of predictive algorithmic scores in decision-making processes to

scored individuals endogenously affect the behavior the algorithm tries to predict, and if so, why?

(2) What is the role of the algorithmic nature of disclosed scores, that is, do scored individuals respond differently when algorithmic systems or human experts perform the assessment?

As an illustration of the scenarios we have in mind, consider an individual A who applies for a job. The responsible HR manager H observes an algorithmic prediction of applicants' future job performance that the HR manager can use when making the hiring decision. H finally decides to hire A who accepts the offer. After the final decision, because of regulations (Parliament and Council of European Union 2016, 2021), H discloses to A that H had access to an algorithmic scoring method that predicted A to be a future low performer. In this scenario, we are interested in the following questions. Does learning about the result and use of the algorithmic scoring process affect the applicant's actual performance on the job, and if so, why? Would the applicant's on-the-job performance be different if a human expert and not an algorithm produced his score? The existence of such side effects associated with the disclosure of algorithmic scoring processes has important implications for organizations and policymakers alike.

One tacit assumption behind the employment of algorithmic predictions is that their use does not affect the outcome they aim to forecast. Yet, theories about individual belief formation processes (Chapman and Johnson 2002, Dana et al. 2007) rooted in psychology and behavioral economics suggest that disclosing algorithmic scores to data subjects may do just that. As a result, algorithmic transparency may endogenously influence the score's accuracy. In our study, we leverage these theories, more specifically, theories on anchoring and moral wiggle room exploitation, to build the conceptual foundation of how disclosing scores can affect scored individuals' behavior.

We address our research questions using a comprehensive experiment where we incentivize participants to act according to their true preferences and reveal their true first- and second-order beliefs that underlie their behavior. Our findings show that disclosing the result and use of inaccurate algorithmic assessment processes to scored individuals steers their behavior in the direction of revealed scores, that is, creates self-fulfilling prophecies. The occurrence of self-fulfilling prophecies is unique to algorithmic scoring if the decision-maker who uses the score chooses to override it. This pattern seems to originate from scored individuals' beliefs about why decision-makers override the score: a pattern that may hint at a broader phenomenon we call *second-order algorithm aversion*.

Our results show that disclosing the use of algorithmic scoring processes after the fact in isolation provides a channel through which algorithms can help create the

world they predict. As we discuss, this unintended side effect has important implications for the development of automation bias, the occurrence of feedback loops, and the design of transparency regulations.

The paper proceeds as follows. In Section 2, we present the theoretical foundation of our work and outline our contribution to existing literature. Section 3 explains our study design. We present our results in Section 4. Section 5 concludes with a discussion about our study's implications and limitations and an outlook on future research directions.

2. Theoretical Background

The following sections describe theories that guide our research and summarize related work. We first explain the role of beliefs in individuals' decision-making processes (Section 2.1). Subsequently, we outline our work's contribution to the existing literature (Section 2.2).

2.1. Role of Beliefs in Decision Making

Research provides ample evidence that beliefs are an essential component of decision-making processes. Among others, beliefs about a system's usefulness shape users' adoption of information systems (Davis 1989, Benlian et al. 2012), trustworthiness beliefs determine trusting behaviors in online environments (McKnight et al. 2002, Kim and Benbasat 2006), reciprocity beliefs influence the willingness to contribute to and use knowledge management systems (Bock et al. 2005, 2006), and beliefs about prevailing social norms affect reporting rates for fake news (Gimpel et al. 2021). In this study, we assert that disclosing the result and use of algorithmic scoring processes to data subjects may affect their beliefs about expected or condoned behaviors. These "second-order beliefs" are integral to individuals' decision-making processes (Battigalli and Dufwenberg 2007, Goldstein et al. 2008) and components of popular psychological models such as the *Theory of Planned Behavior* (Ajzen 1989, Mathieson 1991).

When individuals' behavior is at odds with their second-order beliefs, they typically experience a disutility or discomfort, for example, due to feelings of guilt, shame, or anxiety, associated with disappointing others' expectations or violating a perceived standard (Baumeister et al. 1994, Krupka and Weber 2013). Trying to avoid this psychological unease in the first place, individuals tend to behave in a way compliant with what they believe is expected of them. For instance, employees may not only exert effort at work because they anticipate a reward or because their boss monitors them. At least in part, they may also exert effort because they intrinsically want to meet their firm's standards.

Importantly, beliefs are rarely stationary. Individuals frequently adjust their beliefs, and more broadly, their mental constructs, in response to encountering new

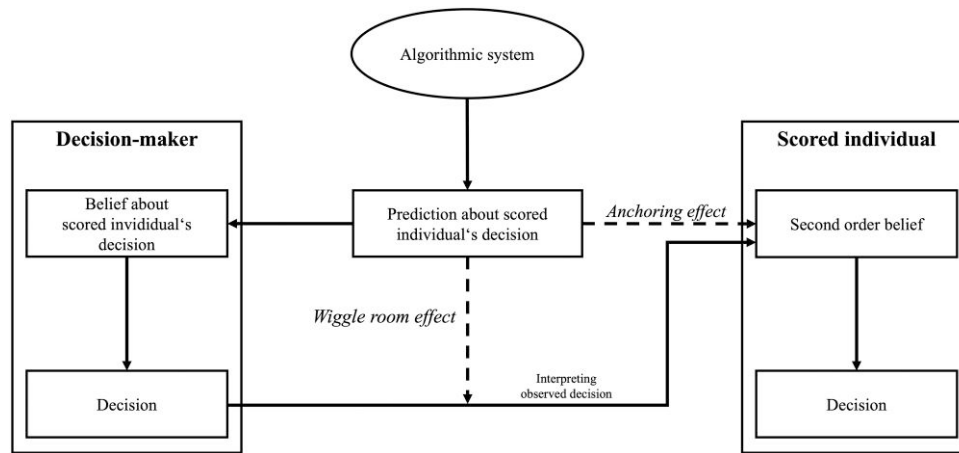
information from their environment (Grether 1980, Vandenbosch and Higgins 1996, Holt and Smith 2009). We argue that the disclosure of predictive algorithmic scores and their use in a decision-making process that affects scored individuals provide information relevant to beliefs. Specifically, under this form of transparency, assessed individuals effectively obtain two novel pieces of information that can change their second-order beliefs: (i) the result of their algorithmic assessment and (ii) a decision maker's awareness of it when making a decision that affects them. We assert and empirically test that these two pieces of information influence assessed individuals' second-order beliefs due to an *anchoring effect* (Chapman and Johnson 2002) and the provision of *moral wiggle room* (Dana et al. 2007), respectively.

Figure 1 illustrates the model we have in mind. The starting point is an algorithmic system that predicts a scored individual's behavior, shaping a decision maker's beliefs about the scored person's behavior and thus their decision. The scored individual observes this decision and, importantly, also the employment of the scoring process. On the one hand, observing a personal score may anchor the scored individual's beliefs about what is expected of them. On the other hand, knowing that the decision maker was aware of this score provides an informational ambiguity to interpret the observed decision's intention in a self-serving way and opportunistically adjust second-order beliefs. By changing the scored individual's second-order beliefs, the algorithmic transparency may endogenously alter the behavior the system tries to predict, thereby influencing whether predictions are "correct."

In the following, we describe anchoring and wiggle room effects in detail, outlining how they relate to the disclosure of algorithmic scoring processes.

2.1.1. Anchoring Effect. Anchoring effects describe a cognitive bias whereby attitudes, behaviors, and beliefs are skewed toward a provided reference point, that is, an anchor (Tversky and Kahneman 1974). The provided anchor is a cue that can be of both a quantitative or qualitative nature (Yasseri and Reher 2022). Previous research has shown that the occurrence of anchoring effects is a very robust phenomenon that spans numerous domains. For instance, anchoring biases occur in the domain of legal decisions (Englich et al. 2006), purchasing decisions (Ariely et al. 2003), valuation decisions (Northcraft and Neale 1987), and even self-efficacy beliefs (Cervone and Peake 1986).

The literature provides different explanations for the occurrence of anchoring biases. One view posits that anchoring biases originate from individuals applying the anchor-and-adjustment heuristic (Tversky and Kahneman 1974). This explanation presumes that individuals use the anchor as an initial starting point for searching the appropriate judgment. Individuals then

Figure 1. Disclosed Algorithmic Scores, Beliefs, and Behaviors

insufficiently adjust their beliefs away from the original anchor value toward a solution that appears subjectively more plausible (Jacowitz and Kahneman 1995).

The currently predominant explanation asserts that a confirmatory hypothesis testing underlies the anchoring bias (Chapman and Johnson 1999, Mussweiler and Strack 1999). According to this notion, individuals consider the provided anchor as a reference point that constitutes a plausible belief in a situation where they assume that an objective standard exists. Once they observed the anchor, individuals selectively form arguments or retrieve the knowledge consistent with the anchor. To do so, individuals effectively activate the accessibility of information, opinions, and experiences in memory that are anchor confirming. Notably, research has found that this retrieval process is independent of the informational relevance of the anchors (Englich et al. 2006). Following the selective retrieval of anchor-consistent knowledge, individuals integrate this information to adjust their beliefs that eventually translate into changed behaviors in the decision-making process (Chapman and Johnson 2002).

A third explanation brought forward reconciles the previous two interpretations. This view on anchoring draws on the processes of attitude change and assumes that anchoring can originate from a relatively non-thoughtful application of the anchor-and-adjust heuristic or a relatively thoughtful confirmatory hypothesis testing. Which process occurs depends on individuals' motivation and capability to exert cognitive efforts (Blanchenship et al. 2008).

We conjecture that disclosing algorithmic scoring processes to scored individuals evokes anchoring effects because the score represents a first reference point about what behaviors are condoned in the context in which the algorithmic scoring occurs. From this initial reference point, scored individuals would then either insufficiently

adjust their judgment about what they should do or retrieve anchor-confirming experiences from memory (conditional on cognitive resources the decision-making process requires). In that sense, the disclosed score can be interpreted as an indication of the expectations of the broader socio-technological environment where the algorithmic scoring process takes place. Notably, the anchoring effect is fundamentally independent of other people's awareness of the score and their ability to use it when making a decision; it only stems from scored individuals' personal knowledge of the score. In the previous hiring example, disclosing to hired candidates that the algorithm predicts them to be a low performer may, independent of the knowledge that the HR manager knew the score when hiring them, set a relatively low initial reference point about performance standards. Insufficiently adjusting from this anchor, the hired candidates adopt low second-order beliefs causing their job performance to converge toward the disclosed score, that is, creating a self-fulfilling prophecy.

2.1.2. Moral Wiggle Room. Through disclosing algorithmic assessments, data subjects not only observe the score as such. Instead, they also become aware that the score may have influenced the decision-making process that has consequences for them. When data subjects learn that the decision-making person was aware of their algorithmic score but effectively chose to override it, they can make opportunistic inferences about the decision maker's expectations and intentions, possibly giving them an excuse to behave selfishly without feeling psychological discomfort. In social situations, individuals care about what involved people expect them to do, feeling guilty if they violate these expectations (Baumeister et al. 1994; Battigalli and Dufwenberg 2007, 2009). Although an aversion to feeling guilty or disappointing others typically evokes expectation-meeting

behaviors, individuals tend to exploit informational ambiguities, that is, “moral wiggle room,” to justify self-ish, expectation-violating behaviors and avoid associated psychological and moral costs (Dana et al. 2007, Grossman and Van Der Weele 2017).

Researchers first studied the tendency to construct and subsequently exploit opportunistic beliefs in the context of honesty and generosity (Dana et al. 2007, Larson and Capra 2009, Grossman 2014). These studies find that subjects in controlled experiments involving distributional decisions often willfully avoid information about how their actions affect others. In their seminal paper, Dana et al. (2007) show that an option to remain ignorant about the consequences of one’s decisions serves as moral wiggle room for behaving selfishly at the expense of others while maintaining a positive self-image. Similarly, Haisley and Weber (2010) report that individuals leverage wiggle room to construct self-serving beliefs that they exploit to justify unfair behavior. Additional studies find people to exploit moral wiggle room in abundant other domains, including reciprocal decision-making processes (Regner 2018), charity donations (Exley 2016), online feedback giving (Bolton et al. 2019), and contributions to carbon offsets (Momsen and Ohndorf 2022).

More broadly, one can understand the construction and exploitation of moral wiggle room as a manifestation of the motivated reasoning phenomenon. Research on motivated reasoning consistently reveals that individuals selectively avoid, distort, or misinterpret information to form or maintain a particular set of self-serving beliefs (Dunning 1999, Balcetis and Dunning 2006, Uhlmann et al. 2009, Epley and Gilovich 2016). These biased beliefs, which feel objective to individuals, enable them to pursue directed goals, that is, intrinsically or extrinsically preferred outcomes (Kunda 1990). In other words, their preferences regarding the result of the reasoning process frequently shape people’s cognitive processes for forming and updating beliefs.

