Bargaining with a Machine

An Experimental Economics Approach on Efficiency and Interpretability

Avishek Anand, Alexander Erlei, Ujwal Gadiraju, Lukas Meub

ABSTRACT

We conduct a large experimental study to examine the effects of algorithmic decision-making (ADM) on human behavior. Our analysis builds on the well established ultimatum game and comprises eight treatments where the sender is either human, a human supported by decision-support system or an autonomous algorithmic agent based on machine-learning. Further, we vary the associated interpretability of the ADM system. The underlying machine learning model is trained by either human or human-machine interactions. We find.... We contribute to both computer science...and economics...Our results clearly indicate the relevance to broaden the scope of behavioral economics to embrace novel forms of human-machine interaction [in case of submitting to an economic journal].

KEYWORDS

Interpretability; Machine Learning; Behavioral Economics; Crowdsourcing; Algorithmic Decision Making; Fairness

ACM Reference Format:

Avishek Anand, Alexander Erlei, Ujwal Gadiraju, Lukas Meub. 2019. Bargaining with a Machine: An Experimental Economics Approach on Efficiency and Interpretability. In *Proceedings of Preliminary Draft (Preliminary Draft DokSem).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

The ever increasing ability of economies and societies all over the world 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 [59], engineering design or urban planning [32].**LM: insert references** 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, reinterpret their role within crucial decision-making processes and even cooperate or compete with autonomous algorithmic agents. Thus, the integration of ADM systems is accompanied by a number of challenges within the human-agent space, 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, institutions and social desirabilities. However, past research in computer and the behavioral sciences has largely neglected these essential aspects of algorithmic decision-making.

This paper examines whether human behavior changes with the introduction of an ADM system, and, if so, how market outcomes are affected. In particular, we ask whether human decision-makers exhibit different social concerns while interacting either with an ADM system or another human using an ADM system, and whether these potential changes subsequently induce welfare gains. Further, we measure the effect of interpretability on human behavior and overall income, thereby contributing to the current discussion around the intelligibility of algorithms by providing rigorous quantifiable data.

To attain a rich setting, our novel experimental approach is based on the ultimatum bargaining game and induces variation in multiple dimensions. We introduce both a decision-support system and an autonomous algorithmic agent, which allows us to explore relevant behavioral variations from different perspectives.[1] We are the first ones to introduce actual machine-learning systems into an economic experiment. Our strict methodological framework adheres to experimental economic standards and builds on existing studies to construct novel, clean counterfactuals. Rather than limiting our analysis to individuals that interact directly with ADM systems, we conduct an exhaustive analysis that also involves individuals that are indirectly affected by potential market disruptions. Finally, we explore how our machine-learning models adjust to human-agent interactions, thus highlighting structural changes in human behavioral patterns and long-term market effects.

In line with Endsley [25], we define decision-support systems as computer systems that are programmed to assist human decision-makers during the decision-process but do not select a course of action autonomously. In contrast, autonomous agents are capable of navigating all decision-making phases on their own: they actively learn from data, generate options, select a course of action and finally implement it based on a specific objective function. Both have potentially highly disruptive effects in various markets and might change human (economic) behavior fundamentally. Yet, they also provoke very different research questions about human decision-making, economic efficiency and social-cognitive factors.

For decision-support systems, many of the relevant issues evolve around human capability and willingness to integrate algorithmically derived information or recommendations into their decision-making processes. From an economic point of view, this entails

Preliminary Draft DokSem, 2018

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

[1] If not mentioned otherwise, all (economic) interactions involving at least one human and one algorithmic entity will be referred to as human-agent interactions. **LM: ausbauen**

questions such as whether the utilization of decision-support systems induces efficiency gains, or what kind of institutional elements promote effective human-machine interactions. Recently, the concept of *algorithm aversion* has raised a lot of interest. In their seminal paper, Dietvorst et al. [19] illustrate that human actors learn differently from observing mistakes by an algorithm in comparison to mistakes by another human actor or by themselves. In several experimental forecasting tasks, human participants were more likely to rely on an inferior human forecaster after watching a statistical algorithm err. In particular, even participants who directly observed the algorithm outperform a human were less likely to use the model after observing its imperfections. Prahl and van Swol [63] find a similar pattern in that participants rejected algorithmic forecasting advice more than ostensibly human advice after receiving bad recommendations. Dietvorst et al. [20] 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. Prior evidence from Oenkal et al. [60] suggests that human decision-makers might generally discount forecasting advice more if they perceive it to come from a statistical model. Moreover, when confronted with advice from both ostensibly human experts and statistical methods, subjects appeared to weight the apparently human advice more heavily. 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 [14, 36, 37, 43, 54]. There is also evidence that experts who utilize decision-support system are more scrutinized by consumers [3, 24, 62, 69]. If these patterns prevent a widespread usage of decision aids that complement human judgments productively and thus ultimately add to social welfare, it would be useful to identify conditions and factors that induce trust in these systems and encourage decision-makers to make use of their potential.

Recent evidence, however, points to humans showing an initial preference for algorithmic decision aids. In Dietvorst et al. [19], the majority of subjects who did not observe an algorithm err prior to their decision exhibited preferences for the statistical model over a human forecaster. Logg et al. [52] conducted six experiments to examine this phenomenon. They coin the phrase *algorithm appreciation* and find that participants weighed identical advice more coming from an algorithm than from another, albeit non-expert, human. These results are consistent with work from Dijkstra et al. [21], in which human subjects evaluated advice from "expert systems" as being more rational. Considering logic problems, subjects were more likely to agree with the same argument when it came from an "expert system" rather than from a "human".

In sum, evidence on the effectiveness and acceptance of decision-support systems is mixed and largely restricted to forecasting tasks. In particular, economic research is almost entirely absent, and many studies forgo proper monetary incentives that adhere to economic standards and have been shown to strongly influence human behavior [70]. One limitation of conventional forecasting tasks is that they do not capture important factors of interpersonal interactions such as social preferences and reciprocity. Thus, we still lack an understanding on how the utilization of decision-support systems affects economic human relations, e.g. through the suppression of fairness concerns in favor of self-interested strategic behavior.

For the case of autonomous agents, the potential consequences could be much more disruptive. Over the last twenty years, behavioral sciences have been able to substantially enhance traditional models of human behavior and subsequently produced many different economic and political insights. While this process has endowed us with a diverse collection of models and theories, they might fail to externalize to human-agent interactions. For instance, it remains an open questions whether and in what way the introduction of autonomous ADM systems disrupts essential behavioral patterns such as trust, reciprocity, reputation and fairness **LM: add references**. These concepts have been shown to crucially moderate human interaction and enhance social welfare, e.g. by complementing incomplete contracts in labor markets **LM: add references**, mitigating the exploitation of information asymmetries **LM: add references** or overcoming social dilemmas in human cooperation [11, 28] **LM: add references**.

Existing evidence, however, suggests that human actors do differentiate between agents and humans and judge them quite differently. Sanfey et al. [67] used functional magnetic resonance imaging (fMRI) to examine the neural processes involved in the decision-making of players in an ultimatum game. When confronted with unfair offers from a computer rather than a human, participants showed less activation in brain regions associated with negative emotions. Further, human responders rejected unfair offers from human proposers significantly more often than from a deterministic algorithm representing a computer. This was also replicated by van 't Wout et al. [73] and Dunn et al. [23]. Nelissen and Zeelenberg [58] find third-party punishment in a one-shot dictator game to be significantly lower when dictators are substituted by random computer-generated transfer-decisions. In a traditional prisoner's dilemma, human-human interactions activated different brain regions than human-computer interactions and humans tended to cooperate more with other humans than with computers [65].

