Economics Letters 244 (2024) 111979

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Ultimatum bargaining: Algorithms *vs.* Humans

Ali I. Ozkes [a,b,*], Nobuyuki Hanaki [c,d], Dieter Vanderelst [e], Jurgen Willems [f]

[a] *SKEMA Business School, GREDEG, Université Côte d'Azur, France*
[b] *Université Paris-Dauphine, Université PSL, CNRS, LAMSADE, France*
[c] *Institute of Social and Economic Research, Osaka University, Japan*
[d] *University of Limassol, Cyprus*
[e] *Department of Electrical Engineering and Computer Systems, University of Cincinnati, United States of America*
[f] *Institute for Public Management & Governance, WU Vienna University of Economics and Business, Austria*

ARTICLE INFO

Dataset link: osf.io/jx3m4, osf.io/xhq9v

Keywords:
Ultimatum bargaining
Human-ai interaction
Social preferences
Fairness
Equity
Explainability

ABSTRACT

We study human behavior in ultimatum game when interacting with either human or algorithmic opponents. We examine how the type of the AI algorithm (mimicking human behavior, optimising gains, or providing no explanation) and the presence of a human beneficiary affect sending and accepting behaviors. Our experimental data reveal that subjects generally do not differentiate between human and algorithmic opponents, between different algorithms, and between an explained and unexplained algorithm. However, they are more willing to forgo higher payoffs when the algorithm's earnings benefit a human.

1. Introduction

Algorithms increasingly influence decision-making processes in many fields, including finance, healthcare, and justice. As they become more prominent in domains traditionally managed by humans, understanding how people perceive and interact with algorithms is crucial (see, for instance, Capraro et al., 2024, for a review of research on the impact of generative AI). Previous studies indicate that interactions with algorithms can differ significantly from those with humans, depending on how transparent and fair the algorithmic decision-making process appears to be.

We study how behavior in ultimatum game (UG) played against algorithms differ from behavior against human opponents. We focus on (i) if algorithms have different objectives (optimising gains or mimicking human behavior), (ii) whether algorithms provide an explanation, and (iii) if there is a human beneficiary of the algorithm's gains.

UG is one of the commonly used tools in experimental economics, alongside dictator game, to study social preferences, fairness, and equity preferences, among others (Brañas-Garza et al., 2014). In their seminal work, Güth et al. (1982) demonstrated how individuals often reject unfair offers, preferring to receive nothing rather than accept inequity, highlighting the role of fairness in economic decisions as opposed to strict rationality arguments that require minimal sharing and acceptance of virtually anything (see Van Damme et al., 2014; Chaudhuri, 2008, for reviews).

The extent to which algorithms mimic human decision-making or optimise selfishly might affect human trust and cooperation. Furthermore, the role of explainability in algorithms, *i.e.*, whether making an algorithm's decision-making process transparent affects human players' willingness to accept its propositions, is central to designing algorithms that are efficient but also socially compatible and accepted in scenarios where fairness concerns are pivotal.

Recent advances suggest that humans generally behave more rationally when interacting with machines, potentially suppressing emotional reactions that typically influence decisions involving humans (March, 2021; Chugunova and Sele, 2022). Zhang et al. (2022) note that people perceive AI as more likely to make utilitarian choices than humans, potentially reducing market bubbles and increasing bargaining efficiency in auctions. Areas of the brain associated with emotional processing are found to be less active when participants interact with algorithms (Knoch et al., 2006). Yalcin et al. (2022) find that people react less positively when an algorithmic decision-maker makes a favorable decision, but this difference disappears for an unfavorable decision.

Erlei et al. (2022) find that most responders favor human opponents over autonomous agents or humans using AI decision aids, demanding higher compensation to contract with autonomous agents and often overriding economic self-interest to avoid algorithms. Wang et al. (2023) find that rule-driven algorithmic decision-making is found

[*] Corresponding author.
E-mail address: ali.ozkes@skema.edu (A.I. Ozkes).

https://doi.org/10.1016/j.econlet.2024.111979
Received 16 August 2024; Received in revised form 11 September 2024; Accepted 12 September 2024
Available online 21 September 2024

to be more fair by subjects in their experiments compared to data-driven decision-making, indicating that the type of algorithm influences perceived fairness and acceptance. von Schenk et al. (2023) find that subjects display higher social preferences when they knew that a human benefited from machines' decisions.

