Computers in Human Behavior Reports 10 (2023) 100290

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Human control redressed: Comparing AI and human predictability in a real-effort task

Serhiy Kandul [b,*], Vincent Micheli [c], Juliane Beck [a], Thomas Burri [a], François Fleuret [c], Markus Kneer [b], Markus Christen [b]

[a] University of St. Gallen, Switzerland
[b] University of Zurich, Switzerland
[c] University of Geneva, Switzerland

ARTICLE INFO

Dataset link: https://doi.org/10.17632/pbbvnjw638.1

Keywords:
Human control
AI predictability
Lunar lander game
Human-computer interaction

ABSTRACT

Predictability is a prerequisite for effective human control of artificial intelligence (AI). For example, the inability to predict the malfunctioning of AI impedes timely human intervention. In this paper, we employ a computerized navigation task, namely, a game called lunar lander, to investigate empirically how AI's predictability compares to humans' predictability. We ask participants to guess whether the landings of a spaceship performed by AI and humans will succeed. We show that humans are worse at predicting AI performance than at predicting human performance in this environment. Significantly, participants underestimate the differences in the relative predictability of AI and, at times, overestimate their prediction skills. These results raise doubts about the human ability to exercise control of AI effectively — at least in certain contexts.

1. Introduction

Using artificial intelligence (AI) in high-stakes environments requires "meaningful human control" (Siebert et al., 2022, Steen et al., 2022, de Sio & van den Hoven, 2018). Ensuring human control of AI serves two primary purposes: prevention of system mistakes, i.e., increasing the overall accuracy of human-AI teams, and attribution of responsibility when mistakes occur, i.e., assuring that someone is held accountable when a system causes (avoidable) damage or harm. Moreover, granting people control is known to reduce "algorithm aversion" (Dietvorst et al., 2018), i.e., it makes people rely more on algorithms and potentially improves their overall performance on a task. Assuring a certain degree of human control is likely to become a legal duty as the proposed EU Artificial Intelligence Act requires companies to guarantee "human oversight" of "high-risk" AI systems (Beck & Burri, 2022).

An essential prerequisite for effective human control is the sufficient *predictability* of AI. Predictability refers to a human's ability to foresee the output of an AI, allowing humans to detect and prevent AI-produced mistakes. Given that AI systems, especially advanced machine learning models, are often "black boxes", doubts abound about whether sufficient AI predictability is feasible. To address the issue of insufficient predictability, researchers have called for more explainable AI (Lipton, 2017, Miller, 2019).[1] While the existing literature investigates AI predictability and discusses how more explainable models improve predictability (Anderson et al., 2019, Chandrasekaran et al., 2018, Guillemé et al., 2019, Ribeiro et al., 2016, Iyer et al., 2018), it is not based on a shared understanding of *sufficient* predictability. Determining sufficient predictability is important for the possible trade-off between the predictability and the performance of an AI system (Bell et al., 2020). If making AI more predictable for humans diminishes its performance, it will be tempting to accommodate less AI predictability. At the same time, a certain level of predictability remains desirable to ensure human control of AI.

We contribute to the discussion on sufficient predictability in three important ways. First, we suggest that human predictability serves as a natural benchmark to assess the sufficiency of AI predictability. If there is a gap between AI and human predictability in the sense that AI predictability is lower, one could argue that controlling AI is more

* Corresponding author.
E-mail address: serhiy.kandul@ibme.uzh.ch (S. Kandul).

[1] "Explainability" refers to a more general understanding of an AI system's functioning, which might or might not lead to better predictability. In a narrow sense, however, predictability does not require explainability. For example, one could predict the AI's performance by simply looking at its accuracy in the past/in the training phase without having to understand how exactly the AI arrives at a particular outcome.

https://doi.org/10.1016/j.chbr.2023.100290
Received 15 February 2023; Received in revised form 14 April 2023; Accepted 19 April 2023
Available online 3 May 2023

difficult than controlling humans and look for possible remedies. Our first research question is thus:

RQ1. How does AI predictability compare to human predictability?

Second, we argue that if — in a specific task — there was a gap between the predictability of an AI system's performance and the predictability of a human operator's performance, and people did not correctly assess the size of this gap in predictability, there would be a reason for heightened concern. The second research question is thus:

RQ2. How well do people assess a potential gap between AI and human predictability?

