Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining

Erlei, Alexander ; Nekdem, Franck Awounang; Meub, Lukas; Anand, Avishek ; Gadiraju, Ujwal

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing

Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining

Alexander Erlei,[1] **Franck Awounang Nekdem,**[2] **Lukas Meub,**[1] **Avishek Anand,**[2] **Ujwal Gadiraju**[3]
[1]Georg-August-Universität Göttingen, [2]Leibniz Universität Hannover, [3]Delft University of Technology
firstname.lastname@wiwi.uni-goettingen.de,[1] franck.awounangnekdem@gmail.com, anand@L3S.de,
u.k.gadiraju@tudelft.nl

Abstract

Recent advances in machine learning have led to the widespread adoption of ML models for decision support systems. However, little is known about how the introduction of such systems affects the behavior of human stakeholders. This pertains both to the people using the system, as well as those who are affected by its decisions. To address this knowledge gap, we present a series of ultimatum bargaining game experiments comprising 1178 participants. We find that users are willing to use a black-box decision support system and thereby make better decisions. This translates into higher levels of cooperation and better market outcomes. However, because users under-weigh algorithmic advice, market outcomes remain far from optimal. Explanations increase the number of unique system inquiries, but users appear less willing to follow the system's recommendation. People who negotiate with a user who has a decision support system, but cannot use one themselves, react to its introduction by demanding a better deal for themselves, thereby decreasing overall cooperation levels. This effect is largely driven by the percentage of participants who perceive the system's availability as unfair. Interpretability mitigates perceptions of unfairness. Our findings highlight the potential for decision support systems to further human cooperation, but also the need for regulators to consider heterogeneous stakeholder reactions. In particular, higher levels of transparency might inadvertently hurt cooperation through changes in fairness perceptions.

1 Introduction

The ever increasing ability of economies and societies to effectively make use of machine learning techniques has lead to a surge in algorithmic decision-making (ADM) systems. Nowadays, algorithmic systems are being implemented in many high-stakes domains such as medical diagnoses (Obermeyer and Emanuel 2016), judicial sentencing (Angwin et al. 2016), urban planning (Glaeser et al. 2016) or hiring decisions (Liem et al. 2018). Compared to human actors, ADM systems provide enhanced analytic capabilities, increased efficiency and allow for comprehensive data monitoring. As decision-making environments become more complex and dependent on extensive data analytics, human decision-makers will be required to effectively utilize these computational tools and reinterpret their role within crucial decision-making processes. Thus, the integration of ADM systems is accompanied by a number of challenges within the human-agent sphere, calling for new and clearly identified design choices that facilitate human-agent interactions. These challenges are manifold in nature, since one not only needs to identify the limits and ambiguities of current ADM systems, but further combine those with a sophisticated understanding of human behavior, social desirabilities and societal expectations.

This paper exploits the widely studied ultimatum game to examine how the introduction of an algorithmic decision-support-system (DSS) affects the behavior of those who are using the system and those who are affected by the system. The game is commonly used to represent a range of bargaining situations such as purchasing decisions, price setting or firm negotiations. Subjects are assigned one of two roles: the proposer and the responder. The proposer receives a sum of money and subsequently makes an offer on how to divide that money between both players. Responders can either accept or reject that offer. If the responder rejects, both receive nothing, if the responder accepts, the money is split according to the proposer's offer. Within this framework, we ask three main questions:

- **RQ1:** Do human decision-makers integrate advice from a black-box DSS?
- **RQ2:** Do human decision-makers exhibit different social concerns while interacting with another human using an ADM system?
- **RQ3:** Does interpretability mediate system trust and fairness perceptions?

We introduce eight treatments relating to and building on the standard one-shot ultimatum game. Throughout, we explore the impact of (i) a black-box DSS and (ii) a DSS that is accompanied by an explanation, on user behavior and overall economic gains. Further, we investigate how decision-makers without a system react to the introduction of a DSS by quantifying changes in behavior and measuring fairness perceptions.

We find that the introduction of a DSS generally improves

overall market outcomes, which is primarily driven by improved proposer offers. Proposers are willing to use a black-box DSS, but still under-weigh the provided information. The introduction of interpretability appears to increase system usage, while there is evidence for a decrease in system trust as proposers revise their initial offers less. Roughly one-third of responders judges the introduction of a DSS as unfair and ask for more money. This harms responder income, proposer income and overall market efficiency. Endowing responders with additional explanations about the system partially mitigates perceptions of unfairness.

To facilitate further research and for the benefit of the HCOMP community, we publicly release our anonymised data, code, and additional information.[1]

2 Related Literature

Algorithmic Recommendations & Advice-Taking This paper relates to current research evolving around human capabilities and willingness to integrate algorithmically derived information or recommendations into their decision-making processes. From an economic point of view, this entails questions such as whether the utilization of decision-support systems induces efficiency gains, or what kind of institutional elements promote effective human-machine interactions. Prior research from (Oenkal et al. 2009) suggests that human decision-makers might generally discount forecasting advice more if they perceive it to come from a statistical model. Indeed, a long literature documents the supposed human tendency to prefer human experts (both themselves or external sources) over statistical or algorithmic predictions, even if the latter have been shown to be reliably more accurate (Meehl 1945; Grove and Meehl 1996; Grove and Lloyd 2006; Dawes, Faust, and Meehl 1989; Highhouse 2008). Recently, the concept of *algorithm aversion* has raised a lot of interest (see (Burton, Stein, and Jensen 2020) for a review). In their seminal paper, (Dietvorst, Simmons, and Massey 2015) illustrate that human actors learn differently from observing mistakes by an algorithm in comparison to mistakes by humans. In particular, even participants who directly observed an algorithm outperform a human were less likely to use the model after observing its imperfections. (Prahl and van Swol 2017) find a similar pattern in that participants rejected algorithmic forecasting advice more than ostensibly human advice after receiving bad recommendations. (Dietvorst, Simmons, and Massey 2016) propose that algorithm aversion is in part mediated by control, or a lack thereof. Thus, giving human decision-makers the opportunity to process and shape algorithmic output could enhance compliance with decision aids and thereby boost performance. There is also evidence that human decision-makers are particularly averse towards algorithmic systems for tasks that are perceived as subjective (Castelo, Bos, and Lehmann 2019). For our case, this research sheds doubt on user willingness to integrate DSS information while bargaining with another human. In so far as the system provides more accurate predictions than the decision-maker themselves, this would lead to reduced cooperation and efficiency losses. Some recent evidence however also suggests that there are situations in which human actors prefer algorithmic advice or output (Dietvorst, Simmons, and Massey 2015; Logg, Minson, and Moore 2018; Dijkstra, Liebrand, and Timminga 1998).

