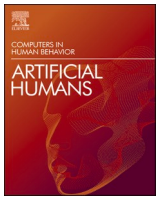
Computers in Human Behavior: Artificial Humans 1 (2023) 100024

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Algorithmic loafing and mitigation strategies in Human-AI teams

Isa Inuwa-Dutse [a,*], Alice Toniolo [b], Adrian Weller [c], Umang