If people believe that they correctly assess relative predictability, they will be interested in remedies to improve the predictability of an AI system. If, instead, people underestimate the predictability gap, they will not want to implement appropriate measures to reduce it and ultimately improve their ability to exercise control of the AI in relative terms.

Finally, we investigate how one's proficiency in a task affects one's ability to predict the performance of the two types of agents (humans and AI). Arguably, people performing well on a given task might better understand what leads to a failed or successful completion of the task. If this were the case, the natural solution to improve predictability would be to train people to perform the task better. However, if the AI or another human applied a different, perhaps unusual strategy to complete a task, and the environment was rich enough to allow for multiple ways of performing the task, improving people's own performance would not necessarily lead to higher predictability.

The third research question we address is:

RQ3. How does peoples' performance in a task affect predictability?.

When discussing the effects of task proficiency, we again look at AI and human predictability and compare objective and perceived predictability.

The remainder of the paper is organized as follows: Section 2 discusses related literature, Section 3 presents the design of our experiment, Section 4 highlights the key findings, Section 5 discusses the results, and Section 6 concludes.

2. Related work

2.1. Human vs. AI performance and behavior

A vast body of literature compares human and AI performance in various tasks. Machine learning algorithms have made tangible progress in tasks like chess or Go (Silver et al., 2016). Recent empirical papers plausibly argue that AI can reach or even exceed human-level performance in more complex tasks, such as image recognition (He et al., 2015, Russakovsky et al., 2015) or writing review articles (Blanco-Gonzalez et al., 2022). The differences in performance between AI and humans are often attributed to differences in learning patterns or cognitive abilities (Kühl et al., 2022). Despite recent technological developments, researchers in cognitive psychology and philosophy of science urge caution in taking evidence of better-than-human performance on a specific task as a sign of "human-like" intelligence or behavior (Cowley et al., 2022, Momennejad, 2023). For example, adapting to sudden changes in the decision environment still poses a significant challenge for AI systems (Lake et al., 2017, Crosby et al., 2019). Therefore, in many areas, AI operates very differently from humans. Such differences may be beneficial in some contexts, but they can lead to failures in others (Kühl et al., 2022).

In our experimental environment, the task is successfully landing a spaceship on a lunar surface. Humans tend to be unsuccessful at navigating, while AI agents trained to perform the navigational task tend to be good at it. AI agents, in particular, navigate differently than humans, adopting different landing patterns and maximizing the chances of landing successfully. Given this finding and the literature, we have adapted the environment of our experiments to the effect that AI *performs* equally to humans while still *navigating* differently than humans. We intuited that this difference in navigation, i.e., in "behavior" if we use anthropomorphic terms, poses a challenge to AI predictability.