Interpretability We contribute to the growing, interdisciplinary literature on interpretability and how it relates to user trust, system utilization and fairness perceptions (see e.g. (Doshi-Velez and Kim 2017) for an overview). Recent studies suggest that interpretability does not necessarily translate into better model utilization or error-detection (Poursabzi-Sangdeh et al. 2018) and might even negatively affect people's accuracy perceptions (Nourani et al. 2019). In other research, accuracy was more important in furthering user trust than explanations (Papenmeier, Englebienne, and Seifert 2019). Yet evidence is mixed, with e.g. (Yeomans et al. 2017) showing that explanations for a recommender system can mitigate distrust and subsequently increase human preferences for the system. Stated model accuracy has also been shown to significantly affect model trust (Yin, Wortman Vaughan, and Wallach 2019). (Lage et al. 2019) demonstrate that the types of complexity users are confronted with interact with the effectiveness of explanations. Thus, there is evidence that interpretability elements do affect human behavior, but the conditions under which they are beneficial remain under-explored.

Algorithmic Decision-Making and Social Norms We draw from the literature on how algorithmic systems affect the effectiveness of human-human cooperation and social norms. In particular, this paper relates to the bargaining literature (Anand et al. 2018). Ultimatum bargaining is one of the most prominent games researched in experimental economics (Gueth, Schmittberger, and Schwarze 1982). Although the game setting seems simple, understanding behavior in this framework remains complex even after decades of research (Güth and Kocher 2014). It has been applied to a variety of issues such as culture (Henrich 2000; Henrich et al. 2005), gender (Gong and Yang 2012), child development (Harbaugh, Krause, and Liday 2003) or human-computer interaction (Sanfey et al. 2003). Importantly, it might be the most transparent tool to demonstrate the importance of social norms, psychology and emotions in real-life negotiations (Roth et al. 1991; van Damme et al. 2014). Prior findings relating to our research question broadly find that humans appear to exhibit less social concerns when interacting with autonomous computer agents (Sanfey et al. 2003; van 't Wout et al. 2006; de Melo and Gratch 2015), but are more cooperative when acting through an agent (de Melo, Marsella, and Gratch 2018).

3 Method and Experimental Setup

Ultimatum Game Our basic framework replicates the simplest design of the ultimatum game, modified by the strategy method. This common procedure has the advantage of providing more data, which is especially useful given our

[1] https://osf.io/rkzj2/?view_only=d891813a74ee48d39213e36aebdcf7f4

need for a large data set to train the ADM systems. A proposer X decides on the distribution of a pie with size p. X receives x and the responder Y receives y, where $x, y \geq 0$ and $x + y = p$. In a simultaneous process, the responder Y decides on a minimum offer z, where $z \geq 0$, and accepts the proposal $(x, y) = 1$ if $y \geq z$. If $z > y$, the responder rejects the offer $(x, y) = 0$. Payoffs are given by $\delta(x, y)x$ and $\delta(x, y)y$, i.e. if the responder Y rejects both earn nothing.

A straightforward solution of the game merely based on monetary outcomes implies that responder Y should accept all positive offers, which gives $\delta(x, y) = 1$ for $y > 0$. This is based on the rational that receiving something is better than receiving nothing, which is particularly true in a one-shot game without reputation being a factor.[2] This is anticipated by the proposers X, which has them offer the minimal positive amount. In consequence, X receives almost the whole pie p and Y receives little more than nothing.

However, prior experiments have shown that the optimal offer by the proposer amounts to 40% to 50% of the pie, since responders often reject lower offers (Camerer 2003; Oosterbeek, Sloof, and van de Kuilen 2004). These findings have led to influential theoretical work integrating other-regarding preferences such as fairness concerns into the traditional *homo oeconomicus* (Bolton and Ockenfels 2000; Fehr and Schmidt 1999).

Participants Participants were recruited via Amazon Mechanical Turk. We restricted the sample to workers from the United States who had completed at least 100 HITs with an approval rate of at least 80%. Participants enrolled on their own accord after reading a brief description of the experiment and having the option to take a look at the survey form. Participants immediately received a $0.50 participation fee on survey completion. Less than 6 minutes were required to complete the survey, and 76% of bargaining interactions were successful. Thus, participants received no less than an hourly wage of USD 8.5/h on average.

All participants first read the same basic instructions and had to answer four comprehensions questions correctly in order to proceed with the ultimatum game. Those who made more than four mistakes were dropped from the experiment (N = 242). We also added an attention check. Our final sample consisted of 1178 observations (45% female).

Procedure Participants then followed a link that randomly assigned them to the role of either a responder or a proposer. Proposers received $1 and were asked to make an offer to their responder on how to split the money between them. Responders chose the *minimum offer* they were willing to accept from the proposer. All treatments (see Table 1) followed this basic setup. We implemented T_0 as a standard one-shot ultimatum game without a decision-support system to gather the necessary training data.