One caveat of these studies is that they relate to human-agent interactions only indirectly, since their "computer" players were not based on sophisticated machine-learning algorithms and presumably did not evoke any perceptions of an "intelligent", autonomous agent. Rather, they were mostly presented as "mindless", generally deterministic and sometimes even random mechanisms. However, the degree of perceived agency and mindfulness has been suggested to crucially influence the way humans interact with ostensibly autonomous agents and algorithmic computer programs [18, 26, 50, 53, 75]. Some research on social decision-making indicates that humans do apply social norms when interacting with autonomous agents [55, 64], but to a significantly lower degree [15, 16]. Participants have been shown to exhibit a higher propensity to cheat when reporting their outcomes to non-human entities [12], make fairer allocation offers to ostensibly human team members [74] and react differently to virtual emotion expressions depending on whether they thought an agent or a human behind the virtual presence [17]. It thus remains unclear how human decision-makers will adapt when confronted with an algorithmic agent, particularly within rigorous economic frameworks, or how individual characteristics and institutions might shape these processes. However, there is considerable evidence suggesting that humans

weigh social factors like other-regarding preferences less when interacting with a non-human actor, which can in turn hurt market efficiency. Even if market efficiency does not depend on the absence of strategically selfish behavior, e.g. in ultimatum bargaining, there might be important consequences for income distribution and inequality. On the other hand, market efficiency could also increase when e.g. consumers are more willing to accept mechanisms like price discrimination as long as they are implemented by an ADM system rather than a human decision-maker. Understanding the manifold ways in which ADM systems and machine-learning change traditional patterns of human behavior will be critical in identifying necessary regulatory adjustments, predicting policy effects and designing systems that produce economically efficient and socially desirable decision-making. Here, we provide substantial novel evidence.

First, we contrast human-human and human-agent interactions in a rigorous experimental setup to map out differences in behavior and how these differences affect overall market efficiency. Contrary to existing studies, our ADM system does not follow a strictly predetermined strategy but learns and engages with each responder individually. This allows us to make compelling inferences about efficiency losses or gains that could translate into real world situations. Second, we analyze how the machine-learning model adapts to potential behavioral shifts triggered by addition of an ADM system. **LM: vermerk** This is particularly relevant for the long-term perspective of markets which are increasingly shifting towards human-agent interactions. Third, subjects know that they are playing with an autonomous algorithmic decision-making entity that is based on machine-learning. We increase the salience of the proposed "intelligence" of the agent, to produce a setting that reasonably describes modern and future economic interactions involving AI. de Melo and Gratch [15] provide the study that most closely resembles our research. However, their experimental setup does not offer clean counterfactuals in line with economic standards.[2] Further, they did not actually deploy a machine-learning algorithm, are not interested in decision-support systems or interpretability and lack effective monetary incentives.

Next to an adequate understanding of the relevant behavioral models that drive human-agent interactions, it is important to consider the institutional framework they are functioning in. These include, but are not limited to, the regulatory structure of a sector, a country or a supranational body like the EU. The current paper identifies requirements regarding *interpretability* as one institutional element of particular importance. An algorithmic system can be defined as interpretable when it succeeds in explaining its inner processes to human actors. Hence, an interpretable system ensures that humans are able to understand how it arrives at a certain outcome [22]. Concerns about interpretability are currently on the forefront of the social and political debate. This has also been recognized by policymakers, as illustrated for example by the EU's new *General Data Protection Regulation* that establishes the right to explanation (Art. 22 GDPR). Due to their inherent complexity, many machine-learning systems are ubiquitous black boxes whose computational processes often cannot be understood by any human actor, let alone those who are affected by their decisions.**LM: add reference** Specifically in the AI and machine learning community, there has been an increasing realization about the value of interpretable systems nested within a transparent institutional environment, **LM: add references** partially driven by renewed interest in auxiliary determinants of importance such as safety, unbiasedness, non-discrimination or social acceptance.[3] However, rigorous research quantifying the effect of interpretability on the nature of human behavior is scarce. The existing literature either deals with the effectiveness of specific explanations or their ethical value [29, 31, 34, 35]. We argue that, while important, this approach falls short in that it ignores the effect of interpretability on actual human behavior in human-agent environments. Yet, as with every regulative design choice, one can only judge its real-life desirability when taking into account how it will manifest itself in economically and socially relevant outcomes. Without empirical evidence derived from e.g. carefully designed human experiments, one cannot be sure about the theoretically ambiguous implications of increased interpretability. Therefore, we contrast human-agent interactions involving either an opaque or an interpretable ADM system. Our experimental setup addresses the question whether increased interpretability spurs **LM: hier schon mit hypothese/richtung arbeiten?** the utilization of decision-support systems and thus leads to economic gains. Similarly, we ask if human actors are more willing to collaborate with autonomous algorithmic agents when they have more information about them, and if these potential behavioral changes are reflected in economic outcomes. In general, we imagine two different lines of reasoning on subsequent welfare effects.

On the one hand, it seems intuitive that the availability of information and the capacity to understand how an ADM system arrives at a certain solution - i.e. increasing interpretability - would positively affect human behavior. For one, being informed about the concrete parameters of a decision-making situation reduces uncertainty and potentially induces strategic behavior. Moreover, demystifying opaque agents could change the perception of ADM systems - e.g. increasing the tendency to attribute human characteristics to these non-human agents - and thereby influence economic behavior. Other important behavioral variables such as increased trust might spur economic interactions and induce welfare gains. Indeed, first research from Yeomans et al. [76] suggests that humans tend to distrust recommender systems in subjective domains like humor, partially because machine recommendations are harder to understand than human recommendations. Endowing subjects with explanations of the system can, however, mitigate this distrust and subsequently increase human preferences for the system.

On the other hand, it seems plausible that providing additional information might conversely result in negative effects like an erosion of trust. Imagine a system supporting a physician (expert) in diagnosing a patient's (consumer) MRI scan. The physician might generally trust the system based on positive experience and presumptions about its superiority; thus reaching higher accuracy

[2] There was no recipient of computer income, fundamentally changing its role compared to human income. Subjects played in groups of six, which is too small to guarantee anonymity. Although the authors counterbalanced their within-subject design, they did not provide any information about the effect of game and role order, nor did they conduct an analysis that was limited to clean ultimatum game treatment comparisons.

[3] For more on the importance of interpretability, see **insert references**.

in the diagnosis. In contrast, learning about unfamiliar or unexpected features used by the agent might cause distrust and has the physician stick to their own assessment. Consequently, increased interpretability could diminish the efficiency of economically vital consumer-expert interactions. And while decreasing opaqueness might increase human attribution of anthropomorphic characteristics, it could also highlight certain traits that appear inherently mechanical, thereby making the difference between an agent and a human more salient. In turn, human decision-makers might be primed to further differentiate between agents and humans, which could weaken the influence of efficiency-enhancing social norms and concerns. (Insert reference ref to people are too lazy to differentiate between comp and human -> now primed to see decision-maker as computer -> efficiency losses possible) **LM: and example how this turns into negative outcomes**)

Recently, there have been attempts to standardize the current, rather arbitrary approach towards the evaluation of interpretability. Particularly, Doshi-Velez and Kim [22] lay out a taxonomy for different evaluation approaches. Although they are mostly focused on evaluating the quality of different types of explanations, their categorizations can also be applied to a wider range of issues. The authors present three different evaluative approaches: application-grounded, human-grounded and functionally grounded. Embedding our study into this taxonomy, we perform our analysis on the general user level, i.e. use a human-grounded approach. Therefore, our claim is not a specific one, but a generalizable and more abstract one. In line with the majority of economic models, we aspire to ultimately identify replicable behavioral patterns that capture some essence of human decision-making and can thus be applied to a wider range of regulatory issues. Eventually, this will allow us to construct suitable behavioral models that capture the uniqueness of human-agent environments and thus allow for accurate policy predictions as well as satisfactory design decisions.