2.2. Human-AI collaboration: evaluation and prediction

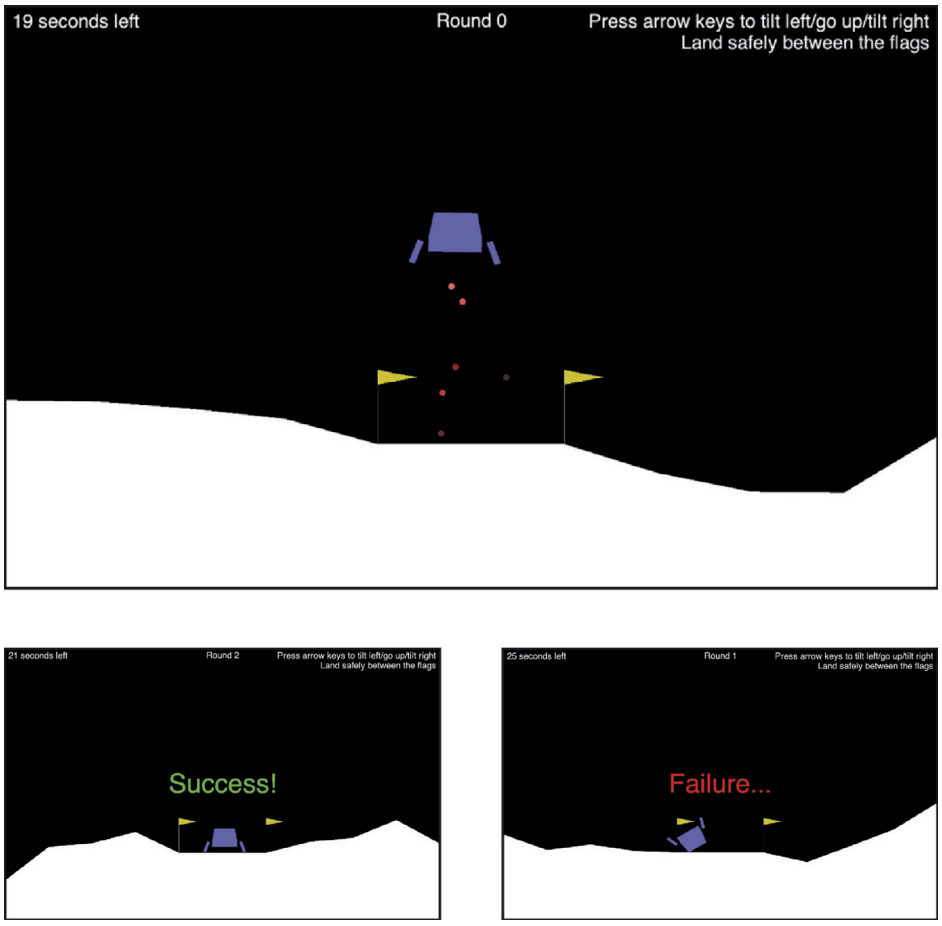
According to recent proposals, AI systems need to be studied "not as engineering artefacts, but as a class of actors with particular behavioral patterns and ecology" (Rahwan et al., 2019). Accordingly, empirical work in human-robot interaction has explored how humans evaluate machine behavior on the one hand and how humans predict machine behavior on the other.

Concerning the evaluation of machine behavior, extant studies have reported interesting and quite diverging findings. Malle et al. (2015), who were among the first to call for research in moral human-robot interaction, reported that people find consequentialist actions more morally acceptable for robots than humans in sacrificial dilemmas. Liu and Du (2021) explored the context of autonomous vehicles and found that, when accidents occur, people judge AI driving systems more responsible than human drivers. They are also blamed more (see also Hong (2020)). de Graaf and Malle (2019) showed that people employ similar mentalizing explanations for robot and human behavior, suggesting that they are "comfortable considering robots as having beliefs and knowledge, as being rational" (p. 245). Shank et al. (2019) explored evaluations of AI-driven agents modelled on real-life scenarios, such as a chatbot tweeting hate speech. In their studies, however, artificial agents were deemed less blameworthy than human agents engaging in identical actions (see also Shank and DeSanti (2018)). Kneer and Christen (2023) find that autonomous weapon systems in military contexts are also deemed less morally responsible than human agents when committing a war crime. Other work, by contrast, shows that in situations of reckless decision-making, AI-driven systems are blamed to similar extents as individual human agents and corporations (Stuart & Kneer, 2021) and that the more sophisticated the AI model, the higher people's propensity to ascribe moral blame to them (Kneer & Stuart, 2021). Overall, it seems that the evaluation of AI v. human behavior is strongly context-sensitive.

As regards people's ability to predict machine behavior adequately, which has attracted less attention, researchers have found that humans do well, e.g., in image classification (Zhou & Firestone, 2019). However, people's ability to adequately predict AI behavior is much weaker in more complex environments such as real-time strategy games (Anderson et al., 2019). In the experiment reported below, we asked participants to predict the overall outcome – success or failure – and thus the capability of AI rather than the type of action the AI would engage in.

2.3. Appropriate reliance in AI decision making

The issue of sufficient predictability that we aim to address in this paper is related to the notion of *appropriate* predictability, which concerns a human's ability to predict the successes and failures of an AI system. Appropriate predictability ultimately determines the appropriate level of reliance on a system, that is, the need to rely on the correct functioning of the system and intervene when the system fails (Schemmer et al., 2023). Inappropriate predictability might occur, for example, when people are good at recognizing a proper outcome but remain blind to system errors. In such cases, even if the predictability of AI were comparable to humans' predictability, there could still be *insufficient* predictability. This insufficient (or inappropriate) predictability, in turn, could lead to inappropriate reliance on the system ("overtrusting"), which some empirical studies have documented (Glikson & Woolley, 2020, Siau & Wang, 2018). To account for inappropriate/insufficient predictability, our environment allows for both types of mistakes: mispredicting success or failure.



