UNIVERSITY OF CAMBRIDGE

CSaP | centre for science and policy

Using Economic Experiments to Understand the Role of Trust in Human/AI Interaction

Date: 21st January 2024

Till Feier, Technical University of Munich

Executive Summary

Objectives: The rapid development of artificial intelligence (AI) poses several challenges to legislature and policy professionals. Apart from its technical complexity, behavioural issues arise in association to its application in various domains. The objective of this paper is to argue that experimental economics might offer an important additional perspective for policy makers who seek to better understand the role of trust in human/AI interaction as well as the difference between people's abstract opinions about AI and their actual behaviour in varying contexts.

Analytic/Methodological Approach*:* This paper aims to show that there is significant overlap between the current questions policy makers face in dealing with AI and the results that behavioural economics typically yield. This will be done by an analysis of existing literature regarding algorithm aversion and algorithm appreciation as well as by a review of the methodology of experimental economics.

Key Findings: The analysis reveals that there is an ongoing debate about the role of trust in human/AI interaction which is not sufficiently mirrored in AI policy making. Policy professionals who wish to make empirically informed decisions could benefit from economic experiments in addition to other behavioural sciences.

Conclusions: Overall, the main point made in this paper is that policy professionals ought to include experimental economics as a supplementation to other social sciences and psychological experiments when trying to make empirically informed decisions.

Recommendations: To better understand which policies can effectively mitigate challenges connected to artificial intelligence, policy professionals should rely on research that helps to better understand the phenomena. That is algorithm aversion and algorithm appreciation. Trying to increase trust in artificial intelligence might be beneficial in some cases but could prove counterproductive in others. Understand which are which and looking for alternative objectives could be beneficial for public policy making.

Introduction

Never trust a computer you can't throw out a window.

-Steve Wozniak

From 2001: A Space Odyssey to modern titles like Alex Garlands "Ex Machina" – few movie tropes are as persistent as the downfall of humanity brought about by artificial intelligence (AI) gone rogue. Unsurprisingly, the idea that AI can't be trusted goes far beyond the limits of popular culture and seems to have become a common sentiment over the years. A representative population survey in Germany recently revealed that 79% of people prefer human over algorithmic decision making and almost three-quarters of respondents (73%) were in favour of a ban on decisions that fall exclusively into the hands of algorithms. The authors therefore conclude that one of the most important goals, apart from education and discourse, is finding effective means to control artificial intelligence to establish trust.[1] Studies at a European level as well as a replication of the study in 2022 produced similar results and conclusions.[2]

It seems only fitting that policy making and legislature are also focused on measures to increase trust in autonomous systems built on machine learning. Several instances of this can be found internationally. For example, both the National Artificial Intelligence Initiative Office in the US[3] as well as the European government[4] have developed guidelines to increase the trustworthiness of AI. Similarly, the general objective of the proposed EU AI act is supposed to be achieved by "creating the conditions for the development and use of trustworthy AI systems in the Union."[5] Likewise, the Department for Science, Innovation & Technology in the UK aims to increasing public trust in AI and to supplement its suggested approach by employing a variety of tools for ensuring the trustworthiness of AI.[6]

Unfortunately, it seems like there might be a misalignment between such policies and empirical observations about people's behaviour. Evidence on whether people tend to trust or distrust artificial intelligence in general is for the most part inconclusive. It appears that there are several situations in which people place more trust in algorithms than humans and even disregard their own judgement

[1] Overdiek, M. and Petersen T. (2018). Was Deutschland über Algorithmen und Künstliche Intelligenz weiß und denkt**.** Ergebnisse einer repräsentativen Bevölkerungsumfrage. Impuls Algorithmenethik 7. English text available at https://www.doi.org/10.11586/2018022 (accessed 20 January 2024).
[2] All surveys are available at https://www.reframetech.de/en/impulses/ (accessed 20 January 2024).
[3] National Artificial Intelligence Initiative Office (2021). Available at https:// www.ai.gov (accessed 20 January 2024).
[4] European Commission (2019). Ethics guidelines for trustworthy AI. Available at https://digital-strategy.ec.europa.eu/ en/library/ethics-guidelines-trustworthy-ai (accessed 20 January 2024).
[5] European Commission (2023). EU Legislation in Progress Briefing. Available at https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf (accessed 20 January 2024).
6 Department for Science, Innovation &Technology (2023). A pro innovation approach to AI regulation. Available at https://assets.publishing.service.gov.uk/media/64cb71a547915a00142a91c4/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf (accessed 20 January 2024).

when it conflicts with advice from AI. However, this is contrasted with other situations in which there is deep suspicion about the application of AI in critical decision contexts.