Our experimental design comprises eight treatments relating to and building on the standard one-shot ultimatum game. The first one simply replicates the traditional ultimatum game in an online setting and is extended by a psychological questionnaire. Treatments two to five expand on this basic setup to include our decision-support system and induce variability in the interpretability of the system. Treatments six, seven and eight substitute the human proposer with an ADM system based on machine-learning that is playing autonomously on behalf of a passive human proposer.

1.0.1 Key results. We find nothing. **LM: the paper is structured as follows**

2 EXPERIMENTAL DESIGN

Ultimatum bargaining is one of the most prominent games researched in experimental economics [38]. Although the game setting seems quite simple, understanding behavior in this framework remains complex even after decades of research and there is a rich literature allowing us to integrate, evaluate and benchmark the relevance of our findings [39]. It has been applied to a variety of issues such as culture [41, 42], gender [33], child development [40] or human-computer interaction [67]. Importantly, it might be the most transparent tool to demonstrate the importance of social norms, psychology and emotions in real-life negotiations [66, 72].

2.1 General Properties of the 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 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.[4] 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 [10, 61]. These findings have led to influential theoretical work integrating other-regarding preferences such as fairness concerns into the traditional *homo oeconomicus* [8, 27].

2.2 The ADM system(s)

Our ADM system takes up one of two roles. It is either implemented as a decision-support system for the human proposer or takes up the role of an autonomous agent proposer. We feel that this case is more relevant in current market interactions than substituting the responders. For instance, consider pricing algorithms that offer customized prices by processing the available data such as characteristics and former behavior of the customer, time of purchase, device used by the customer and so forth **LM: (insert reference)**. Basically, the proposer in the ultimatum game takes the role of such a pricing algorithm, as the algorithm decides on the distribution of the company's surplus from selling a specific product.

2.2.1 Measurement of psychological traits. Across all treatments, participants filled out a 30-item questionnaire taken from the HEXACO-60 personality inventory [4].[5] The data was subsequently used to train a sophisticated and accurate algorithmic model. We used the 10-items for the dimensions *Honesty-Humility*, *Extraversion* and *Agreeableness*. The HEXACO personality model has been successful in predicting and explaining human behavior in a variety of contexts [5]. In particular, *Honesty-Humility* has been shown to explain unselfish behavior in dictator and ultimatum games [46, 48, 78], cooperative behavior in the prisoner's dilemma [77],

[4] 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 $\tilde{y}$ 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.

[5] We omitted 30 questions referring to the dimensions *Emotionality*, *Conscientiousness* and *Openness to Experience* since they are not relevant to our study.

fairness preferences in a redistribution paradigm [45] and unconditional contributions in a public goods game [47].

As a primary driver of prosocial behavior, *Agreeableness* has long been associated with benign behavior in bargaining contexts [6, 51, 79], e.g. predicting higher acceptance of unfair offers in the ultimatum game [48, 71] and less retaliation after prior exploitation [44].

Evidence on *Extraversion* is much more ambiguous, although some studies suggest a link to reward sensitivity and strategic behavior [7, 9, 68]. Overall, we aimed to substantially increase the predictive power of our model and thus allow for meaningful machine-learning, replicating a situation that can realistically approximate potential consequences of decision-support systems and ADM systems for market efficiencies and income distributions.

2.2.2 Programming. [Detailed description of our models and algorithms, insert input from Hannover]

Since our ADM system is based on machine-learning and processes information to maximize the number of successful interactions at the benefit of the proposer, its decisions and predictions might be substantially different from theoretical predictions. That is because theoretical predictions rely on the assumption of perfect rationality by the responder. It is more likely to expect the ADM system to apply a strategy that reflects empirically optimal offers, i.e. a share of about 40 to 50%. Obviously, it is crucial which data is used to train the ADM system. If human-agent interactions in ultimatum bargaining are different from human-human interactions, than exclusively learning from the latter might not fully depict the long-term consequences of implementing ADM systems. As the learning of the ADM evolves, there might be substantial adjustments to responder behavior. We therefore apply a two-step procedure to train our ADM system. First, we run the extended traditional ultimatum game times and feed the results of human-human interactions to our machine learning algorithm. Second, we run another treatment featuring the ADM system as trained by the results from human-agent interactions (see table 1).

- discuss issue of model quality, endogeneity of treatment manipulation, potential need to segment observations
- test performance of algorithm, does it increase expected income?
- pretest of model?

2.3 Interpretability

[Write after we know how the explanations look like]

2.4 Overview Treatments

Table 1 summarizes the first set of treatments of our experimental design. All treatments are variants of the traditional one-shot ultimatum game and designed to provide clean counterfactuals that relate to our hypotheses. Please note, subjects in the pure ADM treatments were still assigned to the roles of either proposer or responder. [6] We employed a between-subject design where all subjects could only participate once and anonymity was guaranteed.

In T_0, the proposer received \$2 and subsequently posed an offer on how to split this money with the responder. The responder indicated the minimum amount of money he or she was willing to accept from the proposer, i.e. the *minimum offer*. If the split offered was at least as high as the minimum offer, the money was allocated according to the proposer's offer. If the proposer offered less than the minimum amount the responder was willing to accept, both players did not earn any money. All treatments followed this basic setup.

In $T_{1.0}$, the proposer could inquire a decision-support system about the likelihood that a specific offer will be accepted by the respective responder. There was no limit for the number of inquiries, but every proposer had to decide on one final offer. Once again, the responder indicated the minimum offer he or she was willing to accept but did not learn about the existence of the decision-support system. Treatment $T_{1.1}$ added the thorough explanations of the algorithm's working mechanism to the proposer instructions and supplementary feedback on their inquiries. Treatment $T_{1.2}$ and $T_{1.3}$ informed responders in the instructions that their respective proposer had the possibility to use a decision-support system. In $T_{1.3}$, we further added the thorough explanations of the algorithm's processes to the responder instructions.

In $T_{2.0}$, we substituted the human proposer with an ADM system that made an autonomous decision on their behalf and was trained on the human-human interactions from T_0. Responders were informed that they would play with an autonomous agent whose surplus went to a passive human observer. In order to correct for and adapt to potential idiosyncrasies of the human-agent environment, we then used the interactions from $T_{2.0}$ to retrain our algorithm. Thus, $T_{2.1}$ captured differences between human-human and human-agent bargaining interactions.$T_{2.2}$ added explanations on the algorithm's working mechanism to the responder instructions.

2.5 Procedure

We recruited subjects via the crowdsourcing platform *figure eight*. In using an online environment, we accepted to cede some control over external subject stimuli in order to access a large and heterogeneous subject pool [2, 30]. Research on the replicability of laboratory experiments online is mixed, although a number of recent studies suggest either no or only small differences between both methodologies [1, 13, 49]. Throughout the experiment, we relied on neutral language and neutral design choices. Figures **x to y** in the appendix depict all relevant screens subjects saw throughout the experiment.