Participants land a spaceship "between the flags" by pressing the respective arrow keys within a time limit. The initial position and the trajectories of the spaceship vary across landings. The countdown, the number of a trial, and the instructions on the keys/goal in the task, and the feedback, i.e. success or a failure, are provided on the screen.

Fig. 1. First stage of the experiment: a landing task.

3. Experimental design

We employed a computerized real-effort task: landing a spaceship in a lunar landscape (the lunar lander environment).[2] The task presents a challenging control problem. A Player plays a video game in which they fly and land a spaceship between two flags on a landing pad. The player has to complete the landing operation within a limited amount of time. The player controls the spacecraft by pressing keyboard keys to fire the main and side engines. Solving the task requires considering physical phenomena, such as acceleration, velocity, and rigid body dynamics. In addition to a real-time decision environment with fine-grained control, the task allows for easy training of potent deep learning agents such as PPO (https://arxiv.org/abs/1707.06347) and existing baselines from the Stable Baselines library (https://stable-baselines3.readthedocs.io/en/master/).

Fig. 1 illustrates the game and its two possible outcomes.

We simplified the game and chose only two possible outcomes: success or failure of the landing operation. We did not, for example, reward participants based on the overall time spent on a trial, the total energy used, and other components of more complex reward functions. Two features of the task are worth mentioning. First, the rules of the game are relatively simple to understand. Second, the task is new to most of the participants. We deliberately chose a task with which participants have little to no prior experience. That the task was new allowed us to measure the predictability of human and AI performance based on the same level of exposure to the game (induced in the experiment).

The AI we are considering in our experiment is a standard artificial neural network trained with classical reinforcement learning (Sutton & Barto, 2018). Starting with a random policy initially, the AI gradually improved the policy through a large number of games.

Most importantly, the behavioral pattern of the AI system and the human operator(s) in the game differ (as confirmed by a pre-test), i.e., AI and humans adopt different strategies to land the spaceship. AI first tends to navigate the spaceship outside of the area designated by the two flags on the screen. It then tries to slide it across the floor to the center of the screen, i.e., to the space between the two flags. Humans, by contrast, often attempt first to stabilize the spaceship in the center of the screen and then land it between the flags. This divergence of behavior emerges "naturally", as we did not guide the AI to land the object in any pre-defined way. We merely set the goal to maximize the chances of successful landings.

We suspect that it is common for humans in many real-life environments to employ different approaches than AI systems to problem-solving and suggest that a difference in the predictability of AI and humans is likely to emerge frequently.

[2] This is a well-known environment with active maintainers. You can find the general environment at the OpenAI gym library here https://github.com/openai/gym and the code we used in the experiment here: https://drive.google.com/file/d/1sJzPLEKBrNy2Oz3NidM27KkBfE1h1I9A/view?usp=share_link.

The experiment proceeded in three stages: In Stage 1, after an introduction to the game and after seeing a few examples of failed and successful landings, participants were asked to play 30 rounds of the game with feedback given between rounds but no payment. This stage allowed us to measure participants' own performance on the task and familiarize them with the game.

In Stage 2, the main stage of the experiment, participants performed the prediction task. For this task, we informed the participants that we had trained an artificially intelligent agent and a human operator to play the same game they had just played and each had played multiple rounds of the game. Before making a guess, participants could watch 50 to 75% of the duration of the landing (random cutoff points for each trial). We randomized the order of successful and unsuccessful pre-recorded landings for each participant. We then fixed the randomly drawn sequence of successful and unsuccessful landings for the AI and the human blocks. Importantly, we did not give the participants any feedback between each landing. We incentivized the prediction task. For a correct guess, participants earned eight points, and for a wrong guess, zero points. At the end of the experiment, we randomly selected two payoff-relevant landings, one per block of predictions (1 point = 1 CHF). In Stage 3, which occurred before participants learned about their performance in Stage 2, we elicited their beliefs about their performance in the prediction task. Participants submitted a number between 0 and 20 to indicate the number of presumed correct predictions. We then compared participants' answers with their true performance in the prediction task. Participants earned another 2 points if they guessed their performance correctly in each block (with +/-1 tolerance regarding the number guessed).