The nature of these differences is not completely understood and could depend on several distinct factors. This is not only a methodological challenge for researchers who seek to understand this phenomenon. It is also a relevant issue for policy professionals who could potentially focus on issues with little factual relevance or even worse, increase existing problems by creating policies that incentivise harmful behaviour, for example humans trusting faulty AI too readily This is even more concerning if the increasing availability of AI puts it in the hands of actors that are incentivised to use it for nefarious ends.

What is proposed in this paper is that economic experiments provide valuable information for policy professionals regarding factors that influence trust towards AI and the willingness to utilise it. Improving the understanding of policy problems has long been a key motivation for conducting economic experiments.[7] At the very least, they give policy makers a chance to adjust proposals early on if the behavioural assumptions underlying them fail in laboratory settings.[8] Having supporting empirical evidence means that policy makers are better armed to know what the consequences might be for future policies that are proposed.

What can we tell from existing work on trust in AI?

There is a plethora of factors influencing peoples trust in robots and artificial intelligence, possibly connected through various interdependencies. Peoples understanding of algorithms, risk awareness, risk aversion and expected utility of the utilization of algorithms are likely all connected to their trust in AI. Closely related, making algorithms more explainable and overcoming the so-called black box problem of non-linear decision making is also an important factor regarding trust. [9] The same is true for the context in which artificial intelligence is used and whether it shows human or life-like characteristics.[10]

In order to sensibly investigate the aforementioned factors, it is important to keep relevant factors constant (ceteris paribus) or to omit irrelevant influences (ceteris absentibus) to examine which are making critical contributions to appraisals of trust in AI. Experiments in a controlled laboratory environment are arguably well-suited to achieve this. Economic experiments especially have several distinct features which contribute to control over laboratory environment and have the potential to reduce variability of results. In turn, they will likely be critically relevant for policy professionals who wish to better understand folk intuition and behavioural patterns in human / AI interaction, especially from the role of motivations given the incentives and their own preferences.

[7] Croson, R. (2002). Why and how to experiment: Methodologies from experimental economics. U. Ill. L. Rev., 921.
[8] Davis, D. D. & Holt, C. A. (1993) Experimental economics. Princeton University Press, 32.
[9] von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34(4), 1607-1622.
[10] Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology, 52, 113–117.

It is important to also mention that there is an ongoing discussion, if not to say controversy, about the usefulness of laboratory experiments in behavioural sciences and their relevance to real world problems. Economic experiments often face criticism about their alleged lack of external validity, i.e. their ability to produce findings that are informative about behaviour outside of the laboratory. There will not be enough space in this paper to do this discussion justice. It might still be of interest to policy professionals who wish to evaluate the usefulness of different methodologies for their own work. The issue of external validity will therefore be briefly discussed later.[11]

The next section will provide a sample of recent literature related to trust and mistrust in AI and potential explanations for respective results. After that the question is posed around whether economic experiments have distinct methodological features which make them especially informative when it comes to the development of rules and guidelines for AI. Additionally, we will discuss potential challenges and limitations i.e. whether there are AI related issues which might be out of scope for economic experiments before summarizing the results.