Subjects enrolled to the experiment on their own accord and received a base payment of \$0.5 on completion. We decided to include detailed instructions and two examples designed to clarify the

[6] Otherwise, introducing ADM would have changed both the proposer and the recipient of the proposer's income, which would be the experimenter without this procedure. In that case, treatment effects could not be unambiguously assigned to the replacement of a human decision-maker as sharing income with the experimenter might substantially alter the context of the task.

treatment	proposer			responder		interpretability		trained by	subjects
	role	DSS	agent	role	DSS_info	proposer	responder		
T_0	active	-	-	active	-	-	-	-	188
$T_{1.0}$	active	yes	-	active	-	-	-	1	-
$T_{1.1}$	active	yes	-	active	-	yes	-	1	-
$T_{1.2}$	active	yes	-	active	yes	yes	-	1	-
$T_{1.3}$	active	yes	-	active	yes	yes	yes	1	-
$T_{2.0}$	passive	-	yes	active	-	-	-	1	-
$T_{2.1}$	passive	-	yes	active	-	-	-	6	-
$T_{2.2}$	passive	-	yes	active	-	-	yes	6	-
total									$\sim$ 1600

Table 1: Overview treatments in one-shot game setting

rules of the game while minimizing any confounding anchoring effects. Afterwards, participants answered a number of demographic questions and three control questions.

Following wrong answers, participants were shown a short explanation why their solution was incorrect. Participants were only allowed to proceed after answering all control questions correctly, and those who failed to answer a question correctly more than once were expelled from the experiment (N=?). Participants then followed a link to either the responder or proposer decision screen. Both screens had a drop-down menu with a total of 41 options on how to split the $2 (5 cent steps). After finalizing their bargaining decision, subjects received a code that forwarded them to the HEXACO questionnaire. In between, we inserted an attention check where subjects were simply asked to select the word "BALL" from a drop-down menu. Those who failed to pass the test were dropped from the analysis (N=?). After finalizing the 30-item questionnaire, subjects were released and paid out according to their and their partner's decisions in the bargaining experiment. Participants who gave the same answer at least X consecutive times during the questionnaire were dropped from the subsequent statistical analysis (N=?). Additionally, $T_{1.0}$ to $T_{1.3}$ and $T_{2.2}$ included a manipulation check measuring the extent to which subjects felt they understood how the algorithmic system worked.[7]

We conducted treatments in numerical order. To avoid any confounding effects due to the specific order of treatments, we kept the time between sessions as short as possible.

Proposers earned Dollar x and responders Dollar x on average. The experiment lasted about x minutes on average with very little variation within or between treatments. Table 2 summarizes the demographics of the participants. Participants were paid via the internal procedures on *figure eight*.

2.6 (Preliminary) Hypotheses

[Possibly: Run pretest or simulations of machine-learning model to validate its prediction accuracy. If we can show that our ML-model is accurate enough, we don't need any assumptions.]
Our hypotheses are predicated on the following two assumptions:

(1) If *all proposers fully exploit the decision-support system* and consequently *implement the option with the highest expected income, overall income increases* compared to traditional human-human settings.
(2) The autonomous agent is *at least as good in predicting responder behavior as human proposers.* **LM: warum ist es hier nur ein größer gleich. Ich würde die beiden Annahmen gerne symmmetrisch formulieren**

These assumptions are necessary to make sensible proposition about income and income distribution. A decision-support system that provides, on average, harmful advice, would have very different implications for the success of ultimatum bargaining interactions and presumably very low relevance for the real world. It would also change the focus of this paper to questions on human overreliance on decision-aids rather than their effective utilization, the consequences for social norms and the role of interpretability in fostering (dis-)trust. Similarly, an algorithmic agent that has worse prediction accuracy than the average human proposer would probably not be implemented in a real negotiation setting. ((We regard these assumptions as very weak, since simple linear statistical models that omit human outliers already improve on average human prediction accuracy.))

To address concerns about the endogeneity of algorithmic prediction quality in our experiment, we partition our data to separately compare those who received advice that, if followed, would have led to a successful negotiation and those who received advice that would have led to an unsuccessful one.

In $T_{1.0}$, when endowed with a decision-support system, subjects should be inclined to use it for two reasons. First, simply providing the option will motivate some out of curiosity. Second, decisions in the ultimatum game are naturally characterized by high uncertainty, which makes the system an inherently useful tool to potentially increase one's income as uncertainty can be substantially reduced. We expect some proposers to largely ignore algorithmic advice, and some to incorporate it into their decision, as e.g. shown by Yeomans et al. [76]. If the system is more accurate than the average human proposer in predicting responder behavior, following its advice will increase the likelihood of successful interactions. **LM: maybe insert footnote pointing to the need to build types for proposers to be analyzed separately**

Hypothesis 1a. Proposers in $T_{1.0}$ decide to probe their decision-support system.

[7] Following the procedure of Yeomans et al. [76], we asked them to state their agreement for two statements: "I could understand why the system thought ....

demographics here

Table 2: Overview demographics

Hypothesis 1b. Compared to T_0, rejection rates in $T_{1.0}$ decrease and overall income increases.

Since the algorithm is programmed to maximize interactions at the benefit of the proposer, we also expect proposer income to increase relative to responder income, skewing the income distribution in favor of the algorithmically supported player.

Hypothesis 1c. Proposers in $T_{1.0}$ earn, on average, a larger share of the $2 than proposers in T_0.

However, hypothesis 1a depends on the assumption that human proposers not only use the system, but alter their own decisions if necessary **LM: accordingly? und warum hier nochmal die Einschränkung nachgestellt? Würde es auch davor ziehen... LM: Brauchen wir nicht insgesamt noch eine Aussage/Überlegung/Kommentar on risk preferences?**. We suspect trust to play a substantial role in this scenario. Here, interpretability could be crucial, since it allows proposers to understand and validate the system's advice. We further expect that variation in interpretability transfers into behavioral variation. This hypothesis is supported by a few preliminary studies on the interplay of explanations and decision-making [insert references]. Thus, we hypothesize that proposers will not only be more willing to utilize the system when they understand it better, but also put higher trust in its output, thereby improving performance.

Hypothesis 2a. The number of proposers who use the decision-support system increase from $T_{1.0}$ to $T_{1.1}$.

Hypothesis 2b. Compared to $T_{1.0}$, rejection rates in $T_{1.1}$ decrease and overall income increases.

Again, improved prediction accuracy regarding responder behavior should further skew the income distribution in favor of the proposer. As uncertainty decreases, proposers are less inclined to overpay.

Hypothesis 2c. Proposers in $T_{1.1}$ earn, on average, a larger share of the $2 than proposers in $T_{1.0}$.

When responders are made aware that their counterpart has the option to utilize an algorithmic support system, we predict significant changes in reciprocal behavior. As is the case with patients who derogate their physicians, attitudes towards proposers might change negatively. **LM: reference imo nötig, auch wenn oben schon vorhanden** Additionally, the algorithm could be perceived as an unfair advantage. Depending on the affective reaction, this potentially induces higher demands. Some subjects might also be anxious about being subjected to an algorithm predicting their behavior, and thus adapt more erratic behavioral patterns. On the other hand, the utilization of a decision-support system might alleviate fairness requirements for the proposers, if responders attribute less responsibility and intent to unfair offers. In such cases, minimal offers would decrease. In line with the literature on the perception of experts that use decision-support systems, we expect responder minimum offers to increase on average. Moreover, since the algorithm is trained on human-human data, any change in responder behavior potentially decreases its prediction accuracy. This in turn can negatively effect the number of successful interactions.

Hypothesis 3a. Responders in $T_{1.2}$ demand higher minimum offers than in $T_{1.0}$ and $T_{1.1}$.