We implemented two treatments: FULL INFO and NO INFO. The treatments only differed in the information provided to the participants on the operator type, human or AI. In the FULL INFO treatment, participants knew whether the AI or a human had performed the landings. The information was continuously displayed on the screen. In the NO INFO treatment, participants did not receive any information on the operator type. In this treatment, participants made their predictions without knowing whether an AI or a human operator performed a particular landing. The comparison between these two treatments enabled us to assess whether any difference between AI and human predictability simply stems from the information on the operator type, i.e., the knowledge that an AI (or a human, respectively) performed a particular landing.

The experiment ended with a questionnaire. We elicited general attitudes and trust towards technology, participants' gaming experience, risk preferences, and socio-demographic data. We also collected open-ended responses about the strategies participants used when making predictions.

For the study, we complied with the ethical standards of the Faculty of Economics and Informatics of the University of Zurich.[3] The study fell into the category of "low risk" as no ethical issue arose, and we, in particular, refrained from using deception in the experiment. Given this, formal IRB approval was not required for the study.

4. Results

The main sessions of the experiment took place in June 2022 at the ETH Decision Science Lab. In total, 119 participants took part in the experiment. The participants were students of local universities recruited with a z-Root software from the pool of students registered at the University Registration Center.

The mean age of participants was 24.6; 56% of them were females. Regarding the highest education degree completed, 38% self-reported a high school, 32% a bachelor, 25% a master, and 2.5% a PhD degree. Around 86% of participants spend less than five hours a day playing video games, and 64% less than one hour per week. The average duration of a session was approximately 1 hour. The average earning was 30 CHF.

Table 1
Share of correct predictions by outcome of the landing in % and p-values of paired t-tests.

	Successes		Failures	
	AI	Human	AI	Human
FULL INFO	61.6	69.8	54.2	65.9
paired t.test	$p < 0.01$		$p < 0.001$	
NO INFO	58.6	65.6	57.5	66.2
paired t.test	$p < 0.001$		$p < 0.01$	

The share of correct predictions is computed as the number of correct predictions divided by the total number of predictions made. The share is computed separately for successful pre-recorded landings (columns 1, 2) and failed pre-recorded landings (columns 3 and 4).

To address our first research question (RQ1), we start by analyzing the share of correct predictions by treatment (FULL INFO v. NO INFO) and operator type red (AI v. human).

Fig. 2 compares the true predictability across operator types and treatment conditions (upper part) and guessed predictability across operator types and treatment conditions (lower part). We can make several observations concerning RQ1. First, we observe lower AI predictability than the human operator ($p < 0.01$, paired t-test). Second, we observe no difference across the treatments. The set of results suggests that the differences in the landing patterns rather than the information about the operator type drive the effect. The prediction levels across all the conditions are above the chance level of 50% ($p < 0.01$ in all four conditions).[4]

Next, we investigate how well participants separately predicted the successful and failed (pre-recorded) landings. Table 1 summarizes the findings.

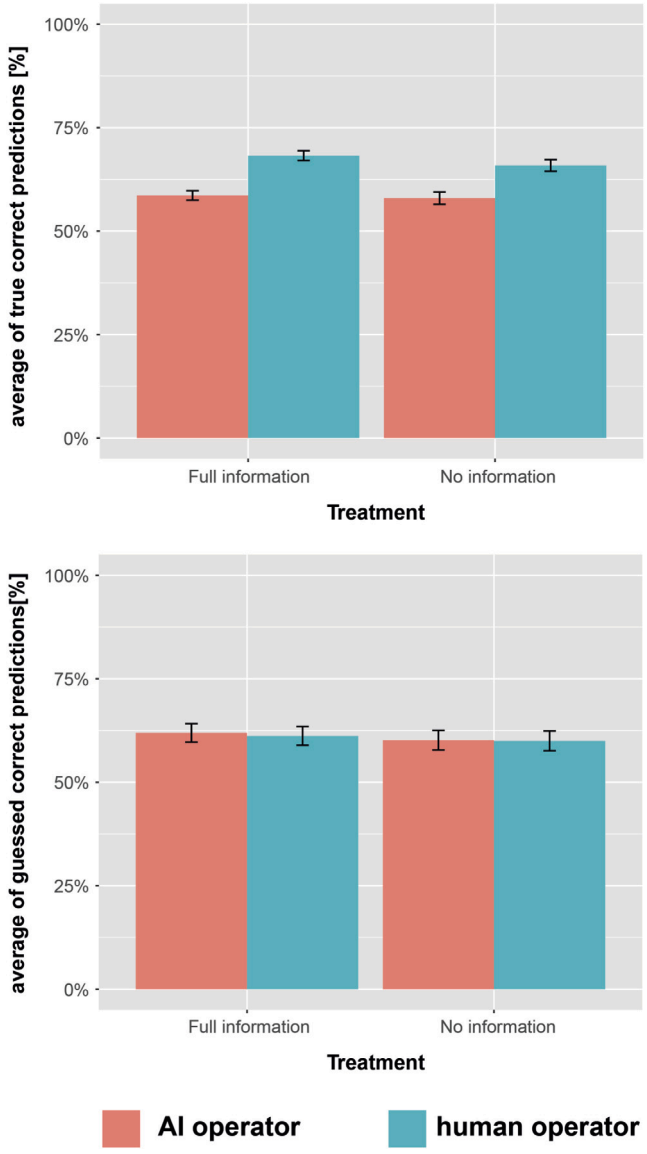
As Table 1 shows, the difference in AI and human operator predictability persists for both successful and failed landings in both treatment conditions. Concerning RQ1, these findings suggest that, in our context, AI's predictability is lower than human operators' predictability.