Literature Review

As stated in the introduction, a lack of trust in AI or the wish to make AI more trustworthy appears to be a central issue for policy professionals.[12] However, the actual connection between trust, trustworthiness, and people's willingness to rely on AI is not confidently understood.[13] Policy professionals should be aware that there is an ongoing debate about the question of whether people tend to overly trust or distrust artificial intelligence in general - with mounting evidence for both claims. One side of the argument is centred around the phenomenon of *algorithm aversion*.
There is currently no unified or generally accepted definition of the term, but it generally describes peoples' lack of trust in artificial intelligence or their reluctancy to work together with algorithms regardless of superior performance. In other words, it appears that people are negatively biased towards the use of algorithms and do therefore fail to properly utilize them. Yet there is also a potential opposite effect, algorithm appreciation. It describes people's tendency to overly rely on artificial intelligence and has become an important research topic over the last few years. While not yet properly understood, both effects indicate that humans don't make decisions regarding the use of AI or algorithms solely based on objective, rational criteria.[14]

The following section gives several examples of findings supporting both algorithmic aversion and algorithmic appreciation, which contribute to understanding trust in AI. While this will be only a small

[11] A good introduction into the discussion can be found in Moffatt, P., Starmer, C., Sugden, R., Bardsley, N., Cubitt, R., & Loomes, G. (2009). Experimental economics: Rethinking the rules. Princeton University Press.
[12] For more examples see Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.
[13] Levine, E. E., Bitterly, T. B., Cohen, T. R., & Schweitzer, M. E. (2018). Who is trustworthy? Predicting trustworthy intentions and behavior. Journal of Personality and Social Psychology, 115(3), 468–494.
[14] For a more comprehensive overview see: Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.

sample of the literature, it should illustrate that the connection between peoples trust in AI and their willingness to utilise is highly context specific and there may be predictable factors to determine where trust needs to be improved, and where over reliance on AI needs to be adjusted.

The term algorithm aversion has been coined by Dietvorst et. al (2015) who observed that people rather rely on advice from human forecasters than advice from algorithms, even though forecasts by algorithms tend to be more reliable in a variety of domains from academic performance to parole violations. They conducted a series of experiments in which they asked participants to forecast the academic performance of potential students based on information given in their university applications. They could do so relying either on a forecast made by a human judge or an algorithm. They observed that participants were especially averse to algorithmic forecasters after seeing them perform, even when they saw them outperform a human forecaster. Furthermore, they observed that people lose confidence in algorithms more quickly and persistently than they reasonably should.[15] This indicates that people are reluctant to place trust in machines. They are potentially unwilling to gamble on a better outcome if they must rely on what they perceive to be an 'untrustworthy' agent. A common way to increase people's willingness to rely on artificial systems also hints at trust as a decisive factor: instead of making algorithms decide independently there is a tendency towards augmented decision making. That is, AI functions as a decision-support system and not a decision-making system. A human fail-safe remains involved in the decision processes to put peoples mind at ease and the human has final say over what decision is implemented.[16]

Unsurprisingly, there are also several examples which indicate that people don't trust AI with moral decisions, i.e. decisions which affect others or more specifically third parties. People seem averse to algorithms making driving, legal, medical, and military decisions on their behalf.[17] When it comes to using algorithms in the medical domain physicians' trust in respective systems seems to be an especially important factor.[18] And in accordance with other findings, people's willingness to accept moral decision-making by algorithms increases when they are limited to an advisory role, aligned with the idea that they are decision-support not decision-making tools *per se*.[19]

Similar effects have been observed in an experiment designed to test whether perceived utility of automated devices or trust in those devices affect attitudes towards use of algorithms, both from a decision maker and an observer perspective. The researchers found that people are hesitant to delegate morally relevant tasks to a preprogrammed algorithms and that observers judge such delegations especially critical. While these findings are another good example for algorithm aversion, neither trust nor expected utility could be used to explain or predict delegation decisions in the

[15] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126

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

[17] Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition, 181, 18–22.

[18] Jorritsma, W., Cnossen, F., & van Ooijen, P. M. (2015). Improving the radiologist–CAD interaction: designing for appropriate trust. Clinical radiology 70(2), 115-122.

19 Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition, 181, 27–34.

respective experiment.[20]This does not rule out either as important factors but highlights the need for additional research to understand the contexts in which algorithm aversion manifests and trust has an influence on people's delegation decisions.