In the context of our study, we contend that disclosing algorithmic scoring processes provides moral wiggle room when the decision-maker seemingly overrides the algorithmic assessment result. Specifically, learning that the decision maker deliberately chose not to adhere to the predictive score creates informational ambiguity about this person’s motives and expectations about choice consequences. Overriding an algorithmic score and its implied recommendation can, in principle, have a multitude of causes, for example, a lack of trust in the system, the absence of explanations for assessment results, or a biased discounting of machine advice (Jussupow et al. 2020). However, scored individuals can opportunistically interpret the decision to override the score as signaling an indifference toward the attitude, trait, or behavior the algorithm aims to predict. As a result, scored individuals can self-servingly adjust their

beliefs about what the decision maker expects them to do. Exploiting this opportunistic belief allows them to behave more selfishly at the expense of the decision maker without feeling guilty about it. For example, applicants who was hired even though an algorithm indicated to the HR manager that they were not suitable for the job might opportunistically conclude that the manager is indifferent about their job performance. This belief allows hired applicants to save effort costs for high performance while avoiding psychological costs from disappointing the HR manager who hired them.

2.2. Contribution to the Literature

Our paper complements three streams of literature.

2.2.1. Algorithmic Transparency. First, we contribute to the growing literature exploring the consequences of algorithmic transparency. Algorithmic transparency broadly refers to disclosing the use, workings, and outcome of algorithmic methods to affected individuals. The increasing adoption of algorithms in consequential domains such as healthcare (Jussupow et al. 2021), finance (Ban et al. 2018), and hiring (van den Broek et al. 2021) has sparked the interest in algorithmic transparency by academics, practitioners, and regulators alike. Proponents of algorithmic transparency argue that it can alleviate negative ramifications associated with using algorithmic systems such as machine biases (Watson and Nations 2019). Artificial intelligence (AI)-developing companies and governments are increasingly enacting policies and regulations that effectively mandate that people be informed about when and how algorithmic systems evaluate, rank, or profile them (Parliament and Council of European Union 2016, 2021; Google AI 2019; Meta AI 2021). Similarly, there is a growing number of transparency-promoting initiatives within academia (Araujo et al. 2018, Dencik et al. 2019).

Much of the research on the ramifications of algorithmic transparency focuses on explanations about why algorithms produce certain outputs, aiming to alleviate problems of accountability (Gregor and Benbasat 1999). Explainability often improves users’ trust in the system (Wang and Benbasat 2007), fairness perceptions (Dodge et al. 2019), task efficiency (Senoner et al. 2021), and understanding of the system’s malfunctions (Rader et al. 2018). However, there is also evidence of potential disadvantages related to informational overload (Poursabzi-Sangdeh et al. 2021), reduced user trust (Kizilcec 2016), and reduced accuracy perceptions (Springer and Whittaker 2018). Several studies also consider the role of disclosing algorithmic prediction performances to users, finding evidence that such a transparency intervention improves system perceptions (Warshaw et al. 2015, You et al. 2022).

Although these studies make critical contributions to our understanding of the ramifications of algorithmic

transparency, they all focus on the consequences for the system user. However, it is also crucial to understand how scored individuals whom the score-influenced decisions ultimately affect respond to algorithmic transparency. After all, these individuals are the most vulnerable stakeholders, especially if they are not even aware of being scored. To the best of our knowledge, no study has previously explored this facet of algorithmic transparency. We aim to make a first step toward filling this gap. We adopt the perspective that the most fundamental form of transparency from the perspective of scored individuals is making them aware of the result and the employment of algorithmic scoring processes in decisions that affect them. This basic form of transparency is the first decisive step to enabling data subjects to identify and challenge inaccurate assessments and avoid possibly severe personal consequences such as not receiving a loan. We investigate whether and how this algorithmic transparency affects the behaviors of data subject that the algorithm aims to predict. This way, we complement related research on algorithmic transparency by shedding light on a central, however, thus far overlooked, channel through which downstream consequences of algorithmic transparency can occur.

2.2.2. Human-Machine Interaction. The second line of research we contribute to examines the ways and consequences of the interaction between humans and machines. A large body of work in the information systems (IS) field explores how (intelligent) decision-supporting systems can affect humans' decision-making performance, providing ample evidence on efficiency-enhancing effects through information structuring (Xu et al. 2014, Jussupow et al. 2021). Notably, there is also evidence of potential risks and downsides often related to influencing users in ways beneficial to system designers (Xiao and Benbasat 2015). Previous research reveals that the employment of decision-supporting machines cannot only influence human behaviors superficially. These machines also seem capable of changing users' more deeply rooted preferences and belief structures. Häubl and Murray (2003) show how the presentation of items included in a recommender system's preference elicitation interface endogenously influenced users' revealed preferences. Using a causal mediation approach in a field experiment, Li et al. (2022) show that decision support systems endogenously change consumers' consideration sets. Adomavicius et al. (2018) provide strong evidence that product recommendations shape people's willingness to pay, emphasizing decision support systems' capabilities to render preferences. Research has also shown that displaying predicted preferences rating to consumers can bias their ex post preference ratings (Adomavicius et al. 2013), which can create considerable biases in feedback

loops that aim to improve predictions over time (Adomavicius et al. 2019).

Our study complements this literature by examining whether the previously identified endogeneity effects do not only occur when algorithmic predictions are about users themselves. We ask whether it is an even broader phenomenon that also occurs when algorithmically scored individuals observe the prediction about themselves but are not the system user. Instead, another person can leverage the prediction to make a better decision that affects the well-being of the scored individuals. This setting reflects the fundamental structure of a wide variety of scenarios such as promotion decisions, bail decisions, loan approval decisions, and hiring decisions. Exploring the occurrence of such endogeneity effects and their underlying mechanisms from the perspective of scored individuals is still lacking in the literature.

2.2.3. Algorithm Aversion and Appreciation. Finally, because we explore whether effects associated with disclosing scores to assessed individuals are idiosyncratic to algorithmic scoring, or occur equally for human expert assessments, our study also relates to the ongoing debate about algorithm aversion and algorithm appreciation. Algorithm aversion refers to human decision makers' reluctance to rely on superior yet imperfect algorithms (and prefer human advice) (Dietvorst et al. 2015), whereas algorithm appreciation describes the inclination to adjust more toward advice from algorithms than humans (Logg et al. 2019). Previous research in this domain dates back decades, producing mixed results. Several studies demonstrate that aversion to algorithms is the dominant phenomenon (Dawes 1979, Mackay and Elam 1992), whereas others find human decision makers to prefer algorithmic over human advice (Sanders and Courtney 1985, Dijkstra 1999). With the rise of contemporary machine-learning systems, the examination of human decision-makers under- or overreliance has seen a considerable resurgence, for example, in the domain of robo-advisory (Ge et al. 2021), automated customer service (Schanke et al. 2021), medical decision making (Chan et al. 2020), and efficient task delegation (Fügner et al. 2021). A large body of work has examined factors influencing the occurrence of algorithm aversion and appreciation (for an excellent review, see Jussupow et al. 2020). Whether human decision makers underuse algorithmic advice depends on their task experience (Prah and Van Swol 2017), unmet expectations about the algorithm's characteristics (Castelo et al. 2019), the perceived loss of decision autonomy (Scherer et al. 2015), and the presence of a "human-in-the-loop" both during the development (Jago 2019) or production stage (Palmeira and Spassova 2015, Dietvorst et al. 2018).

We complement this literature by exploring the existence of second-order effects of algorithm aversion. Although the previous literature considers how people's

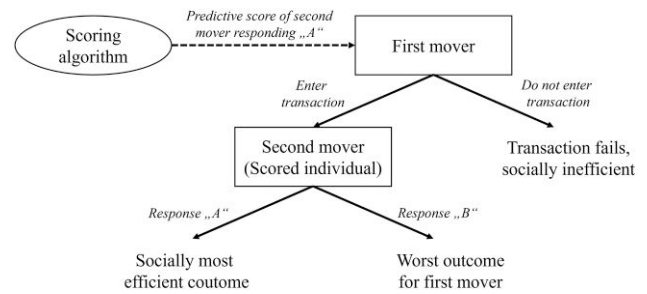
advice-taking differs when the advice comes from a human, an algorithm, or a hybrid system, we examine their response to learning that algorithms or human experts influence decisions that affect them. Do data subjects value or respond to another person's decision differently when this person relied on algorithmic or human expert advice? Does it matter whether data subjects expect the other person to exhibit algorithm aversion or appreciation? To the best of our knowledge, we are the first to explore the existence of such effects.

3. Study Design

Our objective is to explore how and why disclosing algorithmic scoring processes impacts the beliefs and behaviors of scored individuals who are affected by decisions that the score informs. There are several obstacles to exploring this issue and identify causal mechanisms. First, answering our questions requires careful measurement of data subjects' beliefs about the algorithm and the decision-maker's expectations to avoid endogeneity problems. Second, to avoid introducing confounds, the choice to use or disclose algorithmic scoring needs to be out of the control of the decision maker who can use the score. Third, assessments by optimized algorithmic systems are necessarily unique and nonrandom. Hence, in the field, it is virtually impossible to observe data subjects' responses to disclosing counterfactual algorithmic scores and thus identify causal relations. Fourth, the degree to which decision makers can override algorithmic scores depends on a variety of factors that we need to control for, including organizational constraints, personal preferences, and prior task experience. These challenges make it particularly difficult, if not outright impracticable, to answer our core research questions using a natural field setting.

To address these challenges, we developed a novel experimental design. The main task is a sequential transaction that mirrors the fundamental structure of many economic interactions where algorithmic scores about the second moving party inform decisions of the first moving party (see Figure 2 for an illustration of the transaction sequence). Examples include algorithmically supported employment decisions (first mover: HR manager; second mover: applicant), credit approval decisions (first mover: loan officer; second mover: potential borrower), and bail decisions (first mover: judge; second mover: defendant). In our main experiment, participants always take on the role of the second moving party, that is, the scored individual. Using an economic experiment to identify causal effects answers prominent calls for using this methodology in the IS domain (Gupta et al. 2018) and is in line with previous IS studies that successfully used controlled experiments (Jiang and Benbasat 2007, Adomavicius et al. 2018, Fügenger et al. 2022).

Figure 2. Algorithmically Supported Transaction Sequence



Note. Participants on our main study always take on the role of the second mover.

Our preregistered experiment² comprised two consecutive stages. In stage 1, participants answered survey questions that served as a basis to score them algorithmically. In stage 2, participants engaged in the sequential transaction, always taking on the role of the second mover. First movers are subjects from a pilot study. We implemented the transaction as a one-shot investment game (Berg et al. 1995): a reliably and widely used experimental paradigm (Fehr and Fischbacher 2003). Our main between-subject treatment manipulation is disclosing to second movers what their algorithmically generated score is and that the first mover can use it when making the initial decision. To isolate causal mechanisms, we implemented additional control treatments where second movers learn that the first mover did not observe their predictive score or that human experts instead of an algorithm produced the score. In the following, we fill in the experimental details.

3.1. Study Details

3.1.1. Stage 1: Eliciting Personal Information. In Stage 1, participants filled out a questionnaire containing items on 15 personal traits. Twelve of these traits served as the basis for the predictive scoring, and the remaining three serve as additional controls in our analyses (see Table 26 in the online appendix for an overview). At this point, we did not inform participants about the purpose of the questionnaire, allowing us to allay concerns that participants gave intentionally inaccurate answers to (perceivably) outsmart or game the system. Importantly, among items not used for the scoring are two questions about participants' reciprocal preferences from (Falk et al. 2016, 2018) that strongly correlate with people's actual second-mover behavior in standard investment games. We will rely on these measures to determine the accuracy of individual scores.

3.1.2. Stage 2: Investment Game. Stage 2 comprised a one-shot investment game (Berg et al. 1995) with the following basic structure. There are two parties: an investor and a borrower. The investor has 10 monetary units (MU) and begins with deciding whether to keep or

invest the entire 10 MU with the borrower. If the investor keeps the 10 MU, the game ends leaving her and the borrower with a payoff of 10 MU and 0 MU, respectively. If the investor decides to invest, the borrower receives triple the amount, that is, 30 MU. The borrower is then free to keep the whole amount without any repercussions. Crucially, however, the borrower can repay the investor any amount $x \in [0, 30]$ MU. The investor's and borrower's payoffs equal x MU and $30 - x$ MU, respectively. With this structure, the investment game closely mirrors sequential human transactions that require both trust by the first mover (e.g., loan officer, HR manager, supplier) and reciprocity by the second mover (e.g., borrower, worker, buyer), especially in incomplete contract situations (Fehr and Fischbacher 2003, Johnson and Mislin 2011).³

Participants in our experiment always played in the role of the borrower. Investors were participants from a pilot study that we randomly matched⁴ with borrowers to ensure incentive compatibility and thereby the reliable elicitation of preferences and beliefs (Camerer and Hogarth 1999). Borrowers had to indicate how much MU they would repay to an investor *before* knowing whether the investor actually invested. After the repayment decision, participants indicated what they believed an initially investing investor expected as repayment. If their guess did not deviate from the actual belief of the investor by more than five units, participants additionally earned 2 MU. The experiment concluded with a brief questionnaire elaborating on their perceptions of investor behaviors.

3.1.3. Baseline and Treatment Conditions. In our baseline condition, participants played the investment game in Stage 2 as outlined. We introduced our between-subject treatment variations before participants made their repayment decision: participants learned that we generated a prediction about their repayment behavior using a subset of their survey answers. We informed them that the prediction was about whether they would repay more than 10 MU to an investor, a reciprocal prediction because the investor is better off investing, or not, a nonreciprocal prediction because the investor is worse off investing.