To address our second research question (RQ2), we turn to participants' beliefs about their performance in the prediction task. The lower part of Fig. 2 compares participants' guesses regarding the fraction of correct predictions they made. As Fig. 2 shows, participants did not realize the difference in the predictability of AI and human operator ($p = 0.82$ and $p = 0.96$ in FULL INFO and NO INFO, respectively). Interestingly, participants correctly guessed they were better than chance across all the conditions. They slightly underestimated the fraction of correct predictions for the human operator and stayed very close to the true predictability of the AI's landings. Concerning RQ2, the findings suggest that, in our context, people underestimate the predictability gap between AI and human operator we identified above.

Finally, we analyze how participants' performance in the landing task relates to their prediction success rate. We first note that the distribution of participants' performance is skewed towards zero, i.e., a majority of participants failed to land at least once successfully. On average, we observed 10.5% of landings as successful. This finding suggests that participants found the task difficult to perform.

[3] Available at: https://www.research.uzh.ch/en/procedures/ethikkommissionen.html.

[4] Given binary outcomes, the probability of judging the landing outcome correctly by random guess is 0.5. "Naive" guessing, i.e., persistently guessing "success" (or "failure") would result in 60% (40%) of correct guesses. None of the participants in our sample applied a naive guessing strategy, i.e., they all changed their prediction from "success" to "failure", or inversely, at least once during the game.



The share of correct predictions is computed as the number of correct predictions divided by the total number of predictions made. Upper panel: the observed fraction of correct predictions. Lower panel: participants' subjective evaluation of the fraction of correct predictions, i.e. their answers to the question: "Out of 20 landings, how many do you think to have predicted correctly?". Bars represent standard error of the mean.

Fig. 2. Predictability: Share of correct predictions by operator type.

To analyze the effect of task proficiency on prediction success rate, we ran a set of linear mixed-effect models:

$$Y_i = \beta_0 + \beta_1 * Treatment + \beta_2 * OperatorType + \beta_3 * LandingSkill_i + a_i + e_i \quad (1)$$

where Y_i - is the objective or guessed fraction of correct predictions over all trials, respectively;
$Treatment$ - is a dummy for treatment condition (FULL INFO or NO INFO);
$OperatorType$ - is a dummy for operator type (AI or Human)
$LandingSkill$- is the fraction of participants' own successful landings in the practice stage;
a_i - is the individual random effect; e_i - is the error term.

We refer to true predictability as the objective fraction of correct predictions across 20 landings in each block (human and AI) and to guessed predictability or beliefs as participants' own estimation of the fraction of correctly predicted outcomes of the landings in each block (human and AI). We ran the specification in Equation (1) with and without control variables. The controls include age, gender, and education level. Table 2 presents the regression results for the objective and guessed predictability (=beliefs). As Table 2 shows, the coefficient for *Operator* is positive and significant for the true predictability (Models 1 and 2). This confirms our earlier finding about the higher predictability of the human operator compared to AI. Participants' own performance in the landing task and the disclosure of information about the operator type do not significantly affect the true predictability, be it of the AI or the human operator's performance.