Hypothesis 3b. Rejection rates increase and overall income decreases from $T_{1.2}$ to $T_{1.1}$.

$T_{1.3}$ adds thorough explanations to the responder instructions. Strategic responders - conditional on available information - might (try to) anticipate the predictions of the system and thus possibly exhibit very stable and common decision patterns. **LM: ist das die logische folge, die stable patterns** If this holds true, prediction accuracy of the decision-support system could increase, depending on the accuracy of responder predictions, which we expect to increase with increased interpretability. Further, explanations highlight the process through which an algorithm arrives at a solution, thereby making its strategic calculus a salient part of the bargaining situation that could deflect from proposer fairness obligations.

Hypothesis 4a. Responders in $T_{1.3}$ show more uniformly distributed and stable minimum offers than in T_0, $T_{1.0}$, $T_{1.1}$ and $T_{1.2}$. **LM: würde t1.1 einfach auch nennen; kann man more uniformly näher bestimmen oder anders ausdrücken: less heterogeneity, higher uniformity**

Hypothesis 4b. Compared to $T_{1.2}$, rejection rates in $T_{1.3}$ decrease and overall income increases.

Human actors have consistently shown a significantly lower tendency to punish and reciprocate decisions from simple, deterministic computers. **LM: reference** As outlined above, these patterns might crucially depend on context-specific perceptions of the computer, e.g. whether they ascribe agency and intentions. In accordance with past research on human-computer interactions and mind perception, subjects might assign ADM agents based on machine-learning more agency than simple algorithms and deterministic systems, thus possibly equating the system with some form of social entity. **LM: sollten wir das vll nicht nochmal in einem treatment testen? oder wäre das nicht sogar ein eigenständiges papier: Perception of machine agents (ML vs. deterministic)** However, the degree to which social factors from human-human interactions translate into the human-agent sphere remains uncertain. Since the current experiment does not entail a (virtual) presence or learning processes, responder behavior could be especially driven by prior assumptions about machine-learning, and hence hard to predict [insert inferences from realized experimental setup later]. We expect responders to differentiate between an algorithmic and a human proposer, which will be reflected in lower social concerns. Since the human-human training data does not capture the idiosyncrasies of these human-agent interactions, we further expect that the ADM system will overpredict responder demands.

Hypothesis 5a. Responders in $T_{2.0}$ accept on average significantly lower proposer offers than in T_0, $T_{1.0}$, $T_{1.1}$, $T_{1.2}$ and $T_{1.3}$.

Hypothesis 5b. The agent overpays, rejection rates decrease and overall income increases from T_0 to $T_{2.0}$.

With the more accurate training data from $T_{2.0}$, prediction accuracy of the ADM system should increase in $T_{2.1}$. The system should be able to translate relevant variation in human behavior into its offers and thus presumably raise proposer income. Hence, we expect an income distribution that is more heavily skewed in favor of the proposer, i.e. a human observer, than in $T_{2.0}$.

Hypothesis 6a. Mean proposer offers decrease from $T_{2.0}$ to $T_{2.1}$.

Hypothesis 6b. Rejection rates and overall income remain constant from $T_{2.0}$ to $T_{2.1}$.

Hypothesis 6c. Proposers in $T_{2.1}$ earn, on average, a larger share of the \$2 than proposers in $T_{2.0}$.

[Rewrite when we know exactly how the explanations look]

The absence of relevant literature makes it difficult to formulate an adequately substantiated hypothesis about the impact of additional explanations as introduced in $T_{2.2}$. The "computers are social actors" theory poses that humans will automatically treat "machines" in a basic social manner as long as they display some social cues [56, 57]. Increasing the salience of the agent's statistical approach could have the opposite effect, in that it increases the perceived difference between the agent and a human being. Then, responders might be inclined to accept lower offers than before, since social factors increasingly lose their significance. Similar to $T_{1.3}$, some responders might also try to predict agent offers, in which case offers would depend on their expectations of how other human actors behave in human-agent interactions. A third possibility is that responders actually do ascribe more intent and accountability to the agent when endowed with the explanations, because they reveal the "intelligence" of our machine-learning based agent. This hinges on the differences in pre-explanation and post-explanation perceptions about machine-learning and artificial intelligence.

Hypothesis 7. [Rewrite when we know how explanations look. Put more thought into it]

3 RESULTS

3.1 Human-Human Ultimatum Bargaining

From 188 observations in *human-human*, 60% were male, 42% from India, 29% from the USA and 20% from Europe. 52% were older than 35 years old. One participant did not complete the attention check and the corresponding pair was dropped from the analysis. There were no suspicious patterns in the HEXACO questionnaire (see also figure X in the appendix).**LM: würde ich bereits oben in der procedure unterbringen und nicht erst hier in den results...sind ja auch keine results**

The results largely replicate the existing literature and thus instill confidence in our experimental procedure. From 94 total responder-proposer negotiations, 65 resulted in a successful interaction whereas 29 pairs (30%) did not come to an agreement. As shown in figure 1, most proposers opted for an equal distribution offer where both parties receive \$1. The mean proposer offer was ~98 cents and thus exceeded the mean minimum offer of ~90 cents.

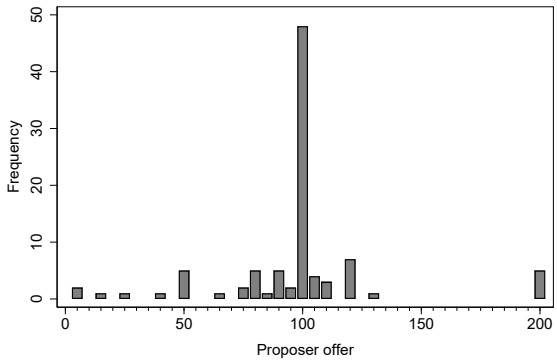

Figure 1: Overview of proposer offers in T_0.

The distribution of responder minimum offers resembles the distribution of proposer offers with a median of \$1. However, there is noticeably more variation around the median, particularly to the left side of the distribution (see figure 2). Proposers regularly offered more than the minimum amount demanded by the responder. Looking at the distribution of responder-proposer discrepancies (figure 3), we identify substantial potential for efficiency improvements. While most pairs reached an agreement and the median is 0, a considerable part of the distribution falls on the right hand side, representing proposer offers that were too small and resulted in a loss of income for both parties.

Even more so, many proposers presumably misjudged the demand position by their responder and hence suffered a possibly avoidable loss. On average, proposers earned ~63 cents compared to the ~75 cents of responders (see also figure X in the appendix).

Result 1. We largely replicate past laboratory results for ultimatum bargaining under the strategy method. Proposers offered on average 49% of their endowment and responders rejected 30% of the offers. We take this as evidence for the validity of the experimental setup.

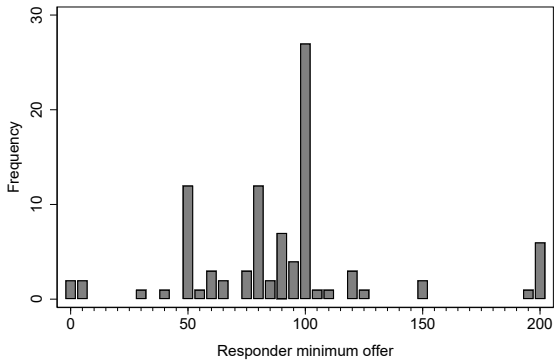

Figure 2: Overview of responder minimum offers in T_0.