Participants in our main treatment condition (*Algo. Public*) learned their personal prediction, that a machine-learning model made the prediction and that the investor knew the prediction before making a decision. We informed participants about the type of the machine-learning model and its workings. Importantly, we used an incentivized strategy method to observe participants' behavior for both possible predictions. The strategy method works as follows: participants had to indicate their repayment for both possible predictions before knowing what the prediction actually was. To determine actual payments, we matched investor and borrower decisions given the true prediction. Because participants did not know the prediction when making their decision, they had a strong incentive to make decisions according to their true preferences to get their most desired outcome for both possible predictions, even the one that *ex post* did not materialize. This procedure allowed us to measure the borrower responses for both actual and counterfactual predictions (see Brandts and Charness (2000) and Fischbacher et al. (2012) for the empirical validity of this strategy method approach).⁵ After making their repayment decision, participants indicated what they believed the investor expected as repayment, again for both possible predictions. Before we revealed their personal prediction, participants guessed the prediction accuracy across (baseline) participants in the experiment, and the investor's belief about the prediction accuracy. If their guesses did not deviate from the correct answers by more than five percentage points, participants additionally earned 2 MU. We also asked participants to answer several questions about their perception of the machine-learning model and a manipulation check question. The experiment ended with informing participants about their prediction and their generated income.

Our study comprised three additional control treatments to isolate causal mechanisms (see Table 1 for an overview). Control treatments merely differed from our main treatment regarding (i) the entity generating the prediction and (ii) the borrowers' knowledge about the investor's awareness of the prediction before making a decision. Specifically, in our *Algo. Private* treatment participants learned about their prediction of the machine-learning model. However, we specifically informed

Table 1. Overview of Experimental Conditions

| Condition | Borrower learns that prediction comes from | | Borrower learns that the investor | |
|----------------------|--|---------------|-----------------------------------|-----------------------------|
| | A machine-learning model | Human experts | Knew the prediction | Did not know the prediction |
| Baseline | X | X | X | X |
| <i>Algo. Public</i> | ✓ | X | ✓ | X |
| <i>Algo. Private</i> | ✓ | X | X | ✓ |
| <i>Human Public</i> | X | ✓ | ✓ | X |
| <i>Human Private</i> | X | ✓ | X | ✓ |

them that the investor had no access to this prediction when choosing to invest; that is, the prediction had not influenced the investment decision. This control treatment, which would have been extremely difficult if not outright impracticable to implement in a field setting, enables us to isolate potential anchoring effects and better understand underlying mechanisms that drive our effects in the main treatment. Our *Human Public* and *Human Private* treatments replicated our *Algo. Public* and *Algo. Private* treatments, respectively. The only difference is that participants in these two control treatments learned that the prediction comes from researchers with great expertise in behavioral science who predefined a rule to determine repayment behaviors, that is, human experts. These two control treatments allow us to isolate the role of the prediction source, or, put differently, whether borrowers behave differently given algorithmic or human expert predictions.

3.1.4. Machine-Learning Model. For the *Algo. Public* and *Algo. Private* conditions, we trained, validated, and tested an actual machine-learning model on a data set comprising 1,048 distinct observations. We collected this data in an incentivized field study that we conducted at a large German university over three years (2016–2019). Students participated online, using a link we distributed via student email addresses. The study included a comprehensive survey and an incentivized sequential social dilemma game that closely resembles the trust game used in our experiment (see online appendix for additional information). Based on their answers in the role of the borrower in the game, we categorize participants as behaving reciprocally or selfish following definitions by Miettinen et al. (2020). We then train a random forest to predict whether an individual is reciprocal, that is, repays more than 10 MU to the investor in the experiment. We developed our model in Python using popular Data Science libraries including Pandas, NumPy, and Sklearn. To avoid imbalance problems, we rebalanced our training data using the Synthetic Minority Over-sampling Technique algorithm. We also recoded our categorical variables into their dummy representation. The final structure of the forest is the result of empirical feature selection⁶ and hyperparameter tuning processes implemented as a grid search in a fivefold cross-validation on the training set. As adjustable hyperparameters we included the number of trees, the learning rate, the share of training data to build a tree, the maximum depth of a tree, and the maximum number of features a tree can use. Overall, we tested 750 hyperparameter combinations. On representative test data that we initially separated from the training data set (85%/15% split), the final model achieves an accuracy, precision, recall, and F1 score of respectively 63%, 70%, 75%, and 72% on average.⁷

3.1.5. Human Experts. The human expert predictions reflect the assessments of real scientists (PhD students,

postdoctoral researchers, and professors) who conduct research in the fields of psychology, economics, or IS. We conducted a pilot study where researchers assessed whether different borrowers behave reciprocally in the trust game that underlies the training of the machine-learning model. To make an informed assessment, researchers always observed 12 borrower characteristics: the same characteristics that the previously outlined random forest leverages. Every researcher assessed 15 random, fictitious borrowers synthesized from our field study data. There is no intermediate feedback. Using the researchers' assessments, we trained, tested, and optimized a decision tree that uses the 12 borrower characteristics to predict the researchers' assessment. This strategy allows us to produce scalable human expert assessments in a live online experiment without having to rely on a mockup. We explained to participants in the *Human Public* and *Human Private* control treatments that the assessments they observe originate from a logical rule predefined by a group of independent human experts.

3.1.6. Investors. We conducted our experiment in an incentive compatible way, that is, borrowers' decisions had real material consequences, so that their behaviors and beliefs reflect their true preferences. To ensure the incentive compatibility regarding the material well-being of investors, their repayment decisions affected the payoffs of real people who played the game in the role of the investor. Specifically, before running our main experiment with the borrowers, we conducted an incentivized pilot study where different participants acted as investors.⁸ The investment game had an identical structure to the one employed in the main borrower study. Participants in this pilot study, made five investment decisions: an investment decision without any further information; two investment decisions respectively assuming that a machine-learning model predicted the matched borrower repay more than 10 MU or not; two investment decisions respectively assuming that human experts predicted the matched borrower repay more than 10 MU or not. We informed investors that borrowers had no opportunity to manipulate predictions in their favor. After making their investment decisions they stated their repayment expectations for each of the five investment scenarios. We also asked them to state their belief about the prediction accuracy of human experts and the machine-learning system. We matched investor decisions from the pilot with borrower decisions in the main experiment to determine payoffs.⁹

3.1.7. Experimental Procedure. We implemented all our studies as computerized online experiments using oTree (Chen et al. 2016) and several Python libraries. For our main experiment with borrowers ($n = 566$) and the pilot with investors ($n = 25$), we recruited participants from the popular platform Prolific.co.¹⁰ We computed

payoffs by matching each borrower with one random investor.¹¹ We paid participants \$0.1 per MU they earned plus a fixed participation fee of \$1. Investors (borrowers), on average, needed approximately 6 (10) minutes to finish the experiment and earned \$3.44 (\$3.26). Participants from our human expert pilot ($n = 35$) were psychology, economics, and IS researchers from two large German Universities that we contacted directly via mail. We did not incentivize them. Researchers took approximately 14 minutes on average to finish the study.

4. Results

We present our results in two steps. First, we explore the impact of disclosing the outcome and the use of algorithmic assessment processes to data subjects on the behavior the algorithm tries to predict. We closely examine second-order beliefs and scores' accuracy to understand underlying mechanisms. Second, we test whether the algorithmic origin of scores plays an idiosyncratic role, creating effects that human expert assessment processes do not evoke.

4.1. Algorithmic Scores, Behaviors, and Beliefs

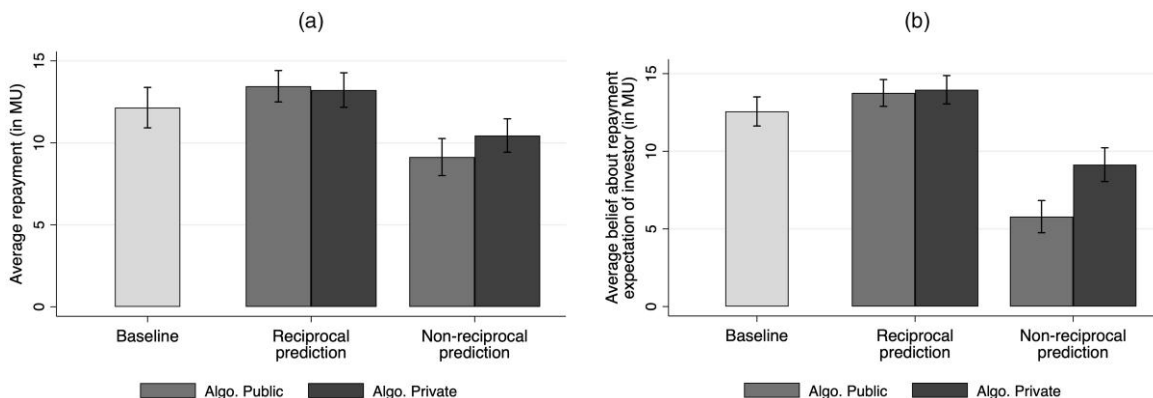
Figure 3(a) depicts how the disclosure of algorithmic scoring processes affected the borrowers' repayment behavior (see Table 4 in the online appendix for summary statistics). The visual illustration suggests that disclosing both the result and the use of algorithmic predictions about borrowers' repayment moral to borrowers endogenously shifted their actual repayment behavior in the direction of their prediction (see *Algo. Public* condition). In the baseline, borrowers, on average, repaid 12.15 MU to the investor. If borrowers became aware that an investor had observed and followed a reciprocal prediction (a prediction that they would repay more than 10 MU), they repaid 13.45 MU on average (+10.7%). Regression analyses reported in Table 2 reveal that this difference is

statistically significant ($p < 0.05$, F test). By contrast, disclosing to borrowers that the investor had observed a nonreciprocal prediction (a prediction that they would repay 10 MU or less) and invested anyway, their repayment equaled 9.13 MU on average. Compared with the baseline, this is a repayment reduction by 24.9% ($p < 0.01$, F test). Together, these findings suggest that disclosing the result and the use of the algorithmic scoring process to assessed individuals steers their behavior toward the disclosed score, that is, creates self-fulfilling prophecies. However, why is this the case? To answer this question we next analyze borrowers' behavior in the *Algo. Private* condition, where they knew that investors had not observed the prediction, and their second-order beliefs.

We first examine the repayment of borrowers in the *Algo. Private* condition to deconstruct the overall self-fulfilling prophecy into the partial effects attributable to learning about the personal score, and learning that the investor had seen this score. When learning about a reciprocal prediction, borrowers in the *Algo. Private* condition repaid 13.22 MU on average. This amount is not significantly different from the corresponding amount in the *Algo. Public* condition (-2% , $p = 0.6$, F test), indicating that the self-fulfilling prophecy for reciprocal predictions stems entirely from becoming personally aware of one's prediction. The information that the investor had been aware of this prediction did not have an additional effect. By contrast, when borrowers in the *Algo. Private* condition learned about a nonreciprocal prediction, they repaid 10.45 MU on average. Compared with the *Algo. Public* treatment, this repayment is significantly larger (+14.5%, $p < 0.04$, F test). Hence, the overall self-fulfilling prophecy for nonreciprocal predictions seems to be due to learning both one's score and the investor's knowledge of it.¹²

Overall, these results reveal two insights. On the one hand, personally learning about their prediction led borrowers to adjust repayments in the direction of the

Figure 3. Disclosing Algorithmic Assessments



Notes. We show borrowers' average repayment decisions and second-order beliefs about what they think the investors expect as a repayment. Different bars represent different experimental conditions. (a) Repayment decisions. (b) Second-order beliefs.

Table 2. Regression Analyses on Repayments and Second-Order Beliefs

| | Repayment (in MU) | | Second-order beliefs (in MU) | |
|--|-------------------------|----------------------------|------------------------------|----------------------------|
| | (1) Reciprocal pred. | (2) Nonreciprocal pred. | (3) Reciprocal pred. | (4) Nonreciprocal pred. |
| <i>Algo. Public</i> (β_1) | 1.64** (0.734) | -2.71** (0.799) | 1.297** (0.625) | -6.822*** (0.683) |
| <i>Algo. Private</i> (β_2) | 1.462* (0.794) | -1.396* (0.789) | 1.48** (0.661) | -3.489*** (0.732) |
| <i>F</i> test: $ \beta_1 - \beta_2 > 0$ | $p = 0.6$ | $p < 0.04^{**}$ | $p = 0.38$ | $p < 0.01^{***}$ |
| Individual controls | Yes | Yes | Yes | Yes |
| <i>N</i> | 352 | 352 | 352 | 352 |
| <i>R</i> ² | 0.094 | 0.096 | 0.045 | 0.175 |

Notes. We depict results from OLS regression models with robust standard errors reported in parentheses. In columns (1) and (2), we use borrowers' repayment decisions (in MU) as the dependent variable. In columns (3) and (4), we use borrowers' second-order beliefs about the investors' repayment expectations (in MU) as the dependent variable. As independent variables, we include treatment dummies. The baseline serves as the reference category. Additionally, we include controls on borrowers' age, gender, academic achievement, risk preference, and general level of reciprocity.