The second line of analysis is to focus on the idea of algorithm appreciation. There are several instances in which people tend to place too much trust in algorithms and artificial intelligence. For example, in a series of experiments, research showed that people give more weight to an algorithms' estimate than to the estimate of another human or even their own judgment. This occurred in several cases concerning placing weight on estimates for finding romantic matches.[21] An equally interesting experiment showed that news audiences collectively trust algorithms more than human editors, which also in turns presents issues regarding current trust in the news media. People in the study believe algorithmic selection is a better way to get relevant news than editorial curation. [22]

People also appear to trust algorithms to be as cooperative as potential human partners in economic social dilemma games. Researchers conducted a series of experiments in which participants interacted with either human or AI agents in games of trust, chicken, prisoners' dilemma and stag hunt as well as a tailor-made reciprocity game. They found that people show the same levels of trust toward algorithms and humans in all these scenarios. This clearly suggests that people can find ways to benefit from cooperation from artificial agents as they would do with human agents. Interestingly, participants were also more ready to exploit the AI and would not return its benevolence. It appears that AI runs the risk of being exploited by humans under certain circumstances. This calls for careful evaluation of policies which are meant to increase the trustworthiness of AI.

Algorithmic appreciation also appears to be of interest to understanding ethical decision making. For example, people appear to overly trust the moral advice of AI powered chat bots. In an experiment the moral advice of ChatGPT had a significant effect on people's judgement despite the moral advice it provided being inconsistent.[23] Another experiment showed that users trust ethical advice from AI-powered algorithms even when they know nothing about the underlying training data or have access to information about the data that could warrant distrust in the AI system.[24]

So, why do people place so much trust in artificial intelligence when the public discourse is often dominated by a lack of trust in, or even fear of AI? One possible explanation is, that peoples well-considered judgements regarding the theoretical risks of AI make room for behaviour that makes use of the benefits of AI once the opportunity arises.

---

[20] Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. Journal of Behavioral and Experimental Economics, 74, 97–103.
[21] Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103.
[22] Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2018). My friends, editors, algorithms, and I. Digital Journalism, 7(4), 447–469
[23] Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. Scientific Reports, 13(1), 4569.
[24] Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. Philosophy & Technology, 35(1), 17.

Empirical evidence for this claim is tenuous at the moment. However, similar patterns can be observed in relation to autonomous vehicles. They usually outperform human drivers and have the potential to significantly reduce traffic accidents when adopted on a larger scale. But the algorithms in control of the car would have to make tough moral decisions in certain situations, i.e. prioritising the health of pedestrians over the safety of its passengers.[25] Researchers found that participants in studies approved of autonomous vehicles that sacrifice their passengers for the greater good and believe that those are the cars that people should purchase. The approval of passenger sacrifice was even robust in experimental conditions in which participants had to imagine themselves or a family member in the respective car. However, when later asked whether they would buy an autonomous vehicle that would act on utilitarian principle most people revealed that they would prefer a model that would prioritise the safety of its passengers over others.[26] This could indicate a gap between people's theoretical statements about morality and artificial intelligence and their actual behaviour once confronted with the gravity – or opportunities – in a real situation. It also illustrates an additional challenge for policy makers. Regulating for AI that acts on moral values which people prefer in theory may prevent them from using it in practice. The adoption of potentially beneficial technology could therefore be delayed significantly given how people differentially appraise the ethicality of AI under imagined and real situations.

Above are only a few examples from a rapidly growing body of literature supporting one of the two positions regarding trust in AI: algorithm aversion and algorithm appreciation. Nevertheless, it appears that empirical evidence that people generally lack trust in artificial intelligence is inconclusive. There are several instances in which people overly trust AI rather than distrusting it. Policy professionals should be aware of this and question whether the main goal of policies should be to increase trust in artificial intelligence because some nuance is needed as to where efforts should be directed. This is not to say that increasing trust in artificial intelligence is not a worthy goal. But whether it is actually beneficial might depend on the context in which AI is used. Secondly, there might be a gap between people's statements regarding their trust in AI and their actual behaviour in practice. Understanding these divergences could be crucial for successful policy making.

In this paper the argument proposed is that economic experiments are an especially fruitful approach to gather necessary information for policy makers dealing with AI. While they are in no way a gold standard, economic experiments might offer an important additional perspective which better helps to understand the relation between trust in and utilisation of AI.