2. Experiments

2.1. Design and hypotheses

In our UG, the proposer offers an amount between 0 and 100 to the responder, and the responder either accepts (resulting in the split being implemented) or rejects it (resulting in both players receiving nothing). The game is implemented with simultaneous choices so that the proposers indicate what they would propose, and the responders state the minimum offer they would accept (MAO).

Each participant plays three rounds as proposer and three as responder (the order depending on the initial random assignment), interacting with both humans and algorithms, the latter being linear regressions. The algorithms with explanations are designed to either mimic human behavior (Mimicking Algorithm, MA) or optimise their own gains (Optimising Algorithm, OA). A third algorithm provided no explanation for its decision-making (No Explanation Algorithm, NA).[1] Additionally, we distinguish between beneficiaries, where the earnings of the algorithmic player goes either to a randomly chosen subject (Token Player, TP) or to no one (No Receiver, NO).

Our experimental design contains six combinations of algorithm interactions: MA_TP, MA_NO, OA_TP, OA_NO, NA_TP, and NA_NO. Each subject interacts (in both responder and proposer role) once with another human subject (HU), once with the algorithm in TP condition, and once in NO condition. In TP, subjects are told "The earnings of the algorithm at the end of the experiment will be received by a random participant in the experiment". and in the NO condition this was replaced by "The earnings of the algorithm at the end of the experiment will be received by no one".

We developed the following hypotheses derived from theoretical and empirical precedents outlined in previous section:

Responders have

H1: lower MAOs against algorithms compared to humans.
H2: lower MAOs against OA compared to MA.
H3: lower MAOs in TP compared to NO.

Proposers offer

H4: the same amount to MA as humans.
H5: less to OA compared to MA.
H6: less to NA compared to humans.
H7: less to algorithms in NO compared to TP.

2.2. Experimental procedures

We analyze data from gender-balanced U.K. subjects recruited via Prolific on 28–30 August 2023 (pre-screened to include information on age, ethnicity, education, income, employment and relationship status, and online shopping experience). Subjects earned on average 11.79 GBP/h, comprising of 1 GBP participation fee and earnings in UG. Payments are made based on a randomly chosen condition (realised as HU). Median time of the experiment was around 5 min. The experiment was implemented on Qualtrics, and the compulsory comprehension check led to dropping of 388 participants (out of 792 total attempts, this leaves 404, as in Table 1). IRB approval is obtained at Vienna University of Economics and Business.

Table 1
Partition of subjects into experimental conditions.

Condition	First part as proposer	Second part as responder	#
P_MA	HU → MA_TP → MA_NO	HU → MA_TP → MA_NO	49
P_OA	HU → OA_TP → OA_NO	HU → OA_TP → OA_NO	50
P_NA	HU → NA_TP → NA_NO	HU → NA_TP → NA_NO	51
P_OA_HU	OA_NO → OA_TP → HU	OA_NO → OA_TP → HU	50
	First part as responder	Second part as proposer	
R_MA	HU → MA_TP → MA_NO	HU → MA_TP → MA_NO	52
R_OA	HU → OA_TP → OA_NO	HU → OA_TP → OA_NO	51
R_NA	HU → NA_TP → NA_NO	HU → NA_TP → NA_NO	50
R_OA_HU	OA_NO → OA_TP → HU	OA_NO → OA_TP → HU	51

Subjects are randomly assigned to one of the eight conditions in Table 1. Apart from P_OA_HU and R_OA_HU, which only differ in the order of roles, subjects started out playing against human agents in both roles. We compared behaviors between human-first and human-last conditions of OA_HU (e.g., P_OA and P_OA_HU) and as there were no significant differences, we did not run further sessions with reverse orders.[2]

3. Results

3.1. Responder behavior

We find no evidence in support of *H1* in our data: Wilcoxon signed-rank (WSR) test $p-$ values with alternative hypotheses $\mu_{HU}^{MAO} > \mu_{TP}^{MAO}$ and $\mu_{HU}^{MAO} > \mu_{NO}^{MAO}$ are 0.996 and 0.917 respectively (where $\bar{x}_{HU}^{MAO} = 42.89, \bar{x}_{TP}^{MAO} = 44.78$, and $\bar{x}_{NA}^{MAO} = 43.58$). See Fig. 1 for the distribution of choices.