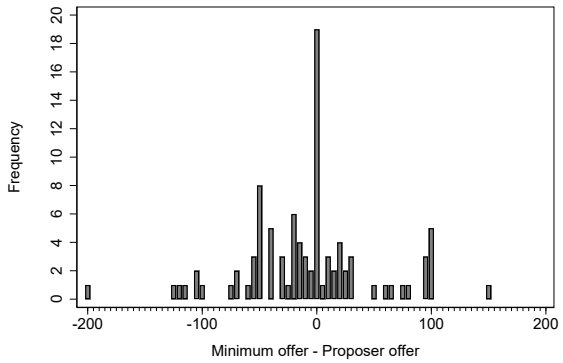

Figure 3: Differences between responder minimum offers and proposer offers in T_0. Positive values represent unsuccessful bargaining interactions. Negative values represent interactions where the proposer offered more than necessary.

After the game, participants answered a 30-item questionnaire that was based on three dimensions from the 60-item HEXACO questionnaire: *Honesty-Humility*, *Extraversion* and *Agreeableness*. The indices show adequate internal consistency and thus appear reliable (Honesty-Humility: Cronbach's α= 0.71; Extraversion: Cronbach's α= 0.86; Agreeableness: Cronbach's α= 0.76). Table 3 and 4 indicate that we largely retained the explanatory power of the three model dimensions. Consistent with the literature, agreeable subjects offered significantly more to responders. We did not find any effects for *Extraversion* or *Honesty-Humility*. Individuals from older age cohorts also appeared to offer less money on average, and the effects were significant for participants between 26 and 45. Regarding minimum offers, we found *Honesty-Humility* to negatively influence responder demands. Thus, individuals who scored high on the *Honesty-Humility* dimension were willing to accept substantially lower proposer offers.

Table 3: Multivariate regression proposer offer

	proposer offer	proposer offer
Honesty Humility	-0.787	0.692
	(6.625)	(7.349)
Extraversion	2.298	2.333
	(4.812)	(5.142)
Agreeableness	14.76**	12.69*
	(6.280)	(6.594)
Male		2.638
		(6.443)
26-35		-32.92***
		(10.90)
36-45		-32.44***
		(11.11)
46-55		-20.25
		(13.71)
56-65		-19.49
		(14.41)
USA		-2.569
		(14.21)
India		1.392
		(14.16)
Europe		-13.37
		(15.15)
Constant	44.75	72.79**
	(30.02)	(34.96)
Observations	93	93
R^2	0.071	0.204

Standard errors in parentheses

The age cohort 18-25 was used as baseline.

* p<0.1, ** p<0.05, *** p<0.01

Table 4: Multivariate regression minimum offer

	minimum offer	minimum offer
Honesty Humility	-13.44*	-14.92*
	(7.588)	(8.657)
Extraversion	7.533	7.878
	(5.920)	(6.657)
Agreeableness	7.123	10.50
	(8.408)	(9.371)
Male		16.15
		(9.843)
26-35		3.075
		(13.60)
36-45		-2.511
		(15.70)
46-55		6.365
		(18.44)
56-65		-6.571
		(25.06)
>65		-18.97
		(34.12)
USA		7.406
		(17.45)
India		-4.468
		(17.76)
Europe		6.196
		(19.03)
Constant	91.33***	73.67*
	(33.87)	(42.22)
Observations	94	94
R^2	0.061	0.109

Standard errors in parentheses

The age cohort 18-25 was used as baseline.

* p<0.1, ** p<0.05, *** p<0.01

4 DISCUSSION

REFERENCES

[1] Ofra Amir, David G Rand, et al. 2012. Economic games on the internet: The effect of $1 stakes. *PloS one* 7, 2 (2012).

[2] Antonio A Arechar, Simon Gächter, and Lucas Molleman. 2018. Conducting interactive experiments online. *Experimental economics* 21, 1 (2018), 99–131.

[3] Hal R. Arkes, Victoria A. Shaffer, and Mitchell A. Medow. 2007. Patients Derogate Physicians Who Use a Computer-Assisted Diagnostic Aid. *Medical Decision Making* 27, 2 (2007), 189–202. https://doi.org/10.1177/0272989X06297391

[4] Michael C Ashton and Kibeom Lee. 2009. The HEXACO–60: A short measure of the major dimensions of personality. *Journal of personality assessment* 91, 4 (2009), 340–345.

[5] Michael C Ashton, Kibeom Lee, and Reinout E De Vries. 2014. The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review* 18, 2 (2014), 139–152.

[6] Anna Baumert, Thomas Schlösser, and Manfred Schmitt. 2014. Economic Games. *European Journal of Psychological Assessment* (2014).

[7] Avner Ben-Ner, Fanmin Kong, and Louis Putterman. 2004. Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving. *Journal of Economic Psychology* 25, 5 (2004), 581–589.

[8] Gary E Bolton and Axel Ockenfels. 2000. ERC: A theory of equity, reciprocity, and competition. *American economic review* 90, 1 (2000), 166–193.

[9] Sean Brocklebank, Gary J Lewis, and Timothy C Bates. 2011. Personality accounts for stable preferences and expectations across a range of simple games. *Personality and Individual Differences* 51, 8 (2011), 881–886.

[10] Colin Camerer. 2003. *Behavioral game theory: Experiments in strategic interaction.* Princeton, NJ: Cambridge University Press.

[11] Colin F Camerer. 2011. *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press.

[12] Alain Cohn, Tobias Gesche, and Michel Andrr Marrchal. 2018. Honesty in the Digital Age. *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3131686

[13] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one* 8, 3 (2013), e57410.

[14] R. M. Dawes, D. Faust, and Paul E. Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674.

[15] Celso de Melo and Jonathan Gratch. 2015. People Show Envy, Not Guilt, when Making Decisions with Machines. *International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China* (2015).

[16] C. M. de Melo, J. Gratch, and P. J. Carnevale. 2015. Humans versus Computers: Impact of Emotion Expressions on People's Decision Making. *IEEE Transactions on Affective Computing* 6, 2 (April 2015), 127–136. https://doi.org/10.1109/TAFFC.2014.2332471

[17] Celso M de Melo, Jonathan Gratch, and Peter J Carnevale. 2015. Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing* 6, 2 (2015), 127–136.

[18] Celso M. de Melo, Stacy Marsella, and Jonathan Gratch. 2018. Social decisions and fairness change when people's interests are represented by autonomous agents. *Autonomous Agents and Multi-Agent Systems* 32, 1 (01 Jan 2018), 163–187.

[19] Berkeley Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126.

[20] Berkeley Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64, 3 (2018), 1155–1170.

[21] Jaap J. Dijkstra, Wim B. G. Liebrand, and Ellen Timminga. 1998. Persuasiveness of expert systems. *Behaviour & Information Technology* 17, 3 (1998), 155–163. https://doi.org/10.1080/014492998119526

[22] Finale Doshi-Velez and Been Kim. [n. d.]. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* ([n. d.]).

[23] Barnaby Dunn, Davy Evans, Dasha Makarova, Josh White, and Luke Clark. 2012. Gut feelings and the reaction to perceived inequity: The interplay between bodily responses, regulation, and perception shapes the rejection of unfair offers on the ultimatum game. *Cognitive, Affective, & Behavioral Neuroscience* 12 (2012), 419–429.

[24] Joseph Eastwood, Brent Snook, and Kirk Luther. 2012. What People Want From Their Professionals: Attitudes Toward Decision-making Strategies. *Journal of Behavioral Decision Making* 25, 5 (2012), 458–468. https://doi.org/10.1002/bdm.741

[25] Mica R Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human factors* 59, 1 (2017), 5–27.

[26] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.

[27] Ernst Fehr and Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114, 3 (1999), 817–868.