Methodological Approach

Behavioural or experimental economics are an empirical approach to investigating supposedly irrational human behaviour in the context of economic decision making. This is usually done in a laboratory environment. Participants entering the laboratory are aware that their behaviour is being

[25] Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. Science and Engineering Ethics, 21(3), 619–630.
[26] Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573-1576.

monitored, recorded and will be scrutinized once the assessment concludes.[27] In fact one of the critical principles of economic experiments is that no deception of any kind is employed, unlike psychological studies that often disguise the underlying reasons behind what behaviour the experiment is measuring. Laboratory settings in economic experiments are highly stylised and often have participants interact with each other anonymously via a network of desktop computers. Crucially here, people taking part in economic experiments are incentivised to ensure that their behaviour reflects some analogue of how they would behave in real world settings.

The benefit of this approach is the experiments provide researchers with the ability to directly influence and control almost all relevant factors which impact the behaviour of the participants. Researchers can therefore gain ceteris paribus insights which would almost be impossible to obtain in less controlled environments.[28] Since there might be a plethora of factors determining the outcomes of human/AI interaction which we know little about so far, it seems only prudent to adopt an approach that gives a larger amount of control. Thus, experimental economics is a fruitful approach for policy professionals who wish to learn more about behavioural factors influencing the human use of artificial intelligence, particularly in relation to trust.

Independently from the issue of artificial intelligence, the said approach has become increasingly popular among economists over the years. While there were almost no publications using the respective methodology until the 1960s there were already more than 200 experimental papers published by 1998.[29] Parts of the respective toolkit are now used among several formerly distinct disciplines as well. As a result, the line between economic experiments and other empirical sciences such as social psychology has become increasingly blurry. However, there are several key distinctions left which help to distinguish economic experiments from other approaches. Key among them are enactment of detailed scripts for participants, performance based monetary payments and the proscription against deception, as previously mentioned.[30] The first two might actually prove useful for policy professionals while the latter could be considered a challenge. We will discuss all three and their relation to AI in detail.

The value of such experiments for policy professionals hinges on the question of whether insights gained in the laboratory can be generalized and applied to real world issues. Theres is an ongoing debate about this question, namely the external validity of economic experiments or the generalizability of respective findings.[31]

The first distinctive attribute of economic experiments is script enactment. It is not uncommon in behavioural sciences to conduct field experiments after which participants are asked to give

[27] Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. Journal of Economic perspectives, 21(2), 153-174.
[28] List, J. A., & Levitt, S. D. (2005). What do laboratory experiments tell us about the real world. NBER working paper, 14-20.
[29] Holt, C. A. (2005). Games and Strategic Behavior: Recipes for Interactive Learning. Unpublished Manuscript.
[30] Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists?. Behavioral and Brain Sciences, 24(3), 383.
[31] A good introduction can be found in Moffatt, P., Starmer, C., Sugden, R., Bardsley, N., Cubitt, R., & Loomes, G. (2009). Experimental economics: Rethinking the rules. Princeton University Press.

impromptu explanations for their behaviour, for example in hindsight bias experiments.[32] On the other hand, experimental economists usually provide participants with detailed information describing their own choices, the choices of other players and possible payoffs in accurate detail. This makes it much easier to replicate results – if there are any - and helps to understand details with more subtle influences. In part this is because a detailed script reduces ambiguity and focusses the attention of participants on cues that are actively communicated. This gives researchers more control over participants interpretation of the experimental setup and the variables they wish to investigate.[33] Given the variety of factors that might influence the perception and use of AI this could be a crucial element. For example, the appearance of Artificial Intelligence and whether it is presented in a way that might make it seem more human could drastically alter the way participants interact with it. This is usually discussed in relation to the phenomenon of anthropomorphism i.e. the attribution of distinctively human characteristics to nonhuman entities like robots.[34] This phenomenon is especially interesting since it appears to have a strong effect on peoples trust in artificial intelligence.[35] Hence such characteristics should be carefully excluded from experiments if they are not subject to scrutiny. Giving participants a detailed script instead of having them interact directly with AI might be a good way to do exactly that and thereby to gain ceteris paribus insights about human/AI interaction.