We then replaced the dependent variable Y in Equation (1) and ran the same specification for the effect of performance on perceived predictability (participants' beliefs). Models 3 and 4 from Table 2 show that the coefficients $Treatment$ and *Operator* remain insignificant. Interestingly, the coefficient $Landing\ skill$ is positive and significant for both

Table 2
Regression estimates: Predictability and beliefs.

	True predictability		Beliefs	
	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.59***	0.67***	0.59***	0.72***
	(0.01)	(0.04)	(0.02)	(0.09)
Treatment: Full INFO	−0.01	−0.01	−0.00	0.01
	(0.01)	(0.01)	(0.03)	(0.03)
Operator: Human	0.09***	0.09***	−0.00	−0.00
	(0.01)	(0.01)	(0.01)	(0.01)
Landing skill	0.01	−0.01	0.23**	0.17*
	(0.03)	(0.04)	(0.07)	(0.08)
Controls	No	Yes	No	Yes
Num. obs.	238	238	238	238
Num. groups	119	119	119	119

Linear mixed-effect model fitted. DV = The share of correct predictions, computed as the number of correct predictions divided by the total number of predictions made. Controls include age, gender, and education. Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

specifications with and without controls: A 10 p.p. improvement in the landing task boosts perceived predictability by about 2 p.p., whereas objectively, predictability is not affected by performance in the landing task. Regarding RQ3, we observe that if one performs better in the landing task, this does not lead to better predictability (of either AI or human operator). Instead, better performance is associated with more optimistic beliefs about one's ability to predict performance successfully. This finding suggests that more skilled participants might be overconfident about their prediction abilities.

5. Discussion

We have explored AI and human predictability in a real-effort task and discovered that in a novel environment, participants are better at predicting humans than AI. In contrast to a broad definition of predictability as a global understanding of AI ("shared mental models"), which requires people to be able to predict concrete outputs or actions of AI, we only required participants to guess the overall success of the task performed by AI or a human. Remarkably, a difference in AI and human operator predictability emerges in this very narrow setting. We attribute the effect on predictability to different behavioral patterns or strategies employed by human operators and AI to fulfill the task.

Lower predictability of AI coupled with peoples' unawareness of any difference in predictability suggests that when predictability is necessary for effective human control, controlling AI might be more challenging than controlling another human. Our findings, however, do not imply that humans should replace AI-based systems. In our setting, we deliberately equalized the accuracy across pre-recorded AI and expert human operator trials to elicit participants' predictions. In many real-life applications, however, AI outperforms humans. In these situations, the relative gain in performance should be weighed against a potential loss in relative predictability. We, therefore, extend the current discussion of the accuracy-predictability or efficiency-explainability trade-offs, which focuses on AI exclusively (Bell et al., 2020) by including the human-to-human dimension. We did not observe any difference in relative predictability for failed and successful outcomes. This finding might be an artifact of our design as any (mis)prediction has the same impact on participants' payoff. In reality, however, when AI performs well, and mistakes rarely occur, supervisors will shift their focus towards detecting mistakes. Further research could investigate how the asymmetry in harm/benefit from the erroneous and flawless operation of AI affects predictability.

Interestingly, participants' actual performance on the landing task was poor compared to the true performance of AI. However, low performance or lack of experience on a task is not a problem per se. In contrast, the fact that participants' own assessments of their ability to predict the outcome of landings performed by others were very close to the true predictability in the experiment suggests that being confronted with a difficult task might help people form correct beliefs about their predictions (and take measures if necessary). It turned out that high-skill performers, i.e., participants who manage to land themselves successfully, actually overestimate their prediction success. These results suggest that improving one's performance on a task does not necessarily result in an objectively better ability to predict the performance of others (AI or a human) on the same task. Instead, it might make people believe falsely that they can predict performance well and therefore effectively exercise control of others.

In our setting, we measured participants' performance in the task after a fixed number of trials. Future research might investigate how exogenous manipulation of skills, e.g., via training, affects the performance and predictability – be it objective or subjective – of AI and human operators.