[28] Ernst Fehr and Klaus M Schmidt. 2006. The economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the economics of giving, altruism and reciprocity* 1 (2006), 615–691.

[29] M Gacto, R Alcala, and F Herrera. 2011. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences* 181 (2011), 4340–4360.

[30] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1631–1640.

[31] S Garcia, A Fernandez, J Luengo, and F Herrera. 2009. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing* 13 (2009).

[32] Edward Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. *American Economic Review* 106, 5 (2016), 114–118.

[33] Binglin Gong and Chun-Lei Yang. 2012. Gender differences in risk attitudes: Field experiments on the matrilineal Mosuo and the patriarchal Yi. *Journal of economic behavior & organization* 83, 1 (2012), 59–65.

[34] Bryce Goodman and Seth Flaxman. [n. d.]. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813* ([n. d.]).

[35] Paul Goodwin and Robert Fildes. 1999. Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making* 12, 1 (1999), 37–53.

[36] William M. Grove and Martin Lloyd. 2006. Meehl's Contribution to Clinical Versus Statistical Prediction. *Journal of Abnormal Psychology* 115, 2 (2006), 192–194.

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

[38] Werner Gueth, Rolf Schmittberger, and Bernd Schwarze. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization* 3, 4 (1982), 367–388.

[39] Werner Güth and Martin G Kocher. 2014. More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization* 108 (2014), 396–409.

[40] William Harbaugh, Kate Krause, and Steven Liday. 2003. Bargaining by children. (2003).

[41] Joseph Henrich. 2000. Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review* 90, 4 (2000), 973–979.

[42] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, et al. 2005. "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and brain sciences* 28, 6 (2005), 795–815.

[43] Scott Highhouse. 2008. Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology* 1, 3 (2008), 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

[44] Benjamin E Hilbig, Isabel Thielmann, Sina A Klein, and Felix Henninger. 2016. The two faces of cooperation: On the unique role of HEXACO Agreeableness for forgiveness versus retaliation. *Journal of research in personality* 64 (2016), 69–78.

[45] Benjamin E Hilbig, Isabel Thielmann, Johanna Wührl, and Ingo Zettler. 2015. From Honesty–Humility to fair behavior–Benevolence or a (blind) fairness norm? *Personality and individual differences* 80 (2015), 91–95.

[46] Benjamin E Hilbig and Ingo Zettler. 2009. Pillars of cooperation: Honesty–Humility, social value orientations, and economic behavior. *Journal of Research in Personality* 43, 3 (2009), 516–519.

[47] Benjamin E. Hilbig, Ingo Zettler, and Timo Heydasch. [n. d.]. Personality, Punishment and Public Goods: Strategic Shifts Towards Cooperation as a Matter of Dispositional Honesty–Humility. *European Journal of Personality* 26, 3 ([n. d.]), 245–254. https://doi.org/10.1002/per.830

[48] Benjamin E Hilbig, Ingo Zettler, Felix Leist, and Timo Heydasch. 2013. It takes two: Honesty–Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences* 54, 5 (2013), 598 – 603. https://doi.org/10.1016/j.paid.2012.11.008

[49] John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental economics* 14, 3 (2011), 399–425.

[50] Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. 2008. Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLOS ONE* 3, 7 (07 2008), 1–11. https://doi.org/10.1371/journal.pone.0002597

[51] Jui-Chung Allen Li and Yeh-Chen Chen. 2012. Personality, Affects, and Forgiving Behavior in Games. (2012).

[52] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2018. Algorithm Appreciation: People Prefer Algorithmic To Human Judgment. *Harvard Working Paper 17-086* (2018).

[53] Kevin McCabe, Daniel Houser, Lee Ryan, Vernon Smith, and Theodore Trouard. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences* 98, 20 (2001), 11832–11835. https://doi.org/10.1073/pnas.211415698 arXiv:http://www.pnas.org/content/98/20/11832.full.pdf

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

[55] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153

[56] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[57] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[58] Rob Nelissen and Marcel Zeelenberg. 2009. Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making* 4, 7 (2009), 543–553.

[59] Ziad Obermeyer and Ezekiel J. Emanuel. 2016. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine* 375 (2016), 1216–1219.

[60] Dilek Oenkal, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 2009. The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments. *Journal of Behavioral Decision Making* 22 (2009), 390–409.

[61] Hessel Oosterbeek, Randolph Sloof, and Gijs van de Kuilen. 2004. Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics* 7 (2004), 171–188.

[62] Mauricio Palmeira and Gerri Spassova. 2015. Consumer reactions to professionals who use decision aids. *European Journal of Marketing* 49, 3/4 (2015), 302–326. https://doi.org/10.1108/EJM-07-2013-0390

[63] Andrew Prahl and Lyn van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36 (2017), 691–702.

[64] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

[65] James K. Rilling, David A. Gutman, Thorsten R. Zeh, Giuseppe Pagnoni, Gregory S. Berns, and Clinton D. Kilts. 2002. A Neural Basis for Social Cooperation. *Neuron* 35, 2 (2002), 395 – 405. https://doi.org/10.1016/S0896-6273(02)00755-9

[66] Alvin E Roth, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review* (1991), 1068–1095.

[67] Alan Sanfey, James Rilling, Jessica Aronson, Leigh Nystrom, and Jonathan Cohen. 2003. The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science* 300, 5626 (2003), 1755–1758.

[68] Anouk Scheres and Alan G Sanfey. 2006. Individual differences in decision making: drive and reward responsiveness affect strategic bargaining in economic games. *Behavioral and Brain Functions* 2, 1 (2006), 35.

[69] Victoria A. Shaffer, C. Adam Probst, Edgar C. Merkle, Hal R. Arkes, and Mitchell A. Medow. 2013. Why Do Patients Derogate Physicians Who Use a Computer-Based Diagnostic Support System? *Medical Decision Making* 33, 1 (2013), 108–118. https://doi.org/10.1177/0272989X12453501

[70] Vernon L Smith and James M Walker. 1993. Monetary rewards and decision cost in experimental economics. *Economic Inquiry* 31, 2 (1993), 245–261.

[71] Isabel Thielmann and Benjamin E Hilbig. 2014. Trust in me, trust in you: A social projection account of the link between personality, cooperativeness, and trustworthiness expectations. *Journal of Research in Personality* 50 (2014), 61–65.

[72] Eric van Damme, Kenneth Binmore, Alvin Roth, Larry Samuelson, Eyal Winter, Gary Bolton, Axel Ockenfels, Georg Dufwenberg, Martin und Kirchsteiger, Uri Gneezy, Martin Kocher, Matthias Sutter, Alan Sanfey, Hartmut Kliemt, Reinhard Selten, Rosemarie Nagel, and Ofer Azar. 2014. How Werner Gueth's ultimatum game shaped our understanding of social behavior. *Journal of Economic Behavior & Organization* 108 (2014), 292–318.

[73] Mascha van 't Wout, René S. Kahn, Alan G. Sanfey, and André Aleman. 2006. Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research* 169, 4 (01 Mar 2006), 564–568.

[74] A. Van Wissen, Y. Gal, B. Kamphorst, and M. Dignum. 2012. Human-Agent Teamwork in Dynamic Environments. *Computers in Human Behavior* 28, 1 (2012), 23–33.

[75] Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel M. Wegner. 2010. Causes and consequences of mind perception. *Trends in Cognitive Sciences* 14, 8 (2010), 383 – 388. https://doi.org/10.1016/j.tics.2010.05.006

[76] Mike Yeomans, Anuj K Shah, Sendhil Mullainathan, and Jon Kleinberg. 2016. Making sense of recommendations. *Preprint at http://scholar. harvard. edu/files/sendhil/files/recommenders55_01. pdf* (2016).