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

score. On the other hand, disclosing that the investor chose to invest despite observing a nonreciprocal prediction, that is, that the investor had effectively overridden the prediction, reinforced the self-fulfilling prophecy. But do these effects originate from, or at least relate to, participants' second-order beliefs? To address this question, we next examine elicited second-order beliefs.

Figure 3(b) reveals that repayment differences across different conditions closely mirror differences in borrowers' second-order beliefs. On average, borrowers in the baseline condition believed that an investor who had sent them his 10 MU expected a repayment of 12.56 MU. Learning about a reciprocal prediction, borrowers in the *Algo. Public* and *Algo. Private* condition believed that the investor expected a repayment of 13.76 MU and 13.96 MU, respectively. Compared with the baseline, these second-order beliefs are significantly higher ($p < 0.05$ for both, *F* test), yet they are not significantly different from each other ($p = 0.79$, *F* test). Learning about a nonreciprocal prediction, borrowers in the *Algo. Public* (*Algo. Private*) condition believed that the investor expected repayment of 5.79 MU (9.14 MU). Both second-order beliefs are significantly smaller compared with the baseline ($p < 0.01$, *F* test), and, more importantly, significantly different from each other ($p < 0.01$, *F* test).

Following these results, privately learning about their prediction shifted borrowers' second-order beliefs in the direction of the observed prediction. This belief adjustment is consistent with our theoretical conjecture that the disclosure of algorithmic scores anchors beliefs and thereby affects repayment decisions.

When learning that the investor invested despite observing a nonreciprocal prediction, that is, had effectively overridden the prediction, borrowers further adjust second-order beliefs in a seemingly opportunistic direction. Consistent with our conjecture about the construction of moral wiggle room, borrowers seem to

exploit the ambiguity about why investors ignored the prediction. They form the belief that the investor expected only a small repayment, although ignoring the prediction could also be meant to credibly convey trust intentions which investors typically expect to pay off (Toussaert 2017).¹³

Revealing that second-order beliefs are central to decision making, we find that repayment decisions and second-order beliefs are highly correlated (Spearman's $\rho \geq 0.29$, $p < 0.01$ for all conditions). Additionally, supporting our interpretation that second-order belief adjustments drive the self-fulfilling prophecies in repayments, we find that the coefficients for treatment variables decline to negligible levels and become statistically insignificant when we include participants' second-order beliefs as additional control variables in our regression analyses (see Table 5 in the online appendix).

Together, these findings are highly consistent with our theoretical conjecture that the disclosure of algorithmic scoring processes to assessed individuals can affect their second-order beliefs and thus the behavior the algorithm aims to predict.

Result 1. Disclosing the result and the use of an algorithmic scoring process to scored borrowers led them to adjust their repayment in the direction of the prediction, that is, created a self-fulfilling prophecy. Our findings are consistent with the notion that an anchoring effect and the exploitation of moral wiggle room shifted borrowers' second-order beliefs and, thereby, created self-fulfilling prophecies.

Result 1 shows that increased algorithmic transparency may yield critical side effects for data subjects, that is, those individuals whom transparency measures typically seek to empower. Data subjects may process the additionally provided information in a heuristic or biased way, creating unforeseen ramifications. At this point, regarding the effects occurring

when investors had overridden nonreciprocal predictions, one may naturally wonder whether this result is an artifact of our strategy method, that is, whether investors actually overrode this prediction. Investors in our corresponding pilot study overrode nonreciprocal predictions in 15.4% of the cases so that the documented effects materialize in a nonnegligible number of transactions. Naturally, from a practical perspective, the occurrence of this effect requires that the human decision maker has the discretion to override algorithms' (implicit) recommendations, which is frequently the case with "humans-in-the-loop" (Cowgill and Tucker 2020) and legally required (Parliament and Council of European Union 2021).

Thus far, we showed that the disclosure of algorithmic scoring processes can evoke self-fulfilling prophecies. However, it remains open whether the effects stem from accurate scores, correct classifications of borrowers' fundamental repayment tendencies, that reinforce existing behaviors, or inaccurate scores, incorrect classifications of borrowers' fundamental repayment tendencies, that attenuate them. For instance, did the disclosure of a nonreciprocal prediction make fundamentally nonreciprocal individuals even more selfish, or did fundamentally reciprocal individuals behave less reciprocally when they learned about a nonreciprocal prediction?

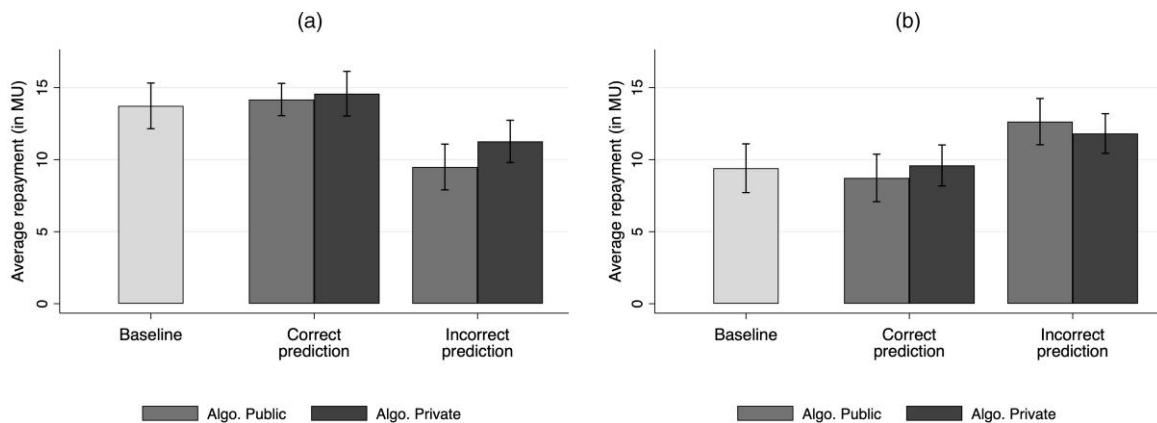
To classify borrowers' ground truth types, that is, their fundamental repayment tendency, we rely on a validated survey item that measures people's pure positive reciprocity on a seven-point scale and is strongly predictive of their repayment behavior in a standard investment game (Falk et al. 2016, 2018).¹⁴ In our baseline, where we use a standard investment game, the correlation between borrowers' answer to the survey item and their actual repayment strongly correlate (Spearman's $\rho = 0.45, p < 0.01$). We classify

borrowers as fundamentally reciprocal types with a tendency to repay more than 10 MU if their survey answer is larger or equal to the median of the distribution of our study. Otherwise, we classify them as fundamentally nonreciprocal types with a tendency to repay at most 10 MU. Providing confidence in this proxy's capability to capture borrowers' fundamental repayment tendency, we find that it correctly predicts whether baseline participants repay more than 10 MU, that is, behave reciprocally, in 63.4% of the cases.

We next examine repayments and second-order beliefs for fundamentally reciprocal types (nonreciprocal types) conditional on whether the prediction says that they will repay more than 10 MU (reciprocal prediction) or repay at most 10 MU (nonreciprocal prediction). We refer to the situations where a borrower's fundamental type, as defined by the validated survey item, and her algorithmic prediction coincide as correct and otherwise as incorrect. Notably, because the disclosure of the prediction appears to affect behaviors endogenously, our definition of a prediction's accuracy captures the hypothetical case where it remains undisclosed. From this perspective, what we refer to as correct (incorrect) prediction is best interpreted as a prediction that (in)accurately classifies a borrower's fundamental repayment tendency, which the prediction's disclosure may reinforce or attenuate. For readability, however, we use the terms correct/accurate and incorrect/inaccurate throughout the rest of this paper.

Repeating our analyses on borrowers' repayment behaviors separately for reciprocal and nonreciprocal types, we find that self-fulfilling prophecies mainly occur for cases where borrowers' fundamental repayment tendencies were incorrectly predicted. Figure 4(a) and (b) provides a visual illustration, whereas Table 3 shows regression analyses. Compared with the baseline, if the algorithm predicted fundamentally

Figure 4. Repayment Decisions by Types



Notes. For fundamentally reciprocal types, correct and incorrect predictions respectively refer to reciprocal and nonreciprocal predictions. For fundamentally nonreciprocal types, correct and incorrect predictions respectively refer to nonreciprocal and reciprocal predictions. (a) Reciprocal types. (b) Nonreciprocal types.

Table 3. Regression Analyses on Repayment Behavior for Different Borrower Types

| | Reciprocal types | | Nonreciprocal types | |
|--|----------------------|------------------------|----------------------|------------------------|
| | (1) Correct pred. | (2) Incorrect pred. | (3) Correct pred. | (4) Incorrect pred. |
| <i>Algo. Public</i> (β_1) | 0.406 (0.989) | -4.070*** (1.11) | -0.574 (1.185) | 3.465*** (1.139) |
| <i>Algo. Private</i> (β_2) | 0.946 (1.141) | -2.383** (1.1) | 0.207 (1.087) | 2.435** (1.088) |
| <i>F</i> test: $ \beta_1 - \beta_2 > 0$ | $p = 0.3$ | $p = 0.06^*$ | $p = 0.25$ | $p = 0.17$ |
| Individual controls | Yes | Yes | Yes | Yes |
| <i>N</i> | 194 | 194 | 158 | 158 |
| <i>R</i> ² | 0.033 | 0.124 | 0.073 | 0.114 |

Notes. We depict results from OLS regression models with robust standard errors reported in parentheses. In all columns, we use borrowers' repayment decisions (in MU) as the dependent variable. Columns (1) and (2) show results for reciprocal borrowers, whereas columns (3) and (4) show results for nonreciprocal borrowers. As independent variables, we include treatment dummies. The baseline serves as the reference category. Additionally, we include controls on borrowers' age, gender, academic achievement, risk preference, and general level of reciprocity.

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

reciprocal borrowers in the *Algo. Public* condition to behave nonreciprocally, they decreased their repayment by 30.9% (13.73 MU versus 9.49 MU; $p < 0.01$, *F* test). Conversely, fundamentally nonreciprocal borrowers whom the algorithm predicted to behave reciprocally increased their repayment by 32.4% (12.46 MU versus 9.41 MU; $p < 0.01$, *F* test). Privately learning that a prediction is not consistent with their fundamental repayment tendency steers both types' behavior toward the prediction, making them behave less in accordance with their fundamental repayment tendencies (for both, $p < 0.05$, *F* test). The awareness that investors had observed a nonreciprocal prediction and chosen to override it evoked an additional adjustment of repayments by fundamentally reciprocal borrowers ($p < 0.06$, *F* test). By contrast, fundamentally nonreciprocal borrowers did not additionally increase their repayment when they learned that the investor had observed a reciprocal prediction. Independent of borrowers' fundamental types, the disclosure of algorithmic predictions that are consistent with borrowers' type did not significantly affect repayment behaviors.

Treatment differences in second-order beliefs (see Figure 6, (a) and (b), and Table 6 in the online appendix) are again consistent with our conjecture that anchoring and wiggle room driven changes in second-order beliefs underlie behavioral differences. Specifically, our results suggest that becoming personally aware of an incorrect prediction changed both types' second-order beliefs in the direction of the prediction (-27.4% and +25.9% for reciprocal and nonreciprocal types, respectively). When fundamentally reciprocal types learned that the investor observed a nonreciprocal prediction but invested anyways, they opportunistically lowered their second order beliefs even more (-35%, $p < 0.01$, *F* test). Notably, we find some evidence that nonreciprocal predictions also shifted nonreciprocal types' beliefs (-22.5%

and -17.3%). However, the effects are substantially less pronounced for these types and do not translate into behavioral changes. The finding that nonreciprocal borrowers do not exploit seemingly opportunistic beliefs is in line with previous research on moral wiggle room showing that informational ambiguities are typically exploited by individuals who, in the absence of an excuse, behave prosocially (Dana et al. 2007); after all, inherently selfish individuals do not need an excuse to behave nonreciprocally in the first place.¹⁵

In sum, our analyses reveal that disclosing the result and use of algorithmic scoring processes to scored individuals can create side effects if algorithmically produced scores incorrectly classify their fundamental behavioral tendencies. We do not find evidence that disclosing correct scores does so too. Thus, documented self-fulfilling prophecies seem to occur because scored individuals act less according to their fundamental (behavioral) inclinations and more as predicted by the inaccurate algorithmic scores. Because self-fulfilling prophecies endogenously alter the observed ground truth labels of data subjects, algorithmic transparency affects the accuracy of the scoring system, allowing it to create the (mis)predicted world without necessarily being noticed by human supervisors. In fact, in our *Algo. Public* condition the prediction accuracy is 19.4% higher compared with the baseline.

Result 2. Disclosing the use and outcome of an algorithmic scoring process to borrowers led to self-fulfilling prophecies when the score was incorrect; that is, it caused borrowers to behave less according to their underlying tendencies and thus endogenously changed the label the score aimed to predict. We do not find self-fulfilling prophecies that reinforce borrowers' fundamental tendencies.

