Test Name: Social Preferences toward AI: Are people willing to share earnings with AI?

Summary: This experiment aims to test people's social preferences toward AI. It proposes to test it by putting human agents and AI agents into a production phase first, and then putting them into a distribution phase where humans can share earnings with their AI counterparts or not. In the production part, agents are not aware of their partner's identity (it can be both AI or human). Before moving to the distribution part, human participants are asked how much they think their partner has contributed to the production (in percentage). Half the participants are randomly selected and informed about their partner's identity while answering the questionaire. At the distribution phase all human participants are informed about their partner's identity. Distribution is done through normal dictator game.

Motivation: The main motivation behind this experiment is to understand the humans' views on AI's role in production. This experiment will aim to test how much value do humans assign to AI's contributions, how much do humans actually share with AI, how much does it differs from the sharing to humans, and does the sharings differ from the value assigned to its production. Besides the main aim of the design, the design also allows us to check if there are differences between human-human production and human-AI production. Furthermore, the design allows checking if there are differences in the sharing when the other player's identity is revealed beforehand or not, and if performances of AI's and humans have same pattern regarding puzzle difficulty.

Literature Review:

Charness et al. (2023b) in a VoxEU column argue generative AI's involvence can increase worker's productivity. Furthermore, Damioli et al. (2021) have found that patenting AI activities have an extra positive effect on labor productivity. Also, Dafoe et al. (2021) argue prioritizing the development of cooperative intelligence in AI, and its complementary skills is the path researchers should take to enhance AI's value. They argue cooperative AI can help solving complex real-world problems. As previous research shows the usage of AI could reshape industries and the world, and it shapes some of them already. Thus inspecting the usage of AI, the roles it can take, and the productivity it can bring are very important and recent research topics. This involvement of AI into production raises a question that how AI's contribution will be evaluated and how willing humans are to share their earnings with AI.

About AI's usage in economic experiments, Christoph March (2019) gathers around more than a hundred papers written about computer players in experiments and

analyses them. In none of the papers computer players acted as an agent in production. In my further research I haven't found any experimental paper that has computer players or artificial intelligence players as a single agent in a production role but I have found one paper that had used AI as a complementary tool in the production process. Noy and Zhang (2023) designed an experiment where one group of participants had access to ChatGPT and the others hadn't. The task was writing, and they have found the participants who had access to ChatGPT were more productive than the participants with no access. They have also found that inequality has been decreased as participants with weaker skills had more benefits from AI usage. However, social preferences toward AI as a single agent are yet to be tested.

Von Schenk et al. (2023), in their systematic review of papers where a distribution with machine is made, found that machine payoffs and shared information about those payoffs differ significantly across studies. They compare payoff informations and their effects on the distribution with machine, and by that they have found that, in general when there is a "human behind the machine", humans are more sharing toward machines, but when no people involved, sharing toward robots decreases. Moreover, this study, with the studies it inspects, shows that sharing with robots occurs even when it is known that the money sent will be "burnt". This experiment design differs from those papers as it will be the first paper that tests social preferences when AI is included in direct production.

Nielsen et al. (2021) argue humans act prosocially when the framework has been found on "computer are social actors". They argue humans perceive machines as somewhat human, applying social categories into them, and having emotions toward machine. Furthermore, a much earlier study on the topic by Nass et al. (1994) argues computers are social actors. Thus, considering AI in social preferences is appropriate.

Gonzalez et al. (2022) and Fraune et al. (2021) show that while robots are in-group in groups of at least 4 that includes both humans and robots, humans favor humans more to robots. However compared to out-group humans, in group robot are more preferred. Thus when AI-human groups are created, humans may have share earnings with AI even if they share more with humans. Furthermore, Makovi et al. (2023) show that paying money to robots could be an adequate proxy for human-AI interactions. In their experiment they show that humans initially don't see robots as needing or desiring money but they also show that humans believe the robots could be designed in a way to desire money. An example of robots desiring money could be chatbots that are programmed to take payments from human counterparts. This finding sets light on the distribution part of this experiment.