6. Conclusion

This paper presents an empirical analysis of AI predictability in a real-effort task. The paper's primary focus is on the comparison between AI and human predictability. The key findings suggest that when AI behaves unlike humans (while maintaining human-level performance), AI predictability is lower than human predictability. Investigating the link between lower predictability, trust an d human control presents an intriguing avenue for future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data from the experiments is available here Kandul, Serhiy; Micheli, Vincent; Beck, Juliane; Burri, Thomas; Fleuret, François; Kneer, Markus; Christen, Markus (2023), AI_human_predictability, Mendeley Data, V1, https://doi.org/10.17632/pbbvnjw638.1.

Acknowledgements

This research has been funded by the Swiss National Science Foundation as part of the National Research Program 77 "Digital Transformation"; Grant Number 187494.

References

Anderson, A., Dodge, J., Sadarangani, A., et al. (2019). Explaining reinforcement learning to mere mortals: An empirical study. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 1328–1334).

Beck, J., & Burri, T. (2022). From 'human control' in international law to 'human oversight' in the new EU act on artificial intelligence (October 3, 2022). In *Research handbook on meaningful human control of artificial intelligence systems*.

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2020). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *2022 ACM conference on fairness, accountability, and transparency (FAccT '22)*.

Blanco-Gonzalez, A., Cabezon, A., Seco-Gonzalez, A., Conde-Torres, D., Riveiro, P. A., Pineiro, A., & Garcia-Fandino, R. (2022). *The role of AI in drug discovery: Challenges, opportunities, and strategies*.

Chandrasekaran, A., Prabhu, V., Yadav, D., Chattopadhyay, P., & Parikh, D. (2018). Do explanations make VQA models more predictable to a human? In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1036–1042). Brussels, Belgium: Association for Computational Linguistics.

Cowley, H. P., Natter, M., Gray-Roncal, K., et al. (2022). A framework for rigorous evaluation of human performance in human and machine learning comparison studies. *Nature: Scientific Reports, 12*.

Crosby, M., Beyret, M., & Halina, B. (2019). The animal-AI olympics. *Nature Machine Intelligence, 1*.

de Graaf, M. M. A., & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 239–248).

de Sio, S. F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, Sec. Ethics in Robotics and Artificial Intelligence*, *5*, Article 15.

Dietvorst, B., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Guillemé, M., Rozé, L., Masson, V., & Termier, A. (2019). Agnostic local explanation for time series classification. In *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)* (pp. 432–439).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE international conference on computer vision (ICCV)* (pp. 1026–1034).

Hong, J. W. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self- driving car accidents in experimental settings. *International Journal of Human-Computer Interaction*, *36*(18), 1768–1774.

Iyer, R., Li, Y., Li, H., Lewis, M., Sundar, R., & Sycara, K. (2018). *Transparency and explanation in deep reinforcement learning neural networks*In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*. Association for Computing Machinery.

Kneer, M., & Christen, M. (2023). Responsibility gaps and retributive dispositions: Evidence from the US, Japan and Germany, https://doi.org/10.13140/RG.2.2.34656.15367.

Kneer, M., & Stuart, M. T. (2021). Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 407–411).

Kühl, N., Goutier, M., Baier, L., Wolff, C., & Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, *76*, 78–92.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, Article e253.

Lipton, Z. C. (2017). *The mythos of model interpretability*.

Liu, P., & Du, Y. (2021). Blame attribution asymmetry in human–automation cooperation. *Risk Analysis*.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 117–124).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Momennejad, I. (2023). A rubric for human-like agents and NeuroAI. *Philosophical Transactions of the Royal Society, Section B*, *378*.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., & Crandall, J. W. (2019). *Nature*, *568*(7753), 477–486.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Russakovsky, O., Deng, J., & Su, H. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). *Appropriate reliance on AI advice: Conceptualization and the effect of explanations*.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648–663.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Academy of Management Annals*, *14*(2), 627–660.

Siebert, L. C., Lupetti, M., Aizenberg, E., et al. (2022). Meaningful human control: Actionable properties for AI system development. *AI and Ethics*.

Silver, D., Huang, A., Maddison, C., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489.

Steen, M., van Diggelen, J., & Timan, T. (2022). Meaningful human control of drones: Exploring human–machine teaming, informed by four different ethical perspectives. *AI and Ethics*.

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human–Computer Interaction*, *5*(CSCW2), 1–27.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd edition). The MIT Press. http://incompleteideas.net/book/the-book-2nd.html.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*.