To elaborate on whether anchoring and moral wiggle room mechanisms underlie the reported findings,

we conducted additional analyses inspired by previous work on moderators. Regarding anchoring, previous work in psychology provides empirical evidence that people's susceptibility to anchoring depends on their Big-Five personality trait "openness-to-experience." Specifically, people with a high openness-to-experience score are significantly more likely to insufficiently adjust away from anchoring cues relative to people low in these traits because these traits lead to the activation of confirmatory search and selective accessibility mechanisms of anchoring (McElroy and Dowd 2007, Zong and Guo 2022). In line with these findings, belief adjustments in the direction of predictions in our *Algo. Private* condition are more pronounced for participants with an above-median openness-to-experience. For both types of participants, the coefficient for *Algo. Private* is larger if they possess a high compared with a low openness-to-experience score (−4.71 versus −3.82 for reciprocal types and 4.74 versus 1.42 for nonreciprocal types; see Tables 7 and 8 in the online appendix). We also find stronger *Algo. Private* treatment effects for participants who score relatively low on an adapted version of Rosenberg self-esteem scale by Brailovskaia and Margraf (2020) (−5.56 versus −1.53 for reciprocal types and 3.02 versus 1.16 for nonreciprocal types; see Tables 9 and 10 in the online appendix). This finding is in line with recent experiments by Zong and Guo (2022), who use the Rosenberg scale to proxy for participants' self-confidence and report that only individuals with a low score exhibited significant anchoring effects in a judgement task. In sum, the treatment effect we attribute to anchoring is stronger for participants with a high openness-to-experience score and a low self-esteem score, two factors that previous studies found to moderate people's susceptibility to anchoring.¹⁶

Regarding the moral wiggle room effect, psychologists have previously demonstrated a connection between self-esteem and strategies of self-presentation. Low compared with high self-esteem individuals are more inclined to avoid exhibiting bad qualities or disappointing others (Baumeister et al. 1989, Leary and Baumeister 2000). One implication of these findings is that constructing moral wiggle room to avoid psychological costs associated with violating others' expectations and behaving selfishly, that is, disappointing others, should be more attractive to people with relatively low self-esteem. Hence, in the context of our study, we would expect additional changes in reciprocal types' second-order beliefs in our *Algo. Public* compared with the *Algo. Private* condition to be primarily driven by low self-esteem participants. Indeed, we find that additional second-order belief adjustments of reciprocal types with relatively low scores on our Rosenberg self-esteem scale (Brailovskaia and Margraf 2020) are larger than the ones of their high self-esteem counterparts (−5.27 MU versus −2.68 MU; see Tables 11 and 12 in the online appendix). Hence, it

appears that low self-esteem participants are more inclined to leverage the information that an investor ignored a nonreciprocal prediction to form opportunistic beliefs about this person's repayment expectations, that is, construct moral wiggle room.

Taken together, additional findings from the subsample analyses provide some support for the notion that anchoring and moral wiggle room effects underlie observed second-order belief adjustments. However, reported findings are by no means conclusive evidence that these are the only mechanisms at work or that there are no other explanations for the results. Therefore, we discuss other potential mechanisms in the final section of this paper.

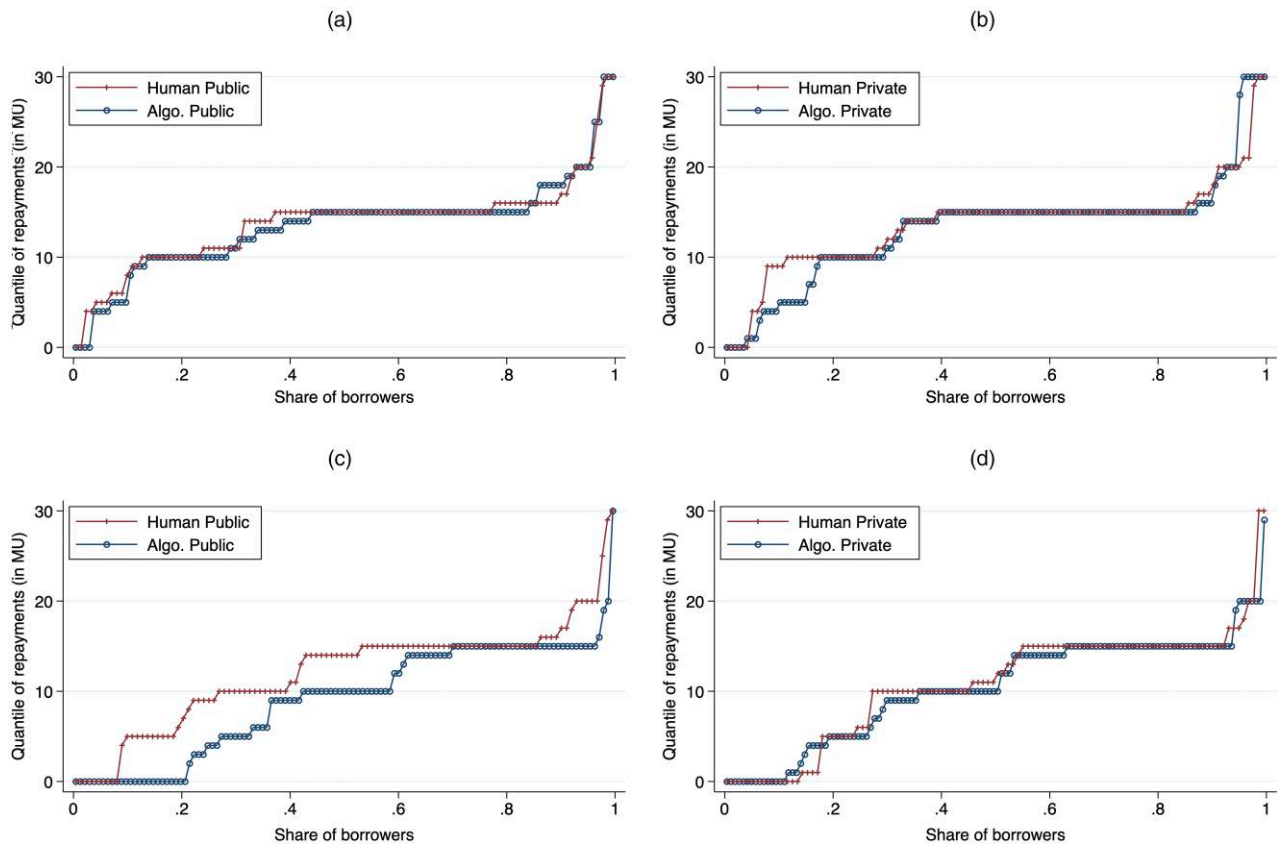
4.2. Algorithmic Factor

Results 1 and 2 depict considerable ramifications associated with making algorithmic scoring processes transparent to data subjects. However, at this point, the reader may wonder about the importance of the prediction's algorithmic origin: Do the reported effects exclusively occur when algorithmic systems produce the prediction or do our results reflect more general prediction effects that also occur when predictions come from human experts? To answer this question, we conclude our analyses by comparing the outcomes from our algorithmic scoring treatment conditions with those from conditions where human experts produce the scores.

Figure 5(a)–(d), contrasts the cumulative distributions of repayments when predictions come from the algorithmic model and human experts. We depict results conditional on the prediction and type of disclosure. We provide corresponding illustrations for second-order beliefs in Figure 8, (a)–(d), in the online appendix. Table 4 and Table 13 (online appendix) show summary statistics and regression analyses, respectively. Our results show that anchoring effects occurred independently of the prediction's origin. By contrast, the exploitation of moral wiggle room appears to be unique to the disclosure of nonreciprocal predictions by the algorithm.

Borrowers in the *Human Public* and *Human Private* conditions who learned that human experts predicted them to behave reciprocally, respectively, repaid 13.77 MU and 13.58 MU on average (Figure 5(a) and (b)); average second-order beliefs equaled 13.77 MU and 14.44 MU, respectively. Neither repayment decisions nor second-order beliefs differ from corresponding results in the *Algo. Public* and *Algo. Private* conditions ($p \geq 0.51$ for all, F test). We also find no significant differences in repayments and second-order beliefs between the *Human Private* and *Algo. Private* conditions for nonreciprocal predictions ($p \geq 0.98$ and $p \geq 0.16$, respectively, F test). The absence of treatment differences for these measures suggests that the anchoring effect is a more general consequence associated with the disclosure

Figure 5. (Color online) Repayment Distributions for Human Expert and Algorithmic Predictions



Notes. We show cumulative distributions of borrowers’ repayment decisions, contrasting repayments for the disclosure of human experts and algorithmic assessments. (a) Reciprocal prediction, public conditions. (b) Reciprocal prediction, private conditions. (c) Nonreciprocal prediction, public conditions. (d) Nonreciprocal prediction, private conditions.

of scoring results and not an idiosyncrasy of algorithmic scores.

A different result emerges when we look at the exploitation of moral wiggle room (Figure 5(c)). Borrowers in the *Human Public* condition who learned that the investor had observed a nonreciprocal prediction repaid 12.03 MU on average. This amount is significantly larger than borrowers’ corresponding repayments on the *Algo. Public* condition (+33%, $p < 0.01$, F test). Intriguingly, when comparing associated second-order beliefs, we do not find significant differences between algorithmic and

human expert predictions (respectively, 5.79 MU and 6.37 MU; $p = 0.37$, F test). Hence, disclosing that the investor had observed and overridden a nonreciprocal prediction led borrowers to construct seemingly opportunistic second-order beliefs independent of whether the prediction came from an algorithm or a human expert. However, only in the former case did borrowers exploit this opportunistic belief and self-servingly reduce their repayment compared with the baseline.

Why did borrowers not exploit the moral wiggle room in the *Human Public* condition? An explanation in

Table 4. Summary Statistics on Repayment Behavior and Second-Order Beliefs

| | Baseline | Prediction | <i>Algo. Public</i> | <i>Algo. Private</i> | <i>Human Public</i> | <i>Human Private</i> |
|-----------------------------|-----------------|---------------|---------------------|----------------------|---------------------|----------------------|
| Repayment (in MU) | | | | | | |
| | 12.15 (6.25) | Reciprocal | 13.45 (5.26) | 13.22 (6.11) | 13.77 (4.90) | 13.58 (5.09) |
| | | Nonreciprocal | 9.13 (6.22) | 10.45 (5.92) | 12.03 (5.90) | 10.74 (6.32) |
| Second-order belief (in MU) | | | | | | |
| | 12.56 (4.74) | Reciprocal | 13.76 (4.74) | 13.96 (5.30) | 13.77 (3.87) | 14.44 (4.55) |
| | | Nonreciprocal | 5.79 (5.72) | 9.14 (6.31) | 6.37 (5.17) | 10.27 (6.31) |

line with our data relates to borrowers' perceptions about the investor's intentions behind overriding a prediction. A large body of work in psychology and economics demonstrates that others' intentions matter when people interact with each other (Dufwenberg and Kirchsteiger 2004, Charness and Levine 2007). Perceptions of others' intentions have also been shown to influence people's inclination to exploit moral wiggle room (Friedrichsen et al. 2022). In line with previous observations from this literature (McCabe et al. 2003), it is possible that borrowers in the *Human Public* condition did not opportunistically exploit the constructed wiggle room because they perceived the decision to override the nonreciprocal prediction as genuine kindness which they felt obliged to reciprocate. Evidence in support of this notion originates from borrowers' beliefs about investors' perceptions of the reliability of the observed predictions.

When looking at borrowers' beliefs about investors' expectations of the prediction accuracy, we find that borrowers, on average, believed investors to be algorithm averse. Specifically, borrowers, on average, believed that investors perceived human expert predictions to be more accurate than algorithmic ones (respectively, 63.6% versus 73.1%; $p < 0.05$, F test), even though borrowers themselves did not expect an accuracy difference (62.8% versus 63.8%, $p = 0.97$, F test). Importantly, for nonreciprocal assessments in the *Human Public* and *Algo. Public* conditions, we find that borrowers' repayments increased with their beliefs about investors' accuracy expectations ($p < 0.05$, F test; see Table 15 in the online appendix). This result suggests that borrowers were less inclined to exploit moral wiggle room when they thought the investor had overridden a highly reliable prediction. Because overriding a highly accurate prediction is, in expected terms, more costly to an investor, borrowers might have seen overriding human expert predictions as a credible signal of investors' genuine kindness toward them. In line with previous research, this perception of genuine kindness could have prevented borrowers in the *Human Public* condition from behaving opportunistically (McCabe et al. 2003, Von Siemens 2013). Conversely, borrowers in the *Algo. Public* treatment had a lower belief about investors' accuracy expectations so they did not see overriding algorithmic outputs as a credible signal of genuine kindness they needed to reciprocate. Answers to additional survey questions on the perceived kindness and social pressure to adhere to algorithmic or human advice provide some support for this notion. Compared with the other treatments, borrowers in the *Human Public* condition most strongly agreed with statements that investing despite a nonreciprocal prediction signals genuine trust that they need to reward (see Table 16 in the online appendix).

In sum, our results suggest that the algorithmic origin of disclosed predictions matters for borrowers' willingness to exploit morale wiggle room created by an

investor's decision to override a prediction. That is because borrowers expect investors to be algorithm averse, leading them to perceive the overriding decision as an act of kindness.

Result 3. Human expert and algorithmic predictions similarly evoked anchoring effects. However, the algorithmic origin of a prediction mattered for the exploitation of constructed moral wiggle room. Borrowers seem to believe that overriding human expert compared with algorithmic predictions is a (more) genuine signal of kindness that limits the exploitation of moral wiggle room.