The experiment would be unique as there haven't been any paper that has used AI in production phase as a single agent. Although there have been papers about social preferences toward AI, no paper has real-effort production beforehand. This experiment

design could make us understand how sharings are distributed when AI and humans work in a group but without direct help to each other by solving puzzles. Related to puzzles, Balogun et al. (2024) and Cremer (2007) show that AI could be effective at solving puzzles. Thus, in this experiment solving puzzles to create a value seems an adequate task for AI. As Immorlica et al. (2024) argue, AI should be implemented as an economic agent to better understand their impact on strategic behavior etc. This experiment would be helpful on understanding AI's role as an agent.

Various other papers have contributed to the literature. Some of them are: Charness et al. (2023a), Krafft et al. (2016), Dafoe et al. (2020), Roese and Amir (2009), Gal et al. (2004), Korinek (2024), Bansal et al. (2019), Sade (2019), Siemon (2022).

Design:

2-Phases: Production and Distribution

Production phase:

There are two different groups. Group 1 is human-human, and group 2 is human-AI. Random assignment. Different groups go through the same production phase.

2-player 2-stage puzzle solving: Interactive game. Moving objects to reach the door by player object. Different objects have specific movement patterns. Players are not aware whether their partner is AI or human in this part. An image depicting the game could be found at the end of the file.

2 players are solving puzzles simultaneously. There will be 6 puzzles stages. Players are solving different puzzles but puzzles have similar difficulty at the same stage. At each stage there is a time limit, if the puzzle is solved within the time limit the player's group wins money, if the puzzle is not solved within the time limit nothing is won and that stage ends and the next one starts. If a player can solve the puzzle and the other in the group can't, the player who solved the puzzle still contributes to the group's earnings. How fast the puzzles are solved doesn't matter. After both player's solve their puzzle or timeout, the first stage is completed and the next stage starts. The next stage has the same rules, but puzzles get harder after each stage to see the responses and performance differences in harder tasks. At each stage participants can observe their team's total earnings up to that point. After each stage a question is asked to players about the difficulty of the task. After all stages are completed and time is up, the value has been produced. The total value produced will be equal to the number of puzzles solved by each player in a group (or its multiplication by an arbitrary number). There would be approximately 6 puzzles in total with 5 minutes to solve for each puzzle. The game

number and time are selected to not bore participants in very long games and also to assess performance effectively in few games.

The puzzles will be assigned randomly from a pool of puzzles. There will be a total of 24 puzzles grouped by their difficulty. 4 puzzles for each difficulty level. The difficulty level of puzzles will be evaluated by organizing independent sessions, before experiments start, specificly for letting people (who are not participants in the experiment) solve the puzzles. The success rate of solving the puzzles will alow us to asses their level.

Before moving on to the distribution, participants will be given the info about the team's performance. Then, the participants participate in a questionaire in which they are asked how much they think the other player contributed (in percentage) to the total value they produced. Half the participants are informed about the other player's identity (human or AI?) and the other half is not.

Distribution phase:

Normal dictator game. The other player's identity is shared for all participants in this part. In the human-human matches, the dictator will be chosen randomly but in human-AI matches, the human will be the dictator.

Hypothesis:

1: Human-human production may be less than human-AI production. Moreover, this production difference could increase as the gap between humans and AI could increase as puzzles get harder. These are due to AI performing better at well-structured complex games as shown by Yen et al. (2023). They show humans may have biases in tasks whereas AI could adhere better to algorithm rules. Since human performance could drop more than AI does as tasks get harder. And since time is irrelevant for this experiment (as long as tasks are solved), human-human production might relatively decrease more, as AI is expected to make less mistakes. Thus general productivity might differ from human-human to human-AI.

2: Human-human matches might result in more relative sharing than human-AI matches as humans favor humans, but the difference might not be high as humans will be aware that AI is programmed to earn money. As Makovi et al. (2023) showed that humans can empathize with robots trying to earn money. However as shown by Von Schenk et al. (2023), humans will be more willing to share with other humans in general.