[2]While this represents the weakly dominant strategy for Y, all distributions (x, y) can be established as equilibrium outcomes. For multiple equilibria consider a certain threshold $\bar{y}$ for acceptance by the responder Y, such that $[(x, y), \delta(\tilde{x}, \tilde{y}) = 1]$ if $\tilde{y} \geq y$ and $\delta(\tilde{x}, \tilde{y}) = 0$ otherwise.

Treatment	Role	DSS	DSS_info	Interpretability	Resp_info
T_0	both	-	-	-	-
$T_{1.0}$	proposer	yes	-	-	-
$T_{1.1}$	proposer	yes	-	-	not informed
$T_{1.2}$	proposer	yes	-	yes	-
$T_{1.3}$	proposer	yes	-	yes	informed
$T_{1.4}$	proposer	yes	-	yes + accuracy	-
$T_{2.0}$	responder	-	yes	-	-
$T_{2.1}$	responder	-	yes	yes	-

Table 1: Treatment overview. "Role" refers to the primary unit of observation. We additionally collected responder data in $T_{1.0}$ and $T_{1.4}$.

Treatments $T_{1.0}$ - $T_{1.4}$ were primarily concerned with proposer behavior. After making their first offer to the responder, proposers subsequently learned about the decision-support system and were then allowed to query it and make a revised offer.[3] This intervention allows us to directly infer changes causally elicited by the DSS while deploying a conservative measurement of system usage that reflects that in reality, most people use well-functioning heuristics and decision rules in bargaining and negotiation environments (Gigerenzer and Gaissmaier 2011; Allison and Messick 1990). In general, it is likely that DSS will predominantly augment already existing human decision processes, which are anchored in historical behavioral patterns.

Following (Yeomans et al. 2017), after making their revised offer, proposers were asked to indicate their agreement to statements about the explainability and their understanding of the system as well as whether they would have made a different offer if the responder was (not) aware of the DSS.[4]

$T_{1.0}$ and $T_{1.1}$ only differed in the information proposer received about the responders. In $T_{1.0}$, proposers did not know whether or not responders were aware of the DSS. In $T_{1.1}$, they were informed that the responder does not know about the system.[5] Thus, $T_{1.1}$ functions as a robustness check controlling for any e.g. fairness-related confoundings in the utilization of a DSS arising from potential transparency for responders. It could, for instance, be that proposers who assumed their responder to have knowledge of the system would be more reluctant to use it, since (i) the utilization it-

[3]We selected a featureless model that solely focuses on responder minimum offer distributions and searches for the fixed offer that maximizes proposer gains over the training sample. The model provides feedback on two dimensions. First, it calculates the probability that a proposer offer is the best offer, i.e. that it coincides with the responder minimum offer. Secondly, it shows the probability of an offer being accepted. This is based on the normalized cumulative histogram of each accepted offer in the training dataset, the training error and the predicted minimum offer for the responder. We added a bias for each possible offer to cope with the absence of some values in the dataset and to guarantee a strictly increasing probability with an increasing offer. The fixed offer maximizing proposer gains was $0.5, i.e. an even split.

[4]For $T_{1.0}$, the feedback was gathered in a follow-up survey where proposers had to play one more time using the DSS.

[5]Because we gathered both proposer and responder data for $T_{1.0}$ and matched proposers from $T_{1.1}$ with responders from $T_{1.0}$ for payment purposes, there was no deception involved.

self might be made transparent to the responder and (ii) the accuracy of the model might be impeded by such information. Since both situations constitute relevant abstractions of real-life decision contexts, any differences in proposer behavior might shed light on important consequences for DSS utilization across a range of functions.

In $T_{1.2}$, $T_{1.3}$ and $T_{1.4}$, we endowed proposers with an explanation of the model's functioning to increase interpretability. While $T_{1.2}$ only added information on the model's process, $T_{1.4}$ further gave explicit information on the model's accuracy. The example below depicts a system-level explanation (Gilpin et al. 2018) of the model's functioning presented to proposers in $T_{1.2}$.

"*The system was trained using 100 prior interactions of comparable bargaining situations. It learned a fixed optimal offer, by testing each possible offer on prior bargaining situations and was selected as the one that provided the highest average gain to proposers. Using the same process, the system also constructed an interval that judges offers that deviate from its recommendation.*"

In addition to the explanation in $T_{1.2}$, proposers in $T_{1.4}$ also received information regarding the accuracy of the model, as shown in the example below.

"*Following the AI System's recommendations, proposers can gain 80% of the pie left by responders. Following the AI System's recommendations, proposers can have 95% of their offers accepted. The probability of an offer being accepted is higher than 50% when the offer is greater than or equal to the recommended offer.*"

Finally, in $T_{1.3}$, proposers again received only the process-related information and further learned that their responder *does know* about the system.[6] Our intention was to capture any feedback effects induced by proposer expectations regarding responder behavior. For instance, proposers could expect responders to perceive the system as an unfair advantage and demand more money. Similar to $T_{1.1}$, $T_{1.3}$ thus provides a robustness check for our results from $T_{1.2}$.

$T_{2.0}$ and $T_{2.1}$ focus on responders as our main unit of observation. Similar to proposers, responders in both treatments first indicated their minimum offer without being aware of the proposer DSS, were then made aware and had the option to revise their initial minimum offer. While responders in $T_{2.0}$ did not receive any additional information on the DSS, we endowed responders in $T_{2.1}$ with the same explanation as proposers in $T_{1.2}$. Following the final minimum offer, responders where asked to indicate their agreement to four statements on a seven point Likert-scale.

Measures

To quantify changes in proposer and responder behavior as a response to the introduction of a decision-support system, we define three main measures.