H2 does not find support in data either: subjects' MAOs are not higher against MA compared to OA. The $p-$value in the Wilcoxon–Mann–Whitney test (WMW) with the alternative hypothesis $\mu_{MA}^{MAO} > \mu_{OA}^{MAO}$ is 0.106, where $\bar{x}_{MA}^{MAO} = 44.64$ and $\bar{x}_{OA}^{MAO} = 42.74$.

H3 is also not supported: no significant differences between MAOs in TP and NO conditions. The $p-$ value (WSR) with the alternative hypothesis $\mu_{NO}^{MAO} > \mu_{TP}^{MAO}$ is 0.808. To summarise, *H1–H3* do not find support in data. We observe that subjects' MAOs (i) are not lower if they play against an algorithm and do not depend on (ii) if the algorithm is maximising gains or mimicking humans, (iii) or if there is a human beneficiary behind algorithms or not.

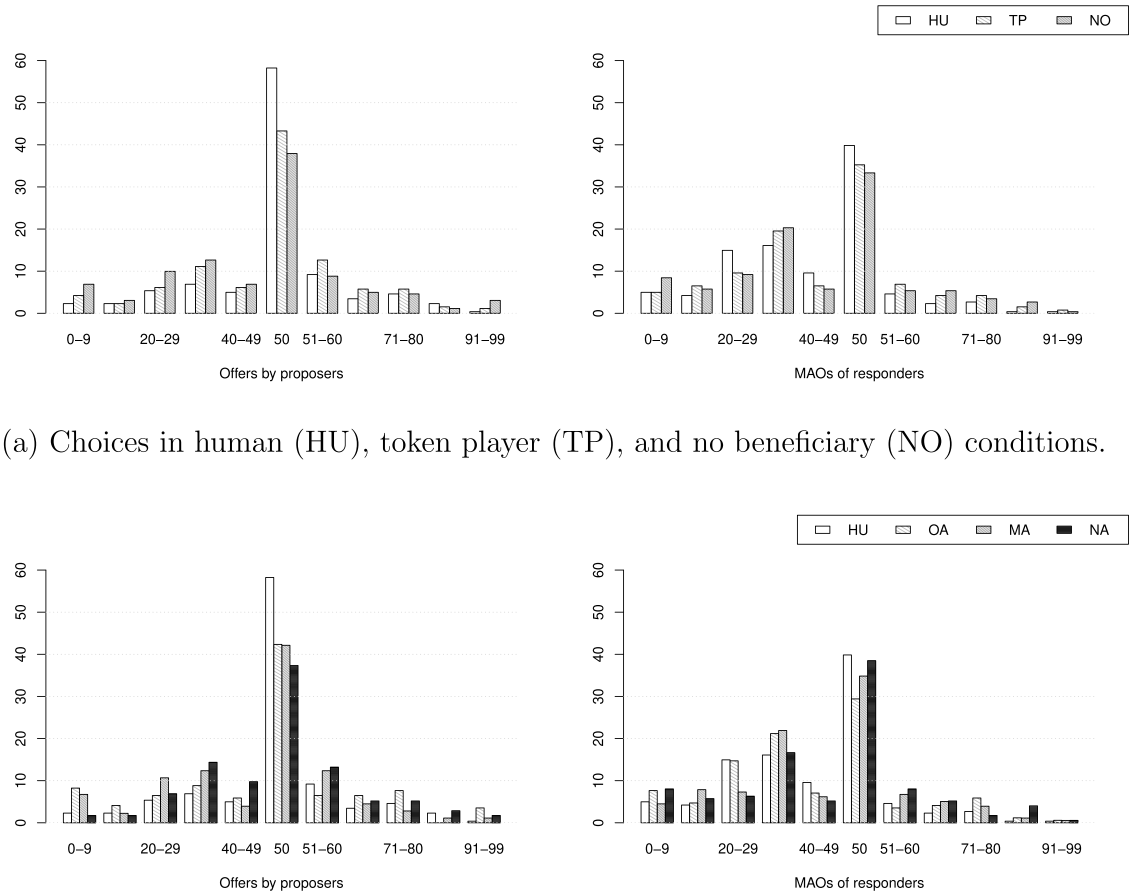
3.2. Proposer behavior

As can be seen in Fig. 1, proposers offer equal split $(50 - 50)$ more to humans (Fisher tests yield $p < 0.05$ for HU *vs.* any algorithmic condition).