3: Human-AI absolute share might be greater than human-human absolute share if the expected result of AI overperforming human happens. However, if AI overperforms

humans, humans might think more egalitarian in distribution phase as ther could be self-serving bias.

4: Preferences in the group that other player's identity is shared at questionaire will not significantly differ from real distribution percentages in general. However it may differ for the group that didn't have identity info about the other player at questionaire. Sharings may drop if the other player appears to be AI.

Notes:

The payments to AI are explained as: The AI is programmed to earn money for itself. A premium version of AI subscription is used and this has a cost. Also, there could be ethical reasons to share with AI, like AI is entitled to the value it produced. Furthermore, if the share given exceeds the payment made for AI, it will be used in AI development / AI costs.

Participants will be aware that they can be matched with AI or human at the beginning, and AI will be money-seeking if matched.

References:

1. Balogun, G.B., Ibisagba, D., Bajeh, A. Debo T. O., Muyideen A., Peter O. J. (2024, November). Comparative analysis of AI-based search algorithms in solving 8 puzzle problems. Bulletin of the National Research Centre 48, 119. https://doi.org/10.1186/s42269-024-01274-3

2. Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019, October). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7(1), 2-11. https://doi.org/10.1609/hcomp.v7i1.5285

3. Charness, G., Jabarian, B., & List, J. (2023b, October). Scientific experimentation with Generative AI. CEPR VoxEU. https://cepr.org/voxeu/columns/scientific-experimentation-generative-ai

4. Charness, G., Jabarian, B., & List, J. A. (2023a, September). Generation next: Experimentation with ai. NBER. https://www.nber.org/papers/w31679

5. Cremer, D. (2007, June). The application of artificial intelligence to solve a physical puzzle. Indiana University Scholar Works. https://scholarworks.iu.edu/dspace/items/007b8624-7650-433a-a388-1dea746fbcce

6. Dafoe A., Hughes E., Bachrach, Y. Collins T., McKee K. R., Leibo J. Z., Larson K., Graepel, T. (2020 December). Open Problems in Cooperative AI. https://arxiv.org/abs/2012.08630

7. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021, May). Cooperative AI: Machines must learn to find common ground. Nature News. https://www.nature.com/articles/d41586-021-01170-0

8. Damioli, G., Van Roy, V., Vertesy, D. (2021, January). The impact of artificial intelligence on labor productivity. Eurasian Business Review 11, (pp. 1–25). https://doi.org/10.1007/s40821-020-00172-8

9. Fraune, M. R., Šabanović, S., Smith, E. R. (2021, February). Some are more equal than others: Ingroup robots gain some but not all benefits of team membership. Interaction Studies. https://www.jbe-platform.com/content/journals/10.1075/is.18043.fra

10. Gal, Y., Pfeffer, A., Francesca, M., Grosz, B. (2004, January). Learning Social Preferences in Games. Nineteenth National Conference on Artificial Intelligence: July 25–29, 2004, San Jose, California, ed. National Conference on Artificial Intelligence, 226-231. AAAI Press. http://www.aaai.org/Library/AAAI/aaai04contents.php

11. Gonzalez A. C., Fraune M. F., Wullenkord R. (2022, December). Can Moral Rightness (Utilitarian Approach) Outweigh the Ingroup Favoritism Bias in Human-Agent Interaction. In Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22). Association for Computing Machinery, 148–156. https://doi.org/10.1145/3527188.3561930

12. Immorlica, N., Lucier, B., Slivkins A. (2024, May). Generative AI as Economic Agents. SIGecom Exch. https://doi.org/10.1145/3699824.3699832

13. Korinek, A. (2024, December). Generative AI for Economic Research: Use Cases and implications for economists. Journal of Economic Literature. https://www.aeaweb.org/articles?id=10.1257%2Fjel.20231736