System Usage – The system usage is reported by both the ratio of proposers who made at least one request to the DSS and the average number of unique requests made.

System Trust – The system trust is determined using the judge advisor system (JAS) metric's weight of advice (WOA) as reported by (Bonaccio and Dalal 2006) where the proposer is the judge and the decision support system is the advisor. WOA is defined as a function of proposer final offer ($offer_{final}$), the proposer initial offer ($offer$) and the DSS recommended offer (DSS_{offer})

$$WOA = \frac{offer_{final} - DSS_{offer}}{DSS_{offer} - offer}$$

We calculate the WOA for each condition over all participants whose initial offers without the system do not match the system's recommendation, i.e. an even split of 0.5$.[7]

Absolute change – To complement the WOA, we quantify the average absolute deviation of a proposer's (responder's) final offer (final minimum offer) from their initial offer (minimum offer). Since subjects did not receive one recommendation, but were able to inquire the system multiple times and gather information about the expected gains and associated risk for different offers, the WOA potentially does not capture the full range of system-induced behavioral changes.

4 Results

Descriptive statistics of the different games are presented in Table 2 for proposers and Table 3 for responders. In each treatment, proposer offers increased on average after being able to use the DSS. Figure 1 shows cumulative distribution functions plotting the difference between initial proposer offers ($offer$) and revised offers ($offer_{final}$). Across treatments, more proposers increased their offer than vice versa, and there is considerable variation across treatments.

Statistics	T_0	$T_{1.0}$	$T_{1.1}$	$T_{1.2}$	$T_{1.3}$	$T_{1.4}$
N	103	105	105	103	105	102
Mean (initial)	43.59	41.48	44.71	46.99	46.67	44.56
Mean	-	45.67	49.38	48.16	49.29	46.47
Standard Dev.	14.97	14.50	14.59	13.32	14.08	17.37

Table 2: Proposers game statistics. Statistics are reported based on final offers in each treatment unless explicitly stated otherwise.

Statistics	T_0	$T_{2.0}$	$T_{2.1}$
N	103	105	105
Mean (initial)	40.10	41.38	38.91
Mean	-	43.81	41.10
Standard Dev.	18.88	21.13	16.50

Table 3: Responders minimum offer statistics. Statistics are reported based on final minimum offers in each treatment unless explicitly stated otherwise.

For responders, we also find directional increases in the minimum sum that players were willing to accept after learning about the proposer's option to use a DSS. Although there

[6] Since we matched proposers in $T_{1.3}$ with responder data from $T_{2.0}$ for payment, there was no deception involved.

[7] There was no significant difference in the fraction of proposers whose initial offers matched the system's recommendation between treatments [$\tilde{\chi}^2(4) = 2.19, p = 0.700$].

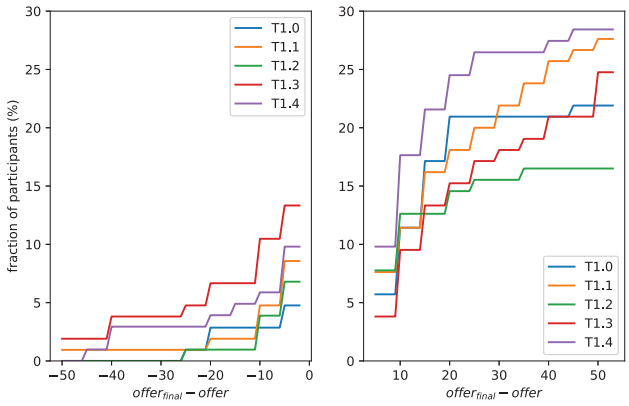

Figure 1: Cumulative histogram curves in percentages of differences in proposers' offers before and after the introduction of the DSS.

was considerable downward adjustment, Figure 2 shows that the majority increased their required share of money, which is diametrical to the changes in proposer behavior.

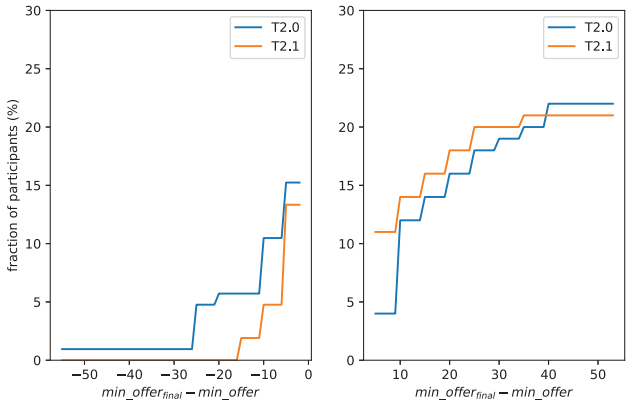

Figure 2: Cumulative histogram curves in percentages of differences in responders' minimum requested offers before and after learning about the DSS.

Main Results

For clarity sake, we examine proposer and responder behavior separately. Any essential inferences for market efficiency or human cooperation will be drawn in the discussion.

Proposer + DSS

To examine whether and to what extent proposers used the black box DSS, we first concentrate on treatments $T_{1.0}$ and $T_{1.1}$. Recall that participants first made an offer without the system, and were then allowed to inquire the system and make a revised offer. Table 4 shows first and second stage proposer offers, the absolute change induced by the DSS, the estimated WOA as well as statistics on system usage.

In both treatments, roughly 90% of proposers inquired the system at least once, with 2.0 and 1.9 inquiries on average respectively. This lead to substantial improvements on a number of indicators. In $T_{1.0}$, on average 31% of first offers would have been rejected by the responders. For revised offers, that share drops to 23%, improving both average proposer (+2.76 cents) and average responder income (+5.81 cents). Thus, introducing a black box DSS significantly reduced rejection rates $[t(104) = 3.12, p = 0.002]$. Despite this, realized rejection rates and income levels were far behind the model's performance (see table 4). Participants could have performed better, had they trusted the model more (WOA = 0.54).