We have partial support for *H4* in our data. Subjects send similar amounts to mimicking algorithms with humans if there is a human

[1] The exact workings of the algorithms, which are not disclosed to participants, are based on different applications of linear regressions (*e.g.*, MA proposer predicts what opponent proposed, OA proposer predicts MAO to offer exactly that, both based on demographics and HU data). For MA subjects are told "The other player is an algorithm. The algorithm is trained on data from human interactions. The algorithm's objective in making its decision is to mimic the choices of subjects interacting with other human subjects who have similar characteristics with you.", whereas for OA the underlined part was replaced with "maximize its earnings by taking into account the choices of" and for NA only the first sentence is given. See instructions in Online Appendix.

[2] As there are no statistical differences in (proposer or responder) behaviors between first part and second part (*e.g.*, first part of P_MA and second part of R_MA), we collapse data from different orders in our analysis. Furthermore, data from 52 subjects who made a choice of offering 100 in proposer role are dropped.



(a) Choices in human (HU), token player (TP), and no beneficiary (NO) conditions.

(b) Choices in human (HU) and optimising (OA), mimicking (MA), and no explanation (NA) algorithm conditions.

Fig. 1. Offers and MAOs (% distributions) in human and algorithm conditions.

beneficiary behind the algorithm ($p_{WSR} = 0.953$), however, they send (on average ~4%) less if there is no beneficiary ($p_{WSR} = 0.014$), where $\bar{x}_{HU} = 48.17$, $\bar{x}_{MA\_TP} = 47.77$, and $\bar{x}_{MA\_NO} = 44.34$ (within-subjects).

H5 and *H6* do not find support in data. Subjects do not send less to OA compared to MA ($p_{WMW} = 0.776$, where $\bar{x}_{MA} = 46.06$ and $\bar{x}_{OA} = 47.93$). They also do not send less to NA compared to humans, regardless of if there is a beneficiary ($p_{WSR} = 0.81$ for TP and $p_{WSR} = 0.243$ for NO, where $\bar{x}_{HU} = 50.64$, $\bar{x}_{NA\_TP} = 51.87$, and $\bar{x}_{NA\_NO} = 49.01$, within-subjects).

Finally, we find support for $H7$. Subjects send (on average ~2.5%) less to an algorithm if there is no beneficiary ($p_{WSR} = 0.001$, with $\bar{x}_{TP} = 49.36$ and $\bar{x}_{NO} = 46.90$, within-subjects). In sum, we find that subjects do not offer less (i) to optimising algorithms compared to minimising algorithms and (ii) to algorithms with no explanations compared to humans. We also observe that they send (i) less to an algorithm when there is no beneficiary and (ii) same amount to humans and algorithms with human beneficiaries.

3.3. Further results

Around %40 of subjects prefer to play against humans (whereas ~%20 prefer algorithms), regardless of algorithmic types they interacted with in UG (see Fig. 2, based on responses in the post-experimental questionnaire). There is also a slight preference for algorithms as responders compared to as proposers.

Do subjects who choose human responders think humans accept lower amounts? The answer is no ($p_{WSR} = 0.981$). Also, subjects who choose algorithmic proposers do not think that algorithms offer higher amounts ($p_{WSR} = 0.828$). On the other hand, subjects who choose algorithmic responders think algorithms accept lower amounts ($p_{WSR} = 0.000$) and subjects who choose human proposers believe that humans offer higher amounts ($p_{WSR} = 0.095$). We find that younger subjects choose algorithms more (for proposer role, 39.8 *vs.* 44.1, $p_{WMW} = 0.001$; for responder role, 39.6 vs. 44.1, $p_{WMW} = 0.001$), whereas there is no gender difference in choosing algorithms over humans (both around 40% human, 20% algorithm, and 40% indifferent). Furthermore, education and income levels also do not differ significantly between those who would choose algorithmic and human opponents.

3.4. Discussion

Subjects in our experiment do not differentiate, in general, between human and algorithmic opponents, between different algorithms, and between explained and unexplained algorithms. On the other hand, they forgo higher payoffs when the algorithm's earnings benefit a human. When asked in a post-experimental questionnaire, subjects express double as large interest in interacting with a human as opposed to an algorithm. Our findings hint that although people might prefer interacting with humans over algorithms in strategic interactions, once in interaction they may make similar choices against algorithms and