Borrowers' belief that investors expect expert predictions to be more reliable than algorithmic ones may reflect a so far overlooked facet of algorithm aversion: beliefs about other people's preference for human over algorithmic advice, or *second-order algorithm aversion*. This second-order algorithm aversion may cause borrowers' to expect that overriding a human expert prediction compared with an algorithmic one is more difficult for an investor. From that perspective, this facet of algorithm aversion may be the reason why borrowers perceive the overriding decision to be a (more) genuine signal of kindness. Interestingly, analyzing the decisions and beliefs of investors reveals that such second-order algorithm aversion would be mistaken. Our incentivized measures show that investors expected assessments from human experts (66%) and the machine learning model (66.9%) to be equally accurate ($p = 0.78$, Wilcoxon signed-rank test). Additionally, we do not find significant differences in investors' likelihood to adhere to an assessment conditional on its source ($p = 0.32$, χ^2 test).

5. Discussion and Conclusion

Understanding the ramifications of revealing algorithmic scoring processes to (involuntarily) scored individuals constitutes a major yet underexplored research problem. We conducted a comprehensive experimental study to systematically examine the impact of algorithmic transparency on behaviors of scored individuals in strategic settings. Our main contribution is to show that disclosing the involvement of incorrect algorithmic scores in decision-making processes to data subjects can steer their behaviors toward their score, creating a self-fulfilling prophecy. When analyzing the mechanisms driving our results, we find evidence that is consistent with our theoretical conjecture that the self-fulfilling prophecy occurred due to an anchoring effect and the exploitation of moral wiggle room that reshaped scored individuals' second-order beliefs. We further show that the occurrence of the self-fulfilling prophecy is a phenomenon unique to algorithmic scoring if the decision maker who used the score chose to override it. That appears to

be the case because scored individuals interpreted decision makers overriding human expert scores as a more genuine sign that they felt obliged to reciprocate. A potential explanation for the difference in the perceived kindness may relate to scored individuals' expectation that decision makers are algorithm averse: a pattern we refer to as *second-order algorithm aversion*.

5.1. Discussion of Results and Implications

Because of our nonspecialized experimental design, reported findings provide insights for a broad range of strategic decision-making scenarios where the algorithmically scored person is not the one to use the score to inform a decision. Instead, it is up to another person's discretion to use the score as a tool to make a decision that ultimately affects the scored person. Examples include hiring decisions where HR managers use algorithmic scores to assess an applicant's suitability for the job, insurance deals where brokers rely on scores to calculate premiums for potential customers, bail decisions where judges invoke algorithmic scores to determine the recidivism risk of defendants, and even payment term decisions where suppliers rely on algorithmic scores to assess their customers' payment behavior. To better understand the direct impact of revealing scores to assessed individuals, we employ a one-shot design, abstracting away from institutional or reputational factors that may constrain data subjects' ability to adjust behaviors in the direction of the score. Hence, our findings represent unadulterated behaviors, and underlying mechanisms, that disclosing scores to assessed individuals brings to light and whose occurrence institutions and other environmental influences refine. The magnitude of the behavioral effects we report in this study will depend on factors such as the threat of legal recourse or public shame so that our results may not generalize equally well to different strategic settings. However, as long as scored individuals have at least some discretion to adapt their predicted behavior, self-fulfilling prophecies can occur. For instance, a hired candidate who learns that the algorithm incorrectly predicts him to be a low performer may not entirely abstain from work due to binding contracts. Yet the candidate may, *ceteris paribus*, exhibit less effort compared with a counterfactual world where the algorithmic scoring process was absent. Similarly, a borrower who obtained a loan despite a low credit score may not completely default on the loan, but make late repayments if learning about the score and its use during the credit approval process. More generally, we expect our documented effects to be more likely to manifest the more discretion the decision maker has to override scores and the more discretion scored individuals have to change their predicted behavior without (serious) repercussions.

Relatedly, the extent to which our findings have important implications also depends on two interdependent

facets of the socio-technological environment where scoring processes are in use: the underlying distribution of assessed types and the error distribution of the algorithmic model. Our reported side effects are particularly noteworthy in settings where algorithmic models have difficulty making correct predictions. Typically, algorithmic models exhibit relatively low prediction performances in domains where there is not (yet) a sufficient amount of high-quality training data available. That is often the case for novel prediction problems (e.g., performance predictions for new fields of activities in a company) or when concept drifts (Žliobaitė et al. 2016) occur that change the underlying data generation process and adversely affect the prediction performance (e.g., risk premium predictions at the beginning of the COVID pandemic).

Notably, given the likely influence of the aforementioned contextual factors, our findings may not generalize equally to each of the examples we consider throughout this paper. Nonetheless, our results have several noteworthy implications whose magnitude of relevance will depend on the previously outlined factors. First, because the disclosure of the outcome and the use of algorithmic scores creates a self-fulfilling prophecy, algorithmic transparency can endogenously increase the predictive accuracy of the system and thus the decision maker's incentives to follow the available score. In our setting, for instance, the self-fulfilling prophecy inflated the machine-learning model's performance. Although the model in our baseline condition achieves an receiver operator characteristics (ROC)-area under the curve (AUC) of 0.567, its performance in the *Algo. Public* condition as measured by the ROC-AUC equals 0.603. As a result, the decision maker risks becoming a passive bystander who simply nods off the machine's implicit decision. Even people who are initially skeptical of the algorithm's predictive accuracy may gradually come to rely on it once they realize that their decisions to override the machine are actually suboptimal. From this point of view, the disclosure of results and the use of algorithmic evaluation procedures may inadvertently encourage blind faith in those results and undermine efforts to keep humans in power, that is, foster automation bias (Wickens et al. 2015).

Second, the self-fulfilling prophecy effect may be especially harmful in environments where algorithmic systems comprise discriminating machine-learning models and undergo repeated retraining based on novel training data they help to produce. If a machine-learning model produces biased outputs and inaccurately scored individuals are more likely to behave according to their score once it is revealed, the prediction becomes automatically more accurate. As a result, it becomes harder to detect whether the system is actually discriminatory as it endogenously affects the label it aims to predict. This endogeneity undermines the decision-makers' capabilities to act as a guard who

interferes if algorithms discriminate. In a dynamic setting where newly created observations are fed back to the system for retraining at a later point in time, the machine-learning model's bias might even increase. That is because the novel data encodes the endogenously reshaped feature-label relationship, which the machine-learning model will pick up on. Against this background, our results suggest that revealing algorithmic scoring processes to data subjects as an isolated transparency measure may foster the occurrence of negative feedback loops that prior work has warned about (Ensign et al. 2018, Cowgill and Tucker 2020).

Third, scored individuals' differential responses to being scored by an algorithm or human experts emphasize the complex, second-order channels through which algorithm aversion can affect human-machine interactions. The mere expectation that other people are averse to following algorithmically generated advice appears to affect the behavior of scored individuals in our study. This observation more broadly suggests that it matters how organizations advertise their employees' trust in and collaboration with machines, especially if the employment of algorithmic systems in customer-related processes becomes transparent. Our findings indicate that data subjects outside the organization may interpret the motivation why a decision maker overrides their algorithmic score and "gives them a chance" differently, conditional on their belief about his fundamental inclination to rely on the score. Against this background, organizations may be well-advised to combine transparency efforts with an advertisement or public-relations campaign that shows how well employees collaborate with machines (that make their work more manageable).

From a societal perspective, our results indicate that the indiscriminate and isolated implementation of algorithmic transparency measures may create unintended downstream ramifications. Imposing disclosure obligations that force organizations to provide data subjects access to the result of algorithmic scoring processes and inform them where scores affect people's lives might enable algorithms to create the frequently inaccurate world they forecast. This potential side effect raises the question of whether algorithmic transparency after the fact is a suitable means to balance power asymmetries between the scorers and the scored. The fundamental problem may be rooted more deeply in the equal endurance of good and bad information about their customers that organizations secretly use. When poor information feed into scoring algorithms, resulting predictions are generally unfair and may also be blatantly incorrect, evoking the outlined side effects. From this perspective, it might be beneficial if transparency and contestability measures precede the use of data by the algorithmic methods that transform it into scores. For example, an applicant for a position may be required to verify and correct provably inaccurate information that recruiters

have collected about the applicant, independent of the submitted documentation. Only after this correction phase can the company use algorithmic scoring methods to predict an applicant's fit for the job, reducing the likelihood of an inaccurate or unfair evaluation whose disclosure may evoke negative side effects. Notably, effectively including data subjects in the data cleaning process might also sense from a managerial perspective, because it could not only improve the performance of predictive models and prevent the occurrence of the side effects we document but also positively affect the data subjects' trust in the organization.

5.2. Limitations and Future Research

Our study does not come without limitations. In light of increasing regulatory requirements and private initiatives aiming to enhance algorithmic transparency, we believe these limitations open up fruitful avenues for future research.

First, although presented empirical insights are consistent with our theoretical framework, we acknowledge that they do not constitute conclusive evidence that anchoring and moral wiggle room effects are the (only) mechanisms driving changes in second-order beliefs. Several alternatives come to mind. Considering alternative explanations for the anchoring effect, it appears conceivable that privately disclosing an incorrect prediction affects scored individuals' belief about what social category of people they belong to and, thus, what behavioral norms they are obliged to follow (Krupka and Weber 2013). If such a change in self-categorization had occurred, the more reliable the received signal, the more likely borrowers would have changed their beliefs about what kind of person they are (Bénabou and Tirole 2011). Casting doubt on this rationale, additional regression analyses reveal that the impact of emotional and cognitive trust in the prediction (Komiak and Benbasat 2006), that is, the perceived reliability of the signal, on second-order beliefs is statistically and economically insignificant (see Table 18 and 19 in the online appendix). Another way to rationalize anchoring effects is that scored individuals perceive the algorithm and human experts as authority figures they need to obey because they exert "expert power" (French et al. 1959, Milgram and Gudehus 1978). As expert power typically increases with beliefs about the authority figure's expertise in a relevant domain, we would expect that second-order belief adjustments increase with scored individuals' perception about the prediction accuracy, that is, the authority figure's expertise. However, results from additional analyses are inconsistent with this alternative explanation (see Tables 20 and 21 in the online appendix). A third alternative rationale for effects associated with the private disclosure of predictions we can speculate about is some general preference to prove predictions right. Although we believe that our

experimental design and the incentive compatibility render such a mechanism unlikely, we cannot definitively rule it out, leaving its examination to future research. Regarding the moral wiggle room effect, we see two plausible alternative (or additional) mechanisms. On the one hand, the second-order belief adjustments in the *Algo. Public* condition we interpret to be opportunistic could reflect genuine changes in beliefs about investor motives. If the borrower wants to maximize the sum of the borrower's and investor's utility and takes the investor's decision to override the nonreciprocal prediction as a sign of indifference to the material income, it is rational to reduce repayments. That is because the investor values the material income relatively less. Although we are unable to speak to participants' true motives, our finding that repayments in the *Algo. Public* condition increase with beliefs about investors' accuracy expectations are inconsistent with this explanation. A borrower believing an investor to have high accuracy expectations should see overriding a prediction as a credible signal that the investor is not interested in material income. However, our findings suggest the opposite: repayments increased with borrowers' beliefs about investors' accuracy expectations. On the other hand, it seems possible that borrowers comprehend investors as the product of a hybrid human-machine system rather than just a human aided by an algorithm. In this case, if a borrower is wrongly classified as someone who is not making repayments, they may retaliate against the whole hybrid system. In line with previous insights from the algorithm aversion literature (Dietvorst et al. 2015), the absence of a wiggle room effect for human expert predictions may then reflect a higher willingness to forgive incorrect predictions of a pure human system that makes investment decisions. We cannot rule out that such an effect additionally reduces repayments of reciprocal participants in our *Algo. Public* condition. However, this mechanism does not provide a clear rationale for the observed adjustments of second-order beliefs. Hence, it seems more plausible that such an effect adds to the moral wiggle room mechanism instead of being an orthogonal alternative. Examining the existence of such an effect more generally appears particularly fruitful considering upcoming regulations requiring humans-in-the-loop (Parliament and Council of European Union 2021).

Related to alternative mechanisms driving changes in second-order beliefs, it is also conceivable that the repayment differences in the *Human Public* and *Algo. Public* conditions point to a more deeply rooted phenomenon. The higher repayment for a nonreciprocal, that is, negative, prediction in the human condition may also be related to a more pronounced desire to prove the fundamentally inaccurate predictions of human experts wrong. This stronger desire to defy a prediction may also trace back to borrowers' belief that investors, on average, expect human expert predictions to be more

accurate than the algorithmic ones, that is, their second-order algorithm aversion. It is possible that a sense of accomplishment associated with defying somebody else' (negative) assessments (Deci and Ryan 2013) depends on how reliable investors perceive the assessment to be. This train of thought naturally raises the question of whether the repayment difference would be reversed if borrowers believed investors to exhibit algorithm appreciation. Alternatively, our finding may partially depict a fundamental human need to "prove human haters wrong" but not algorithmic ones. Testing and isolating these different motives and studying the role of second-order algorithm aversion (or appreciation) in more depth constitute a fruitful avenue for future research. Overall, our discussion on the mechanisms behind our documented effects aims to inspire future research in this area, as understanding these mechanisms is crucial for anticipating when and how they may occur.