	$T_{1.0}$	$T_{1.1}$	$T_{1.2}$	$T_{1.3}$	$T_{1.4}$
$offer$	41.48 (15.32)	44.71 (16.78)	46.99 (13.92)	46.67 (17.93)	44.56 (19.18)
$offer_{final}$	45.67 (14.50)	49.38 (14.59)	48.16 (13.32)	49.29 (14.08)	46.47 (17.37)
$absolutechange$	5.619 (13.44)	6.952 (13.00)	2.524 (6.018)	8.333 (14.47)	5.637 (10.36)
WOA	0.54 (0.58)	0.65 (0.67)	0.61 (0.88)	0.81 (0.93)	0.66 (1.3)
$uniquerequests$	1.990 (1.842)	1.886 (1.872)	2.612 (2.365)	1.971 (1.778)	2.608 (2.474)
$requested$	0.905 (0.295)	0.895 (0.308)	0.913 (0.284)	0.886 (0.320)	0.912 (0.285
$rejected_{nodss}$	0.314 (0.466)	0.295 (0.458)	0.262 (0.442)	0.333 (0.474)	0.314 (0.466)
$rejected_{dss}$	0.0571 (0.233)	0.0571 (0.233)	0.0874 (0.284)	0.181 (0.387)	0.0882 (0.285)
$rejected_{final}$	0.229 (0.422)	0.190 (0.395)	0.194 (0.397)	0.257 (0.439)	0.265 (0.443)

Table 4: Behavioral statistics for the proposer-centric treatments. $rejected_{nodss}$ refers to rejection rates based on initial proposer offers, $rejected_{dss}$ refers to rejection rates had all proposers chosen the offer that maximized their expected income according to the DSS and $rejected_{final}$ are the actually realized rejection rates after proposers utilized the system. $requested$ is a dummy variable capturing whether a proposer made at least one request, $uniquerequests$ is the average number of unique inquiries to the system.

Moreover, there are no significant differences between $T_{1.0}$ and $T_{1.1}$, indicating that proposers either assumed that responders would not receive information about the DSS anyway, or that the proposer decision is independent of responder transparency. According to the self-reported data, around 30% of proposers in $T_{1.1}$ agreed that they would have made a different offer, had the responder been informed about the system (see table 5). This suggests that proposers' intuition matched the endowed responder information, and that we'd see differences in proposer behavior depending on whether they receive information that responder *do* know about the system. We will test this intuition in treatments $T_{1.2}$ and $T_{1.3}$.

Result 1: Proposers are willing to use a black box DSS and consequently make better offers. However, since proposers under-weigh the system's advice, rejection rates

and aggregated income remain far from optimal.

	Disagree	Neutral	Agree	Mean	Std. Dev.
$T_{1.1}$	55 (52.4)	18 (17.1)	32 (30.5)	3.4	1.9
$T_{1.3}$	63 (60.0)	24 (22.9)	18 (17.1)	3.0	1.7

Table 5: Proposer answers to the following question: *"I would have made another offer if the responder was (not) informed about the AI System"*

	Disagree	Neutral	Agree	Mean	Std. Dev.
$T_{1.0}$	7 (14)	7 (14)	36 (72)	4.96	1.4
$T_{1.1}$	16 (15.2)	15 (14.3)	74 (70.5)	5.11	1.5
$T_{1.2}$	14 (13.6)	12 (11.7)	77 (74.8)	5.22	1.6
$T_{1.3}$	14 (13.3)	15 (14.3)	76 (72.4)	5.02	1.3
$T_{1.4}$	14 (13.7)	17 (16.7)	71 (69.6)	4.9	1.4

Table 6: Proposer answers to the following question: *"I could understand why the AI System thought the responder would accept a given offer."*

	Disagree	Neutral	Agree	Mean	Std. Dev.
$T_{1.0}$	26 (52)	8 (16)	16 (32)	3.7	1.8
$T_{1.1}$	50 (47.6)	14 (13.3)	41 (39)	3.7	1.7
$T_{1.2}$	61 (59.2)	12 (11.7)	30 (29.1)	3.4	1.9
$T_{1.3}$	46 (43.8)	23 (21.9)	36 (34.3)	3.9	1.6
$T_{1.4}$	46 (45.1)	19 (18.6)	37 (36.3)	3.7	1.7

Table 7: Proposer answers to the following question: *"It is hard for me to explain how the AI-System judged my offers."*

Proposer + DSS + Interpretability

To examine whether the intelligibility of our DSS prevents proposers from using it optimally, treatments $T_{1.2}$, $T_{1.3}$ and $T_{1.4}$ endow proposers with additional information about the system. In $T_{1.2}$, proposers were informed about the model's process and did not receive any notice on whether responders learned about the system. It thus provides the direct counterfactual to $T_{1.0}$. While the share of proposers who inquired the system at least once does not differ between treatments, proposers in $T_{1.2}$ made significantly more unique inquiries to the system $[t(206) = 2.12, p = 0.036]$. Despite higher usage, however, absolute change between the first and the revised offer was significantly lower in $T_{1.2}$ $[t(206) = -2.14, p = 0.034]$ and the WOA does not differ significantly $[t(79) = 0.45, p = 0.653]$. These results suggest that our interpretability intervention did not increase user trust in the system and, if anything, discouraged proposers from integrating the system's advice. This is particularly interesting since system usage increased, meaning that subjects were not deterred from acquiring information by inquiring the DSS, but only from adjusting their initial decision. Self-reported understandability and explainability were not significantly affected by the additional explanations (see tables 6 and 7).

Result 2: Interpretability increases system usage, but decreases system trust. Proposers who received additional information on the model revised their initial offers less intensely and thus missed out on substantial monetary gains.