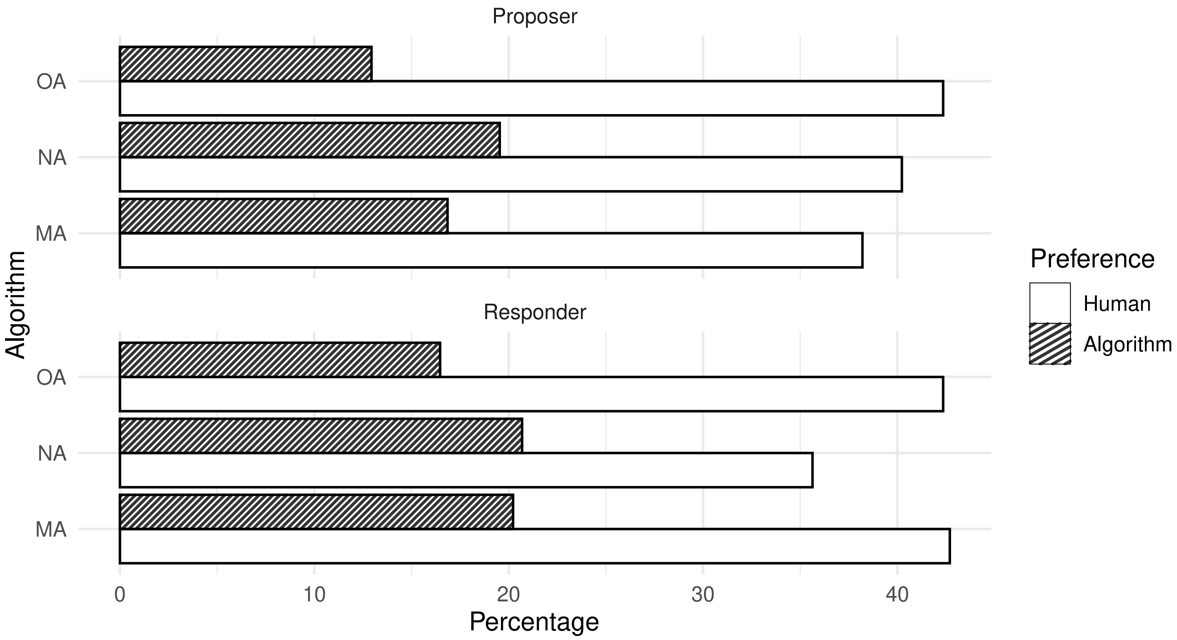


Fig. 2. Preferences for opponents, by algorithmic condition in UG. Remaining answers are indifferent.

humans, and this would be independent of the details of the algorithms' workings and if any explanation is provided.

Data availability

Preregistration and replication materials are available at osf.io/jx3m4 and osf.io/xhq9v.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2024.111979.

References

Brañas-Garza, P., Espín, A.M., Exadaktylos, F., Herrmann, B., 2014. Fair and unfair punishers coexist in the ultimatum game. Sci. Rep. 4 (1), 6025.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., et al., 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. PNAS Nexus 3 (6).

Chaudhuri, A., 2008. Experiments in Economics: Playing Fair with Money. Routledge.

Chugunova, M., Sele, D., 2022. We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. J. Behav. Exp. Econ. 99, 101897.

Erlei, A., Das, R., Meub, L., Anand, A., Gadiraju, U., 2022. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with AI systems. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–18.

Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3 (4), 367–388.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., Fehr, E., 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314 (5800), 829–832.

March, C., 2021. Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. J. Econ. Psychol. 87, 102426.

von Schenk, A., Klockmann, V., Köbis, N., 2023. Social preferences toward humans and machines: A systematic experiment on the role of machine payoffs. Perspect. Psychol. Sci. 17456916231194949.

Van Damme, E., Binmore, K.G., Roth, A.E., Samuelson, L., et al., 2014. How werner Güth's ultimatum game shaped our understanding of social behavior. J. Econ. Behav. Org. 108, 292–318.

Wang, G., Guo, Y., Zhang, W., Xie, S., Chen, Q., 2023. What type of algorithm is perceived as fairer and more acceptable? A comparative analysis of rule-driven versus data-driven algorithmic decision-making in public affairs. Gov. Inf. Q. 40 (2), 101803.

Yalcin, G., Lim, S., Puntoni, S., van Osselaer, S.M., 2022. Thumbs up or down: Consumer reactions to decisions by algorithms versus humans. J. Mar. Res. 59 (4), 696–717.

Zhang, Z., Chen, Z., Xu, L., 2022. Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. J. Exp. Soc. Psychol. 101, 104327.