Second, scored individuals in our study do not obtain explanations about how the employed machine-learning model leverages their personal information to arrive at a prediction. While focusing mainly on system users, not data subjects, previous research on explainable AI (XAI) provides ample evidence that explainability can influence human attitudes toward and perceptions of algorithmic systems (Gregor and Benbasat 1999, Poursabzi-Sangdeh et al. 2021). Disclosing not only their score and its use to data subjects but also informing them about why the algorithm produces a given score may shift perceptions about the system's reliability and thus dynamically interact with second-order beliefs. Analyzing how our results may change when scores come from an explainable algorithm is a fruitful avenue for future studies.

Third, although the strategy method provides unique insights into responses to counterfactual scores, it prevents us from confronting data subjects with one inaccurate score they can choose to contest (at a personal cost). Examining such a setting in more depth is interesting from at least two points of view. On the one hand, it is vital to understand whether and under what circumstances data subjects are willing to challenge inaccurate scores even if the outcome benefits them in the short run. On the other hand, it introduces additional space for wiggle room and opportunistic behavior on the part of the scored individuals; they may strategically refrain from challenging an inaccurate score if it provides wiggle room.

Fourth, we use a one-shot design that does not allow us to study the endurance of reported effects or identify potential spillover effects. Do anchoring effects or constructed opportunistic beliefs persist over time or do scored individuals forget the scoring process relatively quickly? Also, do changed second-order beliefs only apply for the specific context where the scoring process occurs, or do scored individuals transfer these adjusted beliefs to other, related domains?

Fifth, as mentioned before, we abstract away from environmental influences that may constrain the occurrence of (behavioral) self-fulfilling prophecies. One evident avenue for future research is to analyze moderating factors and document real-world settings where our reported effects are most likely to occur. For example, it could be the case that in a different domain with distinct socio-technological influencing factors, the disclosure of accurate predictions also evokes self-fulfilling prophecies that reinforce scored individuals' fundamental types.

Finally, we encourage future research on the pattern we believe may depict second-order algorithm aversion. Our aggregate level evidence on borrowers' expectation that investors believe human expert predictions to be more accurate than algorithmic ones opens a wide range of follow-up questions. Does second-order algorithm aversion (or appreciation) occur in other settings? Does it have explanatory power regarding peoples' response to hybrid human-machine advice? What are the determinants of second-order algorithm aversion (or appreciation)? Do people opportunistically form motivated beliefs that others exhibit algorithm aversion or appreciation or do they project their own perceptions? Answering these questions seems particularly urgent considering that it becomes increasingly likely that people interact with (hybrid) algorithmic systems where other humans are in the loop, for example, due to regulatory requirements, so that second-order algorithm aversion (or appreciation) could moderate economic and social outcomes.

5.3. Conclusion

A concluding remark is worth making. Of course, our work is not meant to be an argument, much less a plea, against making "black box" scoring processes more transparent. Instead, we comprehend our findings as a warning that sunlight all on its own may not be "the best disinfectant". Staying with Louis Brandeis's metaphor, we argue that sunlight without additional protection brings its own dangers, such as sunburn, that need addressing. A high-level interpretation of our results is that the pervasive human inclination to process information in a biased and often self-serving manner may interact with well-intended yet rudimentary transparency efforts in unexpected ways, yielding undesirable outcomes. To design successful algorithmic transparency measures that level power asymmetries, regulators need to address these facets of the human mind by implementing complementary governance and incentive structures. In our setting, data subjects may, for instance, obtain additional information about the score being a probabilistic measure (to mitigate anchoring effects) or explain why a decision maker chose to override a score (to reduce informational uncertainty and thus the construction of opportunistic beliefs).

Acknowledgments

The authors thank seminar participants at Goethe University, the Johannes Gutenberg-University Mainz, University of Mannheim, University of Michigan, the Leibniz Institute for Financial Research Sustainable Architecture for Finance in Europe, and the VHB annual meeting for helpful comments and suggestions.

Endnotes

¹ In the following, we refer to individuals whose personal data are collected, held or processed to generate a predictive score interchangeably as data subjects, scored individuals, or assessed individuals. Similarly, we use the terms scores, assessments, and predictions interchangeably.

² See AEARCTR-0009300 and AEARCTR-0009694.

³ From a game theoretic point of view, when it is public knowledge that individuals are only motivated by their personal material income, there exists a unique subgame perfect Nash equilibrium: Investors correctly anticipate that borrowers do not make a repayment and thus do not invest in the first place, that is, not investing is the dominant strategy. The uniqueness, however, does not hold anymore, when allowing for the existence of social preferences, concerns about other people's material well-being, which seems reasonable in the light of robust empirical research findings (Miettinen et al. 2020). Under perfect information, there exist multiple subgame perfect Nash equilibria where investors always invest when borrowers possess sufficiently strong social preferences and return strictly more than 10 MU; under imperfect information, there also exist multiple perfect Bayesian equilibria.

⁴ The matching procedure has no implications for the robustness of our results; it merely served as a means to ensure incentive compatibility so that the experiment does not represent a mock-up thought experiment but a decision-making scenario that determines participants' personal and other individuals' material well-being. We outline the matching procedure in more detail later.

⁵ To illustrate the unique facet of the strategy method, suppose we randomly matched participant B with investor A who invested even though A observed the prediction that B would not repay more than 10 MU. If we had not used the strategy method but instead asked every participant to decide after they learned their personal prediction, it would have been impossible to observe B's decision following the prediction that B would repay more than 10 MU, that is, the prediction that the algorithm did not actually make.

⁶ We iteratively computed SHAP values of the model and reduced the number of features until we ended up with a set of predictive features that is not overly large so that participants do not get tired filling out the initial survey in stage 1. The superset of features comprised all survey items included in the previous field study (see online appendix). Eventually, we selected 12 characteristics that are sufficiently strong predictors that are easy and fast to elicit in the questionnaire.

⁷ Because we (i) elicited participants' incentivized beliefs about the predictive performance without informing them about the model's true performance and (ii) used the strategy method, the actual model performance can by design not influence our results.

⁸ Pilot participants could not participate in the main experiment.

⁹ The matching procedure worked as follows: At the end of the study, we generated a random integer between 1 and 25 for every borrower. We used the random integer to match borrowers to a unique ID of our 25 investors from the pilot study. Given a borrower's treatment condition and the actual prediction, we selected one of the five investor decisions to determine whether the investor initially invested under the given circumstances. If the investor

invested, we implemented the borrower's decision for the actual prediction. For example, assume a borrower participated in the *Algo. Public* condition and the developed machine-learning model predicts the borrower to repay more than 10 MU. This borrower indicated repaying 15 MU if an investor observes the prediction that the borrower would repay more than 10 MU and invests. Based on the random integer, we matched this borrower with investor number 5, who decided to invest when the machine-learning model predicts a borrower to repay more than 10 MU. Accordingly, in this scenario, we would implement the following decisions to calculate payoffs: the investor initially invests, and the borrower repays 15 MU, leaving both with a 15 MU payoff.

¹⁰ To ensure the quality of our data, we exclude participants who failed our introduced attention check, which is 141 of 707 participants (approximately 20%).

¹¹ Among all borrowers matched with a specific investor, we randomly select one borrower whose choice becomes payoff-relevant for the investor.

¹² We find similar pattern in an additional within-subject design where each participant makes decisions for the baseline, the *Algo. Public*, and the *Algo. Private* conditions, allowing us to include individual fixed effects (see Online Appendix C).

¹³ Additional analyses on treatment heterogeneities provide more direct empirical evidence that the observed changes in second-order beliefs relate to anchoring and morale wiggle room effects. We report these findings and discuss our conjectured mechanism together with alternative mechanisms later.

¹⁴ See the online appendix for more details.

¹⁵ The fact that nonreciprocal types did not further lower their repayment may also originate from a perception of a generally required minimum repayment that is independent of their belief about what the current investor expects back.

¹⁶ Corresponding subsample analyses regarding *Human Private* treatment effects reveal similar, however, less pronounced treatment heterogeneities (see Tables 22–25 in the online appendix).

References

- Adomavicius G, Bockstedt JC, Curley SP, Zhang J (2013) Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Inform. Systems Res.* 24(4):956–975.
- Adomavicius G, Bockstedt JC, Curley SP, Zhang J (2018) Effects of online recommendations on consumers' willingness to pay. *Inform. Systems Res.* 29(1):84–102.
- Adomavicius G, Bockstedt J, Curley S, Zhang J (2019) Reducing recommender systems biases: An investigation of rating display designs. *Management Inform. Systems Quart.* 43(4):19–18.
- Ajzen I (1989) Attitude structure and behavior. *Attitude Structure Function* 241:274.
- Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. *Ethics of Data and Analytics* (Auerbach Publications, Boca Raton, FL), 254–264.
- Araujo T, De Vreese C, Helberger N, Kruijkemeier S, van Weert J, Bol N, et al. (2018) Automated decision-making fairness in an ai-driven world: Public perceptions, hopes and concerns. *Digital Communication Methods Laboratory*. http://www.digicomlab.eu/reports/2018_adm_by_ai/.
- Ariely D, Loewenstein G, Prelec D (2003) "coherent arbitrariness": Stable demand curves without stable preferences. *Quart. J. Econom.* 118(1):73–106.
- Balcetis E, Dunning D (2006) See what you want to see: Motivational influences on visual perception. *J. Personality Soc. Psych.* 91(4):612.
- Ban G-Y, El Karoui N, Lim AE (2018) Machine learning and portfolio optimization. *Management Sci.* 64(3):1136–1154.
- Battigalli P, Dufwenberg M (2007) Guilt in games. *Amer. Econom. Rev.* 97(2):170–176.
- Battigalli P, Dufwenberg M (2009) Dynamic psychological games. *J. Econom. Theory* 144(1):1–35.
- Baumeister RF, Stillwell AM, Heatherton TF (1994) Guilt: An interpersonal approach. *Psych. Bull.* 115(2):243.
- Baumeister RF, Tice DM, Hutton DG (1989) Self-presentational motivations and personality differences in self-esteem. *J. Personality* 57(3):547–579.
- BCC (2021) Predictive analytics: Global markets. Accessed July 30, 2022, <https://www.bccresearch.com/market-research/information-technology/predictive-analytics-market.html>.
- Bénabou R, Tirole J (2011) Identity, morals, and taboos: Beliefs as assets. *Quart. J. Econom.* 126(2):805–855.
- Benlian A, Titah R, Hess T (2012) Differential effects of provider recommendations and consumer reviews in e-commerce transactions: An experimental study. *J. Management Inform. Systems* 29(1):237–272.
- Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games Econom. Behav.* 10(1):122–142.
- Blankenship KL, Wegener DT, Petty RE, Detweiler-Bedell B, Macy CL (2008) Elaboration and consequences of anchored estimates: An attitudinal perspective on numerical anchoring. *J. Experiment. Soc. Psych.* 44(6):1465–1476.
- Bock G-W, Kankanhalli A, Sharma S (2006) Are norms enough? The role of collaborative norms in promoting organizational knowledge seeking. *Eur. J. Inform. Systems* 15(4):357–367.
- Bock G-W, Zmud RW, Kim Y-G, Lee J-N (2005) Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. *Management Inform. Systems Quart.* 29(1):87–111.
- Bolton GE, Kusterer DJ, Mans J (2019) Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Sci.* 65(11):5371–5391.
- Brailovskaia J, Margraf J (2020) How to measure self-esteem with one item? validation of the german single-item self-esteem scale (g-sise). *Current Psych.* 39(6):2192–2202.
- Brandts J, Charness G (2000) Hot vs. cold: Sequential responses and preference stability in experimental games. *Experiment. Econom.* 2(3):227–238.
- Cabral TS (2021) AI and the right to explanation: Three legal bases under the gdpr. *Data Protection Artificial Intelligence* 13:29.
- Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *J. Risk Uncertainty* 19(1):7–42.
- Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *J. Marketing Res.* 56(5):809–825.
- Cervone D, Peake PK (1986) Anchoring, efficacy, and action: The influence of judgmental heuristics on self-efficacy judgments and behavior. *J. Personality Soc. Psych.* 50(3):492.
- Chan L, Doyle K, McElfresh D, Conitzer V, Dickerson JP, Schaich Borg J, Sinnott-Armstrong W (2020). Artificial intelligence: Measuring influence of ai assessments on moral decision-making. *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society* (AAAI Press/ACM, New York), 214–220.
- Chapman GB, Johnson EJ (1999) Anchoring, activation, and the construction of values. *Organ. Behav. Human Decision Processes* 79(2):115–153.
- Chapman GB, Johnson EJ (2002) Incorporating the irrelevant: Anchors in judgments of belief and value. *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, Cambridge, UK), 120–138.
- Charness G, Levine DI (2007) Intention and stochastic outcomes: An experimental study. *Econom. J. (London)* 117(522):1051–1072.
- Chen DL, Schonger M, Wickens C (2016) otree—An open-source platform for laboratory, online, and field experiments. *J. Behav. Experiment. Finance* 9:88–97.