In $T_{1.3}$, proposers learned that responders knew about the DSS. Thus, comparing $T_{1.2}$ and $T_{1.3}$ allows us to examine how proposer expectations regarding responder reactions towards the DSS feed back into the main interaction. While system usage decreased $[t(206) = -2.21, p = 0.028]$, proposers appeared to integrate advice much more into their revised offer. On average, proposers in $T_{1.3}$ revised their initial offer by 8.3 cents (+5.81) $[t(206) = 3.77, p = 0.000]$ and had a substantially larger WOA (+0.2) $[t(81) = 0.99, p = 0.326]$. In line with these results, endowing proposers with even more explanations by adding information about the system's accuracy in $T_{1.4}$ but removing information about responders reverts behavior back to high usage and apparently lower system trust. While the number of unique inquiries to the system increases back to the level of $T_{1.2}$ and is thus significantly higher than in $T_{1.3}$ $[t(205) = 2.13, p = 0.034]$, both the absolute change between offers (-2.69) $[t(205) = -1.53, p = 0.126]$ and the WOA (-0.15) $[t(90) = 0.625, p = 0.534]$ decrease, albeit not significantly. Still, combined results from the three treatments suggests that interpretability increases system usage, but has either no or a slightly negative effect on system trust. Once proposers are informed that their counterpart is aware of the system, they appear to utilize its advice more.

Result 3: Endowing proposers with additional information on the system's accuracy does not increase system trust or usage. However, proposers appear to weigh algorithmic information more strongly once responders are informed about the system.

Responder + DSS

In $T_{2.0}$ and $T_{2.1}$, we focus on responder behavior as a response to the introduction of a proposer DSS. We first find that responders in $T_{2.0}$ who do not receive any explanations about the system substantially revise their minimum offer by on average 10.24 cents. This is associated with a small, directional increase of the average minimum offer. Thus, subjects corrected both upwards and downwards. In line with these results, 39% of responders agreed that they would have chosen a different minimum offer, if the proposer did not have a DSS (see table 8).

	Disagree	Neutral	Agree	Mean	Std. Dev.
$T_{2.0}$	54 (51.43)	10 (9.52)	41 (39.05)	3.47	2.03
$T_{2.1}$	62 (59.05)	12 (11.43)	31 (29.52)	3.08	1.86

Table 8: Responders' agreement to the following statement: *"I would have chosen a different minimum offer if the proposer did not have a recommendation system"*

Comparing rejection rates between the first and the revised minimum offer, we find that the changes induced by the DSS tendentially decrease the overall number of

successful interactions. Whereas responder decisions without knowledge of the system would have resulted in rejection rates of 22.9% in $T_{2.0}$, the implemented rejection rate based on the revised minimum offer was 27.6% $[t(104) = 1.29, p = 0.198]$. Average responder income also decreased marginally (-1.8 cents).

Result 4: Around 40% of responders adjust their minimum offer upon learning that their proposer has the option to use a DSS.

One reason why responder minimum offers tended to increase might lie in their fairness perceptions. To test this, responders indicated whether they agreed that it is unfair that the proposer gets to use a recommendation system. In $T_{2.0}$, 34% agreed with the statement. These 34% made up 62% of all responders who revised their minimum offer upon learning about the system. Further, they adjusted their initial offers significantly upwards by about 7.8 cents on average $[t(35) = 1.69, p = 0.049]$. Feelings of unfairness thus seem to drive higher responder offers, making the bargain overall more risky and less likely to be successful.

Result 5: Roughly one-third of responders judged the introduction of a DSS as unfair and concomitantly demanded significantly more money from the proposer.

	Disagree	Neutral	Agree	Mean	Std. Dev.
$T_{2.0}$	49 (46.67)	20 (19.05)	36 (34.29)	3.47	1.92
$T_{2.1}$	67 (63.81)	12 (11.43)	26 (24.76)	3.09	1.86

Table 9: Responders' agreement to the following statement: *"I think it is unfair that the proposer gets to use a recommendation system."*

Responder + DSS + Interpretability

$T_{2.1}$ endows responders with the same process-related information proposers received in $T_{1.2}$. As a result, responders revised their initial minimum offer by on average 4.2 cents, which is significantly less than in $T_{2.0}$ $[t(208) = 2.84, p = 0.005]$. The share of subjects indicating different minimum offers if the proposer did not have a DSS also dropped by 10 percentage points to 29%. Similar to $T_{2.0}$, these DSS-induced changes increase rejection rates from 20% to 24.7% $[t(104) = 1.91, p = 0.058]$ and reduce responder income by 2.8 cents on average $[t(104) = 1.96, p = 0.053]$. The share of responders who judged the availability of a DSS on the proposer's side as unfair decreased by almost 10 percentage points, leaving roughly 25% of responders. Contrary to $T_{2.0}$, these 25% only make up 37% of all responders who revised their offer, and their average increase in minimum offers (+3.3 cents) is not significant $[t(25) = 1.15, p = 0.26]$. Thus, it appears that the introduction of additional explanation reduced perceptions of unfairness, which might explain why on average adjustments where significantly lower in $T_{2.1}$ than in $T_{2.0}$. Despite this, responders still asked for significantly more once they learned about the system $[t(104) = 2.33, p = 0.022]$ and thereby decreased both their and their proposers average income.

Result 6: Endowing responders with an explanation about the DSS decreases average absolute adjustments of their initial minimum offers as well as self-reported change in behavior due to the system.

Result 7: Endowing responders with an explanation about the DSS decreases perceptions of unfairness.