14. Krafft, P. M., Macy, M., & Pentland, A. (2016, November 2). Bots as virtual confederates: Design and ethics. arXiv.org. https://arxiv.org/abs/1611.00447

15. Makovi, K., Sargsyan, A., Li, W., Bonnefon J-F., Rahwan T. (2023, May). Trust within human-machine collectives depends on the perceived consensus about cooperative norms. Nature Communications 14, 3108. https://doi.org/10.1038/s41467-023-38592-5

16. March, C. The Behavioral Economics of Artificial Intelligence: Lessons from Experiments with Computer Players (2019, November). CESifo Working Paper No. 7926, http://dx.doi.org/10.2139/ssrn.3485475

17. Nass C., Steuer J., Tauber E. R. (1994, April). Computers are social actors. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI

'94). Association for Computing Machinery. (pp. 72–78). https://doi.org/10.1145/191666.191703

18. Nielsen Y., Pfattheicher S., Keijsers M. (2022,February). Prosocial behavior toward machines. Current Opinion in Psychology (Vol. 43). (pp 260-265). https://doi.org/10.1016/j.copsyc.2021.08.004

19. Noy S., Zhang W. (2023, July). Experimental evidence on the productivity effects of generative artificial intelligence. Science 381, (pp. 187-192). https://doi.org/10.1126/science.adh2586

20. Roese N. J., Amir E. (2009, July). Human—Android Interaction in the Near and Distant Future. Perspectives on Psychological Science. (pp 429-434). https://doi.org/10.1111/j.1745-6924.2009.01150.x

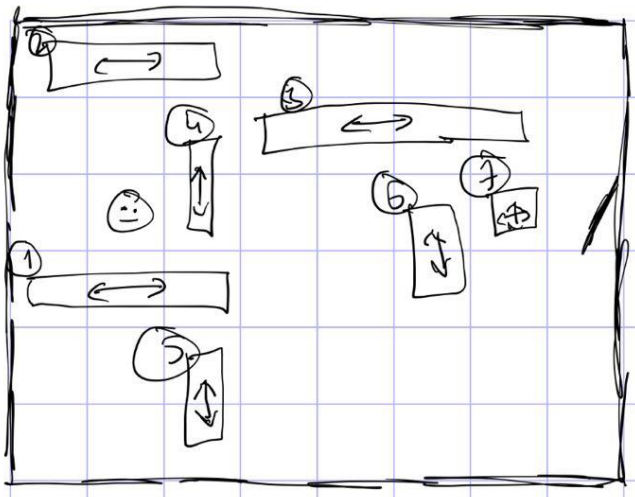
21. Sade, Orly. (2019, May). Robo-Advisor Adoption, Willingness to Pay, and Trust-An Experimental Investigation. http://dx.doi.org/10.13140/RG.2.2.23982.77125

22. Siemon, D. (2022, July) Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. Group Decision and Negotiation 31, 871–912. https://doi.org/10.1007/s10726-022-09792-z

23. von Schenk, A., Klockmann, V., & Köbis, N. (2023). Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. Perspectives on Psychological Science, 0(0). https://doi.org/10.1177/17456916231194949

24. Ren Y., Deng X., Joshi K.D. (2023, August). Unpacking Human and AI Complementarity: Insights from Recent Works. SIGMIS Database 54, 3, 6–10. https://doi.org/10.1145/3614178.3614180

The image depicting the game:



Solution to the game

a) Move 6 down end

b) Move 2 right end

c) Move 4 up end

d) Move 7 down end

e) Move the player to the door

→ The game is solved.

→ There can be much harder puzzles.

→ The Door

→ The Player

1xn objects → Obstacles

1x1 objects can move either way

1xn objects can only move in the axis where its long side rests.

Alternative depiction:

	A	B	C	D	E	F					
1											
2											
3							THE DOOR				
4		THE PLAYER									
5											
6											

Legend:

: Expresses1x1 objects

: Expresses nx1 objects

: Expresses 1xn objects

THE DOOR : Expresses the door

THE PLAYER : Expresses the player