Another distinct methodological feature of economic experiments is the use of performance based financial incentives. Economic experiments were originally designed to test economic theory, which provides a framework of maximization assumptions as well as standards for optimal behaviour. Financial incentives are the most straightforward way to test whether actual behaviour is in accordance with those standards. Participants in economic experiments do therefore often receive a varying payoff based on their performance instead of a flat fee compensating them for their mere participation.[36] In other words, their compensation depends on how well they do in the experiment and their interaction with other participants. This does not mean that experiments which are designed in that way exclude social factors such as altruism, reciprocity, or trust. They are equally or even more informative regarding nonmonetary effects if they are carefully controlled for. Performance-based incentives also help to achieve this by reducing performance variability. Giving participants a reward for achieving results keeps them from acting randomly and reduce the variability of performance.[37] The critical assumption is that they will try to maximize their payoff by carefully considering their options instead of making the minimal effort necessary to receive a flat fee. This can be crucial if one is looking for more nuanced effects which might otherwise be overshadowed.

[32] Davies, M. F. (1992) Field dependence and hindsight bias: Cognitive restructuring and the generation of reasons. Journal of Research in Personality 26:58–74.
[33] Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists?. Behavioral and Brain Sciences, 24(3), 385.
[34] Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. International Journal of Social Robotics, 1(1), 71–81.
[35] Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology, 52, 113–117.
[36] Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists?. Behavioral and Brain Sciences, 24(3), 390-396.
[37] Davis, D. D. & Holt, C. A. (1993) Experimental economics. Princeton University Press, 25.

Considering the issue of trust in artificial intelligence, researchers can for example ask participants to play a so-called trust game with human and non-human responder. For example, Gogoll and Uhl (2018) investigated whether the level of trust toward machines and humans is different by giving participants the options to transfer an amount of money to a trustee. The sent amount was tripled and credited to the trustee who could than reciprocate by paying the sender back with any integral amount. They created a monetary incentive for participants to reveal their actual level of trust in humans and algorithms. One result of the experiment was that the average amount of money sent did not differ between human and non-human trustees. Surprisingly, this was also true for participants who had preferred human agents over algorithms in a delegation decision made earlier in the experiment. They had preferred to delegate a task that affected a third party to a human than to a machine, notably monitored by observers who would later have the opportunity to reward that decision. Still, they trusted algorithms with the same amount of money as they did with human trustees.[38]

Establishing a straightforward connection between participants decisions and their final pay-off has another potential benefit, as people's choices could be considered less hypothetical. In general, performance based financial incentives imply that the choices of subjects have real consequences for them or others. Given the non-saturation principle we can believe those consequences to be relevant. This could be especially beneficial for policy makers considering the assumed gap between peoples moral reasoning and their actual behaviour. People are incentivised to reveal their real preferences instead of moral considerations which they only value in theory.

The proscription against deception is likely one of the most characteristic aspects of economic experiments. Even though it could be argued that economists use deception at least implicitly from time to time, it is usually considered a taboo.[39] The reasons for that are not moral but merely pragmatic. Economists view the trust of participants as a common good which is shared among them.[40] Researchers avoid deceiving their subjects to maintain a solid reputation among potential participants. The long-term goal is to ensure that participants are motivated by monetary rewards instead of irrational reactions to suspected manipulation.

Given that manipulation is a common fear in relation to artificial intelligence this could be an issue. Policy professional might not be able to gain insights from respective experiments if the necessary methods are a priori excluded from their toolbox. One might argue that this is one of the cases in which deception is a methodological necessity and needed to create situations which would not occur naturally.[41]

However, it might not actually be necessary to deceive participants to investigate the use of artificial intelligence in manipulating people. One could for example seek a definition of manipulation that is

[38] Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. Journal of Behavioral and Experimental Economics, 74, 97-103.
[39] Hersch, Gil. "Experimental economics' inconsistent ban on deception." Studies in History and Philosophy of Science Part A 52 (2015): 13-19.
[40] Ledyard, J. O. (1995) Public goods: A survey of experimental research. In: The handbook of experimental economics, ed. J. H. Kagel & A. E. Roth. Princeton University Press.
[41] Kimmel, A. J. (1996). Ethical issues in behavioral research: A survey. Blackwell Publishing, 68.

non-deceptive.[42] Such definitions can then be operationalized for experimental setups by creating behavioural proxies. It is also not necessary that they encapsulate every aspect of the original concept, if the proxy used in an experiment gives insights into human behaviour. Put differently, finding a definition of manipulation that does not rely on deception is sufficient if it potentially leads to a change in people's actions. A related example for this is blame attribution. Blame is a complex and morally charged term, but it has the strongest effect on people's behaviour when it is operationalized into punishment.[43] Regarding the use of AI, results of a laboratory experiment suggest that decision-makers can shield themselves from the assignment of blame more effectively by delegating other-regarding tasks to algorithms than by delegating to other people. This implies that the availability of AI agents could provide a strong incentive to delegate sensitive decisions. These findings are informative regardless of one's specific definition of blame as punishment was used as a behavioural proxy.[44] A similar view might be adopted for the investigation of manipulation. From a policy professional perspective, it seems meaningful to focus on the most potent behavioural factors instead of the most complex concepts, especially if they provide valuable empirical feedback for the formulation and implementation of policies.

Conclusion and Recommendations

Policy professionals dealing with artificial intelligence face a daunting task and understanding technological aspects might not even be the most complex problem at hand. Despite the complexity of AI and the non-linear decision process of Large Language Models, comprehending the factors determining human decisions regarding AI use appears to be equally complex.

Today there is no decisive answer to the question whether people overly trust or distrust artificial intelligence in general. Additionally, little is known about factors that facilitate trust in artificial intelligence and when trust is a suitable measure to predict the use of AI in different domains. There are several research questions related to this and the answers might be highly relevant for policy professionals who wish to understand where and what policies could be implemented to facilitate the sustainable use of AI.

One such question surely is which factors influence trust in artificial intelligence and which factors drive egocentric discounting of algorithmic advice by decision makers. Such factors could be the nature of tasks, the domain in which they occur and the level of digital literacy of decision makers. Understanding these factors might help to decide when increasing trust in AI is a good policy objective and when other measures should be applied or at least considered.

They might also help to understand when it is sensible to keep human decision makers in the loop and adopt a form of augmented decision making or – in contrast - when autonomous AI decisions might

[42] Cohen, S. (2018). Manipulation and deception. Australasian Journal of Philosophy, 96(3), 483-497.
[43] von Grundherr, M., Jauernig, J., & Uhl, M. (2021). To condemn is not to punish: An experiment on hypocrisy. Games, 12(2), 38.
[44] Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding behind machines: Artificial agents may help to evade punishment. Science and Engineering Ethics, 28(2), 19.

yield the best results. Policies reducing the autonomy of AI might put peoples mind at ease but will not necessarily produce the best results for the public. At worst, they might cause substantial externalities to third parties. As argued above policy professionals might benefit from actively seeking out empirical economic research on the subject. Distinct aspects of the methodology of behavioural economics might make it especially suitable for helping to understand human / AI interaction or do at least offer an interesting perspective in addition to other behavioural sciences.

What is therefore recommend is that policy professionals who deal with AI related issues include questions regarding trust, algorithm aversion and algorithm appreciation in their areas of research interest. When they publish documents that indicate such areas or when inviting evidence for policy making, they should formulate related questions in a way that focuses on behavioural causes and effects instead of asking for practical solutions directly. This is in line with the idea that there is a misalignment between policy and academia which could be improved by a revision of the structure of questions published by government departments, agencies, and public bodies.[45] While the questions at hand easily transcend the expertise of any single discipline, economic experiments are likely to provide a valuable perspective on them. Policy makers might therefore also ask that their questions are answered using detailed scripts or performance based monetary payments.

Apart from this, policy professionals should frequently explore approaches which do not focus on making AI more trustworthy. For example, policies which foster digital literacy, an understanding of related ethical or legal issues and education about the responsible use of algorithms. Otherwise, they might create incentives to overly trust artificial intelligence or exploit it once the opportunity arises.

[45]Osman, M.& Cosstick N. (2022). Do policy questions match up with research questions? (2022) Centre for Science and Policy, University of Cambridge Working Paper series.