5 Discussion

The question whether human decision-makers effectively utilize algorithmic systems and integrate them into their decision processes will be a key factor in harnessing the economic potential of machine-learning models. Moreover, these developments will depend crucially on a society's institutional framework, like novel transparency or interpretability regulations. Still, only little is known about how human decision-makers use decision-support systems, whether or under what conditions interpretability does lead to better decision-making, and how changes in behavior translate into human cooperation and market outcomes. Our experiments provide initial results to help tackle and advance our current understanding of these issues.

RQ1: Do human decision-makers integrate advice from a black-box DSS? We show that users are willing to integrate advice from a black box DSS. As a result, both the share of successful interactions and aggregated income increase, with responders receiving disproportionally large gains. However, proposers under-weigh the system's advice, and thereby lose out on substantial economic gains. Knowledge about responder non-transparency does not appear to further trust in the model, although a significant share of roughly 30% agreed that they would make different offers conditional on the responder having transparent knowledge of the system. This was confirmed in a later analysis, where notice about responder transparency decreased system utilization but increased offer adjustments in line with the system's recommendations. We can therefore derive two additional insights. First, despite some willingness to use the system, thinking about ways to increase user trust still holds many economic benefits, and second, transparency regulations on behalf of stakeholders affected by a DSS should take potential feedback effects into account.

RQ2: Do human decision-makers exhibit different social concerns while interacting with another human using an ADM system? In our experiment, responders were only affected by the DSS through the proposers decision. Nevertheless, we find that responders increase the minimum offer they are willing to accept once they learn about the system. Without an explanation, this effect is primarily driven by a subset of responders who judge the system's availability as unfair. It therefore seems likely that perceptions of unfairness are at least partially responsible for the change in minimum offers. Another possibility is that fairness expectations change when humans are assisted by a decision-support system. Instead of demanding more money because people feel unfairly treated, it might be

that they either frame offer fairness differently conditional on the availability of a DSS, or that people are harder to convince that an offer is fair when it is derived with the help of a support system. This might explain why explanations were effective in reducing feelings of unfairness.

RQ3: Does interpretability mediate system trust and fairness perceptions? Endowing responders with an explanation decreases the share who find it unfair that the proposer can use a DSS, but responders on average still revise their minimum offer upwards. This negatively impacts overall market outcomes, since the share of successful interactions as well as aggregated income decrease. Explanations might mitigate, but not negate negative feedback effects.

For proposers, interpretability induces higher system usage, which does not translate into higher system trust. If anything, those who receive additional explanations weigh the model's advice less intensely and make on average less effective offers. We also do not find a difference between only process-related, and full interpretability that offers information on the system's accuracy. In sum, the effects of interpretability on overall market outcomes are ambivalent, and tend to affect human stakeholders differently depending on their role within the bargaining situation.

Practical Implications

A failure to successfully negotiate or cooperate often entails losses for all actors involved. Our results highlight the potential of decision support systems to increase the efficacy of human-human interactions, which benefits everyone, irrespective of the allotment of the system. Nevertheless, insufficient system trust still causes economic losses, which in reality might be represented by products not being sold, a breakdown of vital negotiations such as unions vs. firms or trade negotiations. Regulators should therefore be interested in constructing an institutional framework that facilitates system trust – conditional on the system being safe and useful. Here, we find that interpretability and transparency rules might induce ambivalent effects on market outcomes. In particular, those who are not endowed with a system appear to behave under different social concerns once they observe that another actor uses a DSS. As a result, forcing businesses, firms or organizations to disclose whenever algorithms are used to augment or substitute decision-making could have unintended consequences for all sides. Simply increasing the interpretability of a system does not automatically improve decision-making or human cooperation, but could assist in alleviating perceptions of unfairness.

Caveats and Limitations

The simplicity of the selected model made it easy to explain the training and prediction process. However, simplistic models may be perceived as inaccurate by the participants. On the other hand, while the model alone is simplistic, it doesn't directly provide the best offer proposers should make. Efforts were invested to make the ADM interface simple and understandable.

We used a specific interpretability intervention that for various reasons might not generalize towards all interpretability instruments. Further research is needed to determine the optimal design of explanations conditional on the social context. Moreover, the provided explanation did not increase self-reported measures of understandability. Explanations that make people feel like they understand more about a system's processes could have different effects. Still, despite no measurable effect on understandability, subjects in our study received additional information that clearly explained how the DSS was trained and achieved different predictions. As such, it did increase interpretability and opened the black box for participants. Future research should consider how interpretability and understandability interact in the context of human behavior.

Using a DSS increased proposers' income on average, but still caused some participants to switch from a good offer to a bad one. This effect remained with the introduction of explanations on how the system was trained and works. Decision support systems being able to give the user personalized explanations for the same predictions could improve overall market outcomes.

Since WOA is only partially suited as a measure of trust for a DSS that offers a range of information rather than a single recommendation and allows decision-makers to integrate their personal preferences e.g. with regard to risk-taking, we base a lot of our analysis on the average absolute deviation. We thereby acknowledge that the question of system trust does not necessarily depend on improvements in decision-making. Rather, an increase in system trust should be reflected by a higher impact, and thus measured by changes in behavior induced by the system. It is reasonable to assume that some decision-makers will be willing to decrease their expected gains for a chance to earn more than an even split. Increased system trust could decrease the uncertainty attached to such a decision and thereby motivate offers that are more unequal than a proposer's initial offer. This would be interpreted as a decrease in system trust by the WOA-metric, but captured by the average absolute deviation.

6 Conclusion

In this paper, we studied how the introduction of a DSS affects human-human interactions in a bargaining context. We show that effects are heterogeneous depending on the human stakeholders involved. Introducing a black-box system increases the efficacy of human cooperation and overall market outcomes, but users under-weigh the system's advice and thereby do not exploit its full potential. Furthermore, economic gains are exclusively driven by improved user decisions, whereas people who do not have a DSS themselves become more demanding upon learning that their counterpart has the option to use one. This effect appears to be significantly influenced by perceptions of unfairness and decreases cooperation levels as well as market efficiency.