- Cowgill B, Tucker CE (2020). Algorithmic fairness and economics. Research paper, Columbia Business School, New York.
- Dana J, Weber RA, Kuang JX (2007) Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Econom. Theory* 33(1):67–80.
- Dastin J (2018) Amazon scraps secret ai recruiting tool that showed bias against women. *Ethics of Data and Analytics* (Auerbach Publications, Boca Raton, FL), 296–299.
- Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Inform. Systems Quart.* 13(3):319–340.
- Dawes RM (1979) The robust beauty of improper linear models in decision making. *Amer. Psych.* 34(7):571.
- Deci EL, Ryan RM (2013) *Intrinsic Motivation and Self-Determination in Human Behavior* (Springer Science and Business Media, Boston).
- Dencik L, Redden J, Hintz A, Warne H (2019) The ‘golden view’: Data-driven governance in the scoring society. *Internet Policy Rev.* 8(2):1–24.
- Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych. General* 144(1):114.
- Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3):1155–1170.
- Dijkstra JJ (1999) User agreement with incorrect expert system advice. *Behav. Inform. Tech.* 18(6):399–411.
- Dodge J, Liao QV, Zhang Y, Bellamy RK, Dugan C (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proc. 24th Internat. Conf. on Intelligent User Interfaces* (ACM, New York), 275–285.
- Dufwenberg M, Kirchsteiger G (2004) A theory of sequential reciprocity. *Games Econom. Behav.* 47(2):268–298.
- Dunning D (1999) A newer look: Motivated social cognition and the schematic representation of social concepts. *Psych. Inquiry* 10(1): 1–11.
- Englich B, Mussweiler T, Strack F (2006) Playing dice with criminal sentences: The influence of irrelevant anchors on experts’ judicial decision making. *Personality Soc. Psych. Bull.* 32(2):188–200.
- Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2018) Runaway feedback loops in predictive policing. *Proc. Conf. on Fairness, Accountability and Transparency*, 160–171.
- Epley N, Gilovich T (2016) The mechanics of motivated reasoning. *J. Econom. Perspective* 30(3):133–140.
- Exley CL (2016) Excusing selfishness in charitable giving: The role of risk. *Rev. Econom. Stud.* 83(2):587–628.
- Falk A, Becker A, Dohmen TJ, Huffman D, Sunde U (2016) The preference survey module: A validated instrument for measuring risk, time, and social preferences. *IZA Discussion Paper, Institute for the Study of Labor, Bonn, Germany*.
- Falk A, Becker A, Dohmen T, Enke B, Huffman D, Sunde U (2018) Global evidence on economic preferences. *Quart. J. Econom.* 133(4): 1645–1692.
- Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425(6960):785–791.
- Fischbacher U, Gächter S, Quercia S (2012) The behavioral validity of the strategy method in public good experiments. *J. Econom. Psych.* 33(4):897–913.
- French JR, Raven B, Cartwright D (1959) The bases of social power. *Classics of Organization Theory*, 311–320.
- Friedrichsen J, Momen K, Piasenti S (2022) Ignorance, intention and stochastic outcomes. *J. Behav. Experiment. Econom.* 100:101913.
- Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Inform. System Res.* 33(2):678–696.
- Ge R, Zheng Z, Tian X, Liao L (2021) Human–robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Inform. Systems Res.* 32(3):774–785.
- Gimpel H, Heger S, Olenberger C, Utz L (2021) The effectiveness of social norms in fighting fake news on social media. *J. Management Inform. Systems* 38(1):196–221.
- Goldstein NJ, Cialdini RB, Griskevicius V (2008) A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *J. Consumer Res.* 35(3):472–482.
- Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a “right to explanation.” *AI Magazine* 38(3):50–57.
- Google AI (2019) Responsible AI practices: Interpretability. Accessed March 8, 2022, <https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability>.
- Gregor S, Benbasat I (1999) Explanations from intelligent systems: Theoretical foundations and implications for practice. *Management Inform. Systems Quart.* 23(4):497–530.
- Grether DM (1980) Bayes rule as a descriptive model: The representativeness heuristic. *Quart. J. Econom.* 95(3):537–557.
- Grossman Z (2014) Strategic ignorance and the robustness of social preferences. *Management Sci.* 60(11):2659–2665.
- Grossman Z, Van Der Weele JJ (2017) Self-image and willful ignorance in social decisions. *J. Eur. Econom. Assoc.* 15(1):173–217.
- Gupta A, Kannan K, Sanyal P (2018) Economic experiments in information systems. *Management Inform. Systems Quart.* 42(2):595–606.
- Haisley EC, Weber RA (2010) Self-serving interpretations of ambiguity in other-regarding behavior. *Games Econom. Behav.* 68(2): 614–625.
- Häubl G, Murray KB (2003) Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents. *J. Consumer Psych.* 13(1–2):75–91.
- Holt CA, Smith AM (2009) An update on Bayesian updating. *J. Econ. Behav. Organ.* 69(2):125–134.
- Huang Y, Meng S (2019) Automobile insurance classification rate-making based on telematics driving data. *Decision Support Systems* 127:113156.
- Jacowitz KE, Kahneman D (1995) Measures of anchoring in estimation tasks. *Personality Soc. Psych. Bull.* 21(11):1161–1166.
- Jago AS (2019) Algorithms and authenticity. *Acad. Management Discovery* 5(1):38–56.
- Jiang Z, Benbasat I (2007) The effects of presentation formats and task complexity on online consumers’ product understanding. *Management Inform. Systems Quart.* 31(3):475–500.
- Johnson ND, Mislin AA (2011) Trust games: A meta-analysis. *J. Econom. Psych.* 32(5):865–889.
- Jussupow E, Benbasat I, Heinzl A (2020). Why are we averse toward algorithms? A comprehensive literature review on algorithm aversion. *Proc. 28th Eur. Conf. on Inform. Systems* (Association for Information Systems).
- Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence. *Inform. Systems Res.* 32(3):713–735.
- Kim D, Benbasat I (2006) The effects of trust-assuring arguments on consumer trust in Internet stores: Application of Toulmin’s model of argumentation. *Inform. Systems Res.* 17(3):286–300.
- Kizilcec RF (2016) How much information? Effects of transparency on trust in an algorithmic interface. *Proc. CHI Conf. on Human Factors in Computing Systems*, 2390–2395.
- Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Quart. J. Econom.* 133(1):237–293.
- Komiak SY, Benbasat I (2006) The effects of personalization and familiarity on trust and adoption of recommendation agents. *Management Inform. Systems Quart.* 30(4):941–960.
- Krupka EL, Weber RA (2013) Identifying social norms using coordination games: Why does dictator game sharing vary? *J. Eur. Econom. Assoc.* 11(3):495–524.
- Kunda Z (1990) The case for motivated reasoning. *Psych. Bull.* 108(3):480.

- Larson T, Capra CM (2009) Exploiting moral wiggle room: Illusory preference for fairness? a comment. *Judgment Decision Making* 4(6):467.
- Leary MR, Baumeister RF (2000) The nature and function of self-esteem: Sociometer theory. *Advances in Experimental Social Psychology* (Elsevier, New York), 1–62.
- Li X, Grahl J, Hinz O (2022) How do recommender systems lead to consumer purchases? A causal mediation analysis of a field experiment. *Inform. Systems Res.* 33(2):620–637.
- Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Human Decision Processes* 151:90–103.
- Mackay JM, Elam JJ (1992) A comparative study of how experts and novices use a decision aid to solve problems in complex knowledge domains. *Inform. Systems Res.* 3(2):150–172.
- Mathieson K (1991) Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. *Inform. Systems Res.* 2(3):173–191.
- McCabe KA, Rigdon ML, Smith VL (2003) Positive reciprocity and intentions in trust games. *J. Econom. Behav. Organ.* 52(2):267–275.
- McElroy T, Dowd K (2007) Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment Decision Making* 2(1):48.
- McKnight DH, Choudhury V, Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. *Inform. Systems Res.* 13(3):334–359.
- Meta AI (2021) Facebook's five pillars of responsible AI. Accessed March 8, 2022, <https://ai.facebook.com/blog/facebook-five-pillars-of-responsible-ai/>.
- Miettinen T, Kosfeld M, Fehr E, Weibull J (2020) Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *J. Econom. Behav. Organ.* 173:1–25.
- Milgram S, Gudehus C (1978) *Obedience to Authority* (Ziff-Davis Publishing Company, New York).
- Momsen K, Ohndorf M (2022) Information avoidance, selective exposure, and fake (?) news: Theory and experimental evidence on green consumption. *J. Econom. Psych.* 88:102457.
- Mussweiler T, Strack F (1999) Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *J. Experiment. Soc. Psych.* 35(2):136–164.
- Northcraft GB, Neale MA (1987) Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organ. Behav. Human Decision Processes* 39(1):84–97.
- OECD AI (2021) Database of national ai policies. Accessed July 7, 20, 2022, <https://oecd.ai>.
- Palmeira M, Spassova G (2015) Consumer reactions to professionals who use decision aids. *Eur. J. Marketing.* 49(3/4):302–326.
- Parliament and Council of European Union (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. Accessed July 30, 2022, <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
- Parliament and Council of European Union (2021) Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Accessed July 30, 2022, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>.
- Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H (2021) Manipulating and measuring model interpretability. *Proc. CHI Conf. on Human Factors in Computing Systems* (ACM, New York), 1–52.
- Prahl A, Van Swol L (2017) Understanding algorithm aversion: When is advice from automation discounted? *J. Forecasting* 36(6):691–702.
- Rader E, Cotter K, Cho J (2018) Explanations as mechanisms for supporting algorithmic transparency. *Proc. CHI Conf. on Human Factors in Computing Systems*.
- Regner T (2018) Reciprocity under moral wiggle room: Is it a preference or a constraint? *Experiment. Econom.* 21(4):779–792.
- Sanders GL, Courtney JF (1985) A field study of organizational factors influencing DSS success. *Management Inform Systems Quart.* 9(1):77–93.
- Schanke S, Burtch G, Ray G (2021) Estimating the impact of “humanizing” customer service chatbots. *Inform. Systems Res.* 32(3):736–751.
- Scherer LD, de Vries M, Zikmund-Fisher BJ, Wittman HO, Fagerlin A (2015) Trust in deliberation: The consequences of deliberative decision strategies for medical decisions. *Health Psych.* 34(11):1090.
- Senoner J, Netland T, Feuerriegel S (2021) Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Sci.* 68(8):5704–5723.
- Serrano-Cinca C, Gutiérrez-Nieto B (2016) The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending. *Decision Support Systems* 89:113–122.
- Springer A, Whittaker S (2018) What are you hiding? Algorithmic transparency and user perceptions. *Proc. AAAI Spring Sympos.* (AAAI Press, Palo Alto, CA).
- Tambe P, Cappelli P, Yakubovich V (2019) Artificial intelligence in human resources management: Challenges and a path forward. *California Management Rev.* 61(4):15–42.
- Toussaert S (2017) Intention-based reciprocity and signaling of intentions. *J. Econom. Behav. Organ.* 137:132–144.
- Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.
- Uhlmann EL, Pizarro DA, Tannenbaum D, Ditto PH (2009) The motivated use of moral principles. *Judgment Decision Making* 4(6): 479–491.
- van den Broek E, Sergeeva A, Huysman M (2021) When the machine meets the expert: An ethnography of developing ai for hiring. *Management Inform. Systems Quart.* 45(3):1557–1580.
- Vandenbosch B, Higgins C (1996) Information acquisition and mental models: An investigation into the relationship between behaviour and learning. *Inform. Systems Res.* 7(2):198–214.
- Von Siemens FA (2013) Intention-based reciprocity and the hidden costs of control. *J. Econom. Behav. Organ.* 92:55–65.
- Wang W, Benbasat I (2007) Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *J. Management Inform. Systems* 23(4):217–246.
- Warshaw J, Matthews T, Whittaker S, Kau C, Bengualid M, Smith BA (2015). Can an algorithm know the “real you”? Understanding people's reactions to hyper-personal analytics systems. *Proc. 33rd Annual ACM Conf. on Human Factors in Computing Systems* (ACM, New York), 797–806.
- Watson HJ, Nations C (2019) Addressing the growing need for algorithmic transparency. *Comm. Assoc. Inform. Systems* 45(1):26.
- Wickens CD, Clegg BA, Vieane AZ, Sebok AL (2015) Complacency and automation bias in the use of imperfect automation. *Human Factors* 57(5):728–739.
- Xiao B, Benbasat I (2015) Designing warning messages for detecting biased online product recommendations: An empirical investigation. *Inform. Systems Res.* 26(4):793–811.
- Xu J, Benbasat I, Cenfetelli RT (2014) Research note—The influences of online service technologies and task complexity on efficiency and personalization. *Inform. Systems Res.* 25(2):420–436.
- Yasseri T, Reher J (2022) Fooled by facts: Quantifying anchoring bias through a large-scale experiment. *J. Comput. Soc. Sci.* 5(1):1001–1021.
- You S, Yang CL, Li X (2022) Algorithmic vs. human advice: Does presenting prediction performance matter for algorithm appreciation? *J. Management Inform. Systems* 39(2):336–365.
- Žliobaitė I, Pechenizkiy M, Gama J (2016) An overview of concept drift applications. Japkowicz N, Stefanowski J, eds. *Big Data Analysis: New Algorithms for a New Society* vol. 16 (Springer, Cham, Switzerland), 91–114.
- Zong Y, Guo X (2022) An experimental study on anchoring effect of consumers' price judgment based on consumers' experiencing scenes. *Frontiers Psych.* 13:794135.