[77] Ingo Zettler, Benjamin E Hilbig, and Timo Heydasch. 2013. Two sides of one coin: Honesty–Humility and situational factors mutually shape social dilemma decision making. *Journal of Research in Personality* 47, 4 (2013), 286–295.

[78] Kun Zhao, Eamonn Ferguson, and Luke D Smillie. 2016. Prosocial personality traits differentially predict egalitarianism, generosity, and reciprocity in economic games. *Frontiers in psychology* 7 (2016), 1137.

[79] Kun Zhao and Luke D Smillie. 2015. The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review* 19, 3 (2015), 277–302.

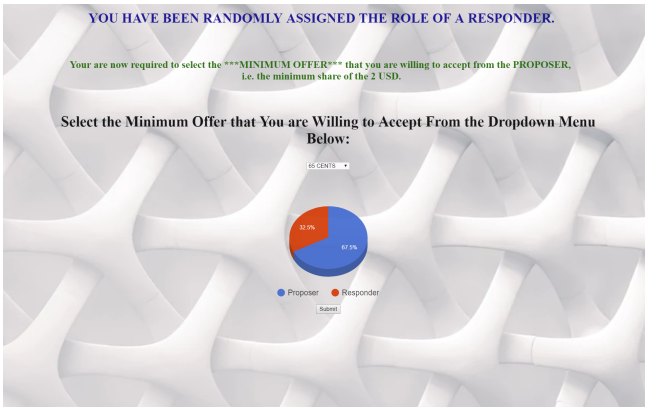

Figure 4: Responder decision screen on Figure8. Subjects could choose any minimum offer from $0 to $2 in steps of 5 cents.

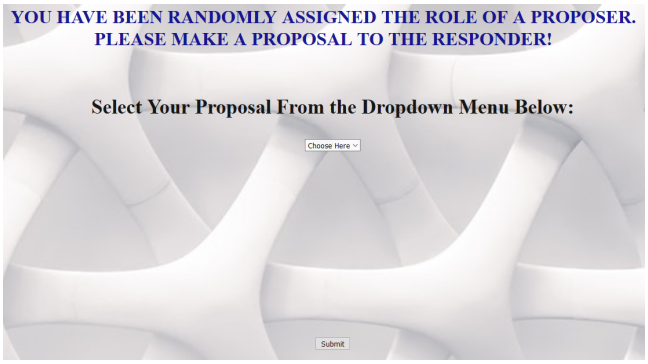

Figure 5: Proposer decision screen on Figure8. Subjects could choose any minimum offer from $0 to $2 in steps of 5 cents.

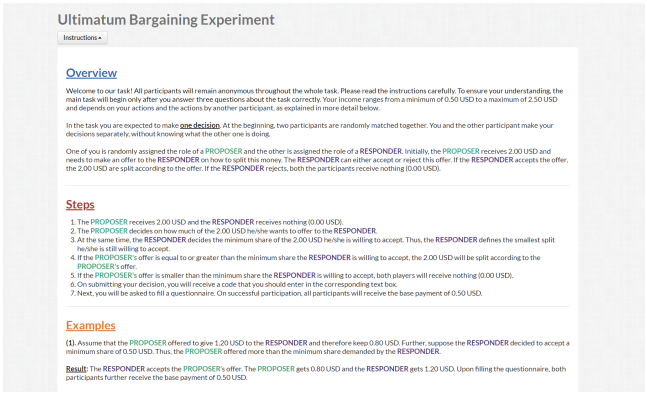

Figure 6: First part of the instruction screen.

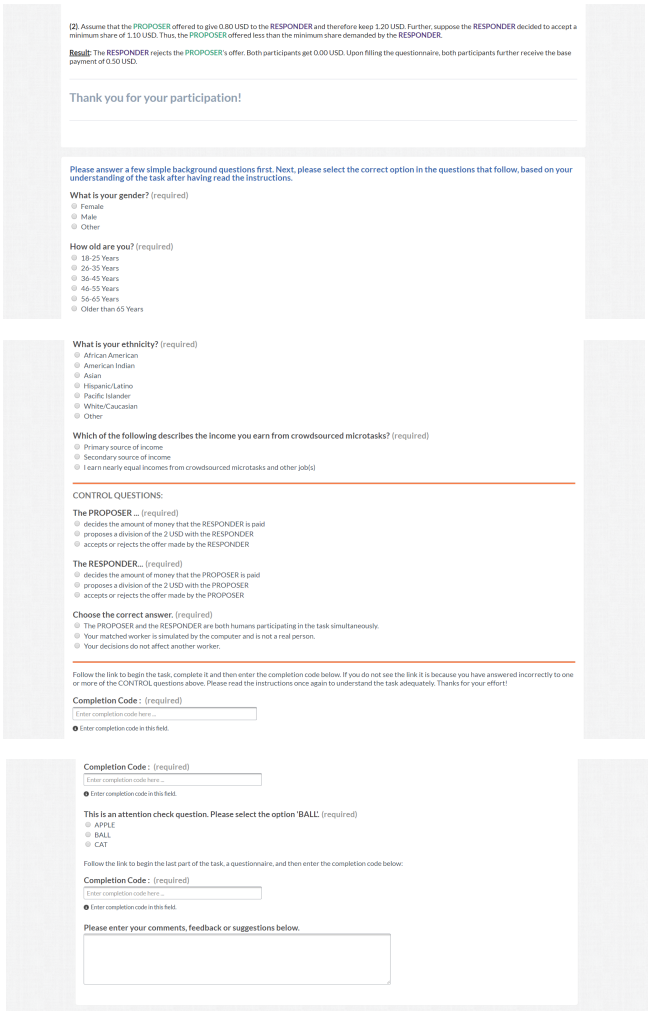

Figure 7: Second part of the instruction screen.

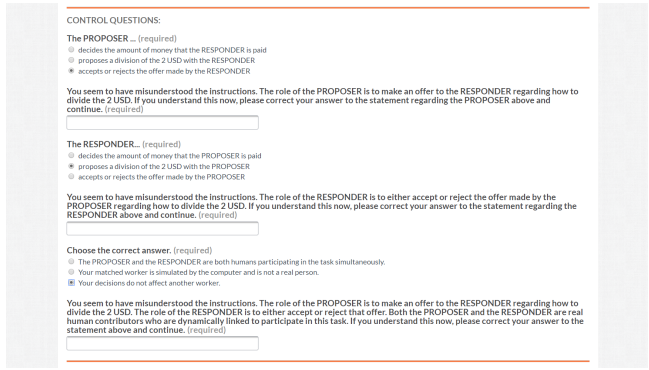

Figure 8: Instruction screen with additional explanations after falsely answering the control questions.

Figure 9: Design of the shortened HEXACO questionnaire.

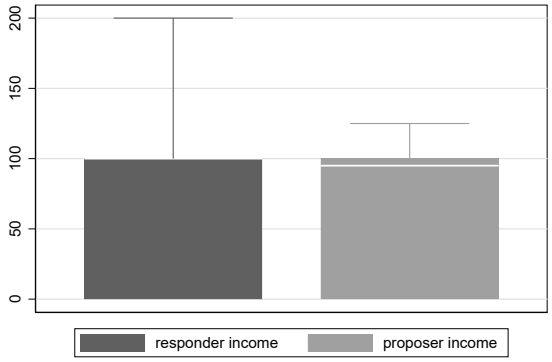

Figure 10: Responder and proposer income in T_0.