On the user side, we find that interpretability increases system usage, but does not lead to higher system trust or better market outcomes. However, perceptions of unfairness from those not having a system available are mitigated by increased interpretability.

References

Allison, S. T., and Messick, D. M. 1990. Social decision heuristics in the use of shared resources. *Journal of Behavioral Decision Making* 3(3):195–204.

Anand, A.; Bizer, K.; Erlei, A.; Gadiraju, U.; Heinze, C.; Meub, L.; Nejdl, W.; and Steinroetter, B. 2018. Effects of algorithmic decision-making and interpretability on human behavior: Experiments using crowdsourcing. In *HCOMP (WIP&Demo)*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. propublica 2016.

Bolton, G. E., and Ockenfels, A. 2000. Erc: A theory of equity, reciprocity, and competition. *American economic review* 90(1):166–193.

Bonaccio, S., and Dalal, R. S. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101(2):127–151.

Burton, J. W.; Stein, M.-K.; and Jensen, T. B. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33(2):220–239.

Camerer, C. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Cambridge University Press.

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

Dawes, R. M.; Faust, D.; and Meehl, P. E. 1989. Clinical versus actuarial judgment. *Science* 243(4899):1668–1674.

de Melo, C., and Gratch, J. 2015. People show envy, not guilt, when making decisions with machines. *International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China*.

de Melo, C. M.; Marsella, S.; and Gratch, J. 2018. Social decisions and fairness change when people's interests are represented by autonomous agents. *Autonomous Agents and Multi-Agent Systems* 32(1):163–187.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2016. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

Dijkstra, J. J.; Liebrand, W. B. G.; and Timminga, E. 1998. Persuasiveness of expert systems. *Behaviour & Information Technology* 17(3):155–163.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Fehr, E., and Schmidt, K. M. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114(3):817–868.

Gigerenzer, G., and Gaissmaier, W. 2011. Heuristic decision making. *Annual review of psychology* 62:451–482.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. IEEE.

Glaeser, E.; Hillis, A.; Kominers, S. D.; and Luca, M. 2016. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review* 106(5):114–118.

Gong, B., and Yang, C.-L. 2012. Gender differences in risk attitudes: Field experiments on the matrilineal mosuo and the patriarchal yi. *Journal of economic behavior & organization* 83(1):59–65.

Grove, W. M., and Lloyd, M. 2006. Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology* 115(2):192–194.

Grove, W. M., and Meehl, P. E. 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law* 2(2):293–323.

Gueth, W.; Schmittberger, R.; and Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* 3(4):367–388.

Güth, W., and Kocher, M. G. 2014. More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization* 108:396–409.

Harbaugh, W.; Krause, K.; and Liday, S. 2003. Bargaining by children.

Henrich, J.; Boyd, R.; Bowles, S.; Camerer, C.; Fehr, E.; Gintis, H.; McElreath, R.; Alvard, M.; Barr, A.; Ensminger, J.; et al. 2005. "economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and brain sciences* 28(6):795–815.

Henrich, J. 2000. Does culture matter in economic behavior? ultimatum game bargaining among the machiguenga of the peruvian amazon. *American Economic Review* 90(4):973–979.

Highhouse, S. 2008. Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology* 1(3):333–342.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S. J.; and Doshi-Velez, F. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 59–67.

Liem, C. C.; Langer, M.; Demetriou, A.; Hiemstra, A. M.; Wicaksana, A. S.; Born, M. P.; and König, C. J. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable*

and Interpretable Models in Computer Vision and Machine Learning. Springer. 197–253.

Logg, J. M.; Minson, J. A.; and Moore, D. A. 2018. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Meehl, P. E. 1945. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN, US: University of Minnesota Press.

Nourani, M.; Kabir, S.; Mohseni, S.; and Ragan, E. D. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 97–105.

Obermeyer, Z., and Emanuel, E. J. 2016. Predicting the future – big data, machine learning, and clinical medicine. *The New England journal of medicine* 375:1216–1219.

Oenkal, D.; Goodwin, P.; Thomson, M.; Gönül, S.; and Pollock, A. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22:390–409.

Oosterbeek, H.; Sloof, R.; and van de Kuilen, G. 2004. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics* 7:171–188.

Papenmeier, A.; Englebienne, G.; and Seifert, C. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Prahl, A., and van Swol, L. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36:691–702.

Roth, A. E.; Prasnikar, V.; Okuno-Fujiwara, M.; and Zamir, S. 1991. Bargaining and market behavior in jerusalem, ljubljana, pittsburgh, and tokyo: An experimental study. *The American Economic Review* 1068–1095.

Sanfey, A.; Rilling, J.; Aronson, J.; Nystrom, L.; and Cohen, J. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300(5626):1755–1758.

van Damme, E.; Binmore, K.; Roth, A.; Samuelson, L.; Winter, E.; Bolton, G.; Ockenfels, A.; Dufwenberg, Martin und Kirchsteiger, G.; Gneezy, U.; Kocher, M.; Sutter, M.; Sanfey, A.; Kliemt, H.; Selten, R.; Nagel, R.; and Azar, O. 2014. How werner gueth's ultimatum game shaped our understanding of social behavior. *Journal of Economic Behavior & Organization* 108:292–318.

van 't Wout, M.; Kahn, R. S.; Sanfey, A. G.; and Aleman, A. 2006. Affective state and decision-making in the ultimatum game. *Experimental Brain Research* 169(4):564–568.

Yeomans, M.; Shah, A.; Mullainathan, S.; and Kleinberg, J. 2017. Making sense of recommendations. *Journal of Behavioral Decision Making*.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 279. ACM.