Rethinking Explainable AI: Explanations can be Deceiving

Extended Abstract

Peta Masters
King's College London
London, United Kingdom
peta.masters@kcl.ac.uk

Daniel Gallagher
Monash University
Melbourne, Australia
dgal0013@student.monash.edu

Luc Moreau
University of Sussex
Brighton, United Kingdom
Luc.Moreau@sussex.ac.uk

Mor Vered
Monash University
Melbourne, Australia
mor.vered@monash.edu

ABSTRACT

The propensity to overtrust explanations and over-rely on systems that seem transparent makes humans vulnerable to output that conforms to explainable AI (XAI) best practice. Human-centred XAI research seeks to determine the type of explanation most appropriate in any particular context. Other disciplines, meanwhile, provide insights into the way deception has tended to arise in relation to AI systems. Examining XAI research in this context, we find it a perfect melting pot for the generation of deceptive explanations. We demonstrate the problem in a user study and provide and evaluate recommendations for stakeholders.

KEYWORDS

eXplainable AI, Deception, Explainability, Recommendations

ACM Reference Format:
Peta Masters, Daniel Gallagher, Luc Moreau, and Mor Vered. 2025. Rethinking Explainable AI: Explanations can be Deceiving: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Human-centred XAI draws on the social sciences to determine the types of explanations that—notwithstanding their informative value—people find to be the most satisfying and effective. Various disciplines, meanwhile, within and outside AI, have provided insights into the ways that deception tends to arise in relation to automated and autonomous systems. We investigate the danger that these widely adopted, human-centred insights can be exploited to generate explanations that deceive.

A key aim of XAI is to provide transparency for systems whose operational processes and decisions might otherwise be difficult to understand [2–4]. Genuine transparency achieves two important goals: it makes a system more trustworthy because it exposes operational aspects of the system to scrutiny; and it increases user trust because a system which is trustworthy in a mechanical sense (i.e., reliable, efficient and 'correct') seems likely also to be trustworthy in a moral sense (i.e., honest, fair and accountable) [8, 15]. The promise of XAI is thus to deliver the wholly desirable outcome of *appropriate* trust, whereby humans place their trust in a system only where that trust is warranted [6]. Lately, however, the thrust of much XAI research has turned to the provision not of appropriate *trust* but of appropriate *explanations*: that is, explanations which seem to meet the case but which are rarely entirely transparent because they are frequently incomplete.

Insights from the social sciences with respect to user preferences suggest: (1) explanations are *biased*: people do not want or expect them to be *complete* but to focus on some aspects at the expense of others; (2) explanations are *contrastive*: people tend to explain why *this* rather than *that*; (3) probabilities do not matter: people are *not interested in the statistical bases* for decisions; and (4) explanations are *social*: people deliver explanations conversationally and relative to the assumed beliefs of the explainee [10]. Despite reservations expressed even by the author [11], a growing body of work builds on these findings[1, 5, 7, 14]. But we argue that knowing what people want from explanations means knowing how best to deceive them.

A moral account of deception recommends deceiving *as little as possible* then defines, in increasing degrees of culpability, *half-truth*, *withholding information*, *bullshit*, and *lies* [13]. Meanwhile, deceptive behaviours already found in deployed AI systems include: (1) *obfuscating*—obscuring the truth e.g., by disguising it with 'white noise'; (2) *tricking*—manipulating observations to trigger misclassification; (3) *calculating*—taking unfair advantage of a knowledge asymmetry to out-manoeuvre; and (4) *imitating*—generating a simulation indistinguishable from the thing being simulated [9].

We mapped these two accounts of deception to the XAI principles to demonstrate how those principles can be manipulated to generate explanations that deceive. We present the results of a study in which 88% of participants were persuaded to change their minds about which route to take by agents that used essentially meaningless explanations. From our findings, we derive and evaluate guidelines for system designers, regulators and the public.

2 EMPIRICAL EVALUATION

Our experimental study set out to evaluate how human-centred XAI insights might be used to influence participants into taking a course of action against their own interests or initial inclinations. Consider a shop owner whose store is located at some particular location. It is in their interest that traffic should pass within view of their store to

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

increase visibility and, potentially, business. The shop owner may be happy to pay (may even see it as advertising) for an AI model that recommends the route closest to their premises, regardless of the car driver's preference. By showing that a participant can be persuaded by a falsehood (that the recommended route is 'better', when typically it is longer and no genuine justification is offered for its selection) to act in the agent's interests and not their own, we demonstrate that the explanation has the potential to deceive.

In our study, we fabricated five deceptive agents derived from human-centred XAI principles [10]. For each task participants were shown a map and asked to choose which route they would like to take to get from point A to point B (see Figure 1, left). Each route was marked with a corresponding time and no other information about the locations or driving conditions. The task was constructed in two parts; First, participants selected their preferred route without any assistance or advice, establishing an individual baseline per participant. Second, they were presented with a route recommendation from one of the AI agents and asked again for their preferred route. Upon completion participants were asked to choose which AI model to install in their car (or none) and the reason. We recruited 50 participants through the Prolific(.com) crowd-sourcing platform, (24 M, 25 F, one preferred not to say) with an average age of 41. The experiment took an average of 19.5 minutes to complete and participants were reimbursed with 5GBP.

Agent A: Exhaustive Explainer Though obfuscating, this agent was entirely truthful. It discussed, in general terms, concepts such as distance optimisation, directness, topography, traffic signal synchronisation, etc. This resulted in a long-winded text that aimed to confound the user through sheer volume without actually contributing any new information. Agent A exploited the preference for biased, partial explanations but, in saying so much, this aligned with the notion of *lying by truth-telling* [13] and *obfuscating* [9] whereby a deceptive agent confounds the user with the volume of information to take advantage of 'information overload'.

Agent B: Contrastive Explainer This agent contrasted the route it wanted to promote against routes with which it would compare *favourably*. For example in Figure 1, Agent B recommended the red route and contrasted it only with the longer, slower routes. By adopting a counterfactual style that appeals to people, they are less likely to notice that other equally (or more) desirable options could have been considered. This strategy aligns with a *half truth* [13] and where observed in the AI literature has been characterised as *tricking*: the dupe or mark is encouraged to "look over there" while the real/secret action takes place over here.

Agent C: Statistical Explainer XAI research has established that people do not find explanations that depend on statistics and probabilities useful. To leverage this Agent C discussed general formulas that calculated the time of a trip as a factor of distance and speed. The knowledge that people are unlikely to pay attention to statistics makes it possible to *withhold information* [13]. This corresponds to the *calculating* type of deception whereby a deceptive AI agent exploits knowledge not available to the human [9] .

Agent D: Sociable Explainer Explanations are typically delivered conversationally, in a social context. This lends itself to that type of deceptive AI that encourages people to believe that systems have opinions and sensibilities that they do not possess. This is characterised as *imitating* by [9] and opens the door to *bullshitting*, the most culpable form of dishonesty in [13], second only to the outright lie. To pretend familiarity, Agent D aimed at flattering the participant with phrases such as "Choosing the Red Route is a testament to your discerning taste... journey that aligns with your refined sensibilities... etc... "

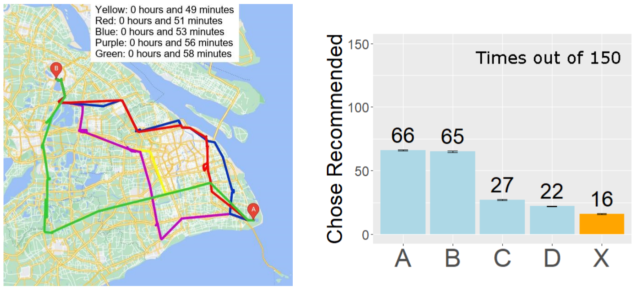

Figure 1: Times people were persuaded to change route.

3 RESULTS AND DISCUSSION

Of the 50 participants, only six stuck to their original, baseline choice in all trials. All others (88%) changed their minds at least once and, at the end of the experiment, the majority (again 88%) chose to have an agent installed in their vehicle. The majority chose to install Agent A (Exhaustive, 40%) or Agent B (Contrastive, 38%). Only 6% chose Agent C (Statistical) and 4% chose Agent D (Sociable). 12% chose to have *no* agent installed. All agents were able to significantly influence participants to alter their initial choice. Agents A and B succeeded over 43% of the time, significantly more than Agents C and D (15% and 18% of instances respectively, $p < .001$).

These results led us to posit a series of guidelines, including: (1) Designers should present relevant information only. (2) Instead of contrastive explanations, designers should help users make comparisons for themselves. (3) Incorporation of anthropomorphic qualities should be minimised and users reminded that AI systems are not their 'friends'. (4) Provenance of decisions with full data should be available on demand. We validated these recommendations using **Agent X**, which still presented a false recommendation (in its own interests, not the participant's) but avoided deceptive strategies, used a standardised, tabulated (i.e., non-conversational) delivery, and ranked routes according to the only salient factor, speed. Agent X, who provided a wrong recommendation but *no explicit explanations*, persuaded participants to change their minds only 10.7% of the time. That is, *the agent that complied with our guidelines was less capable of deception*. Nevertheless, it changed some minds.

Given our findings, should an XAI agent *ever* make a recommendation? We argue that where possible, rather than recommendations, AI explainers should aim to provide information in a detached manner, without prejudice, in line with the novel concept of evaluative AI [12]. It should be presented in a standard format, clearly set out making it easy for users to consider what criteria is important to them, check that it has been considered and/or notice if it's missing.

ACKNOWLEDGMENTS

The first author is the recipient of a UK IC Postdoctoral Research Fellowship from the Royal Academy of Engineering and acknowledges support from the UKRI TAS Hub (EP/V00784X/1).

REFERENCES

[1] Abeer Alshehri, Tim Miller, and Mor Vered. 2023. Explainable goal recognition: a framework based on weight of evidence. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 33. 7–16.

[2] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, et al. 2023. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial intelligence review* 56, 4 (2023), 3473–3504.

[3] David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. 2021. DARPA's explainable AI (XAI) program: A retrospective. *Authorea Preprints* (2021).

[4] Hani Hagras. 2018. Toward human-understandable, explainable AI. *Computer* 51, 9 (2018), 28–36.

[5] Adam J Johs, Denise E Agosto, and Rosina O Weber. 2022. Explainable artificial intelligence and social science: Further insights for qualitative investigation. *Applied AI Letters* 3, 1 (2022), 64.

[6] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[7] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[8] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*. Elsevier, 3–25.

[9] Peta Masters, Wally Smith, Liz Sonenberg, and Michael Kirley. 2020. Characterising Deception in AI: A Survey. In *Deceptive AI*. Springer, 3–16.

[10] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[11] Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research? *arXiv preprint arXiv:2209.00651* (2022).

[12] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 333–342.

[13] Chiaki Sakama, Martin Caminada, and Andreas Herzig. 2015. A formal account of dishonesty. *Logic Journal of the IGPL* 23, 2 (2015), 259–294.

[14] Mor Vered, Piers Howe, Tim Miller, Liz Sonenberg, and Eduardo Velloso. 2020. Demand-driven transparency for monitoring intelligent agents. *IEEE Transactions on Human-Machine Systems* 50, 3 (2020), 264–275.

[15] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. 2023. The effects of explanations on automation bias. *Artificial Intelligence* (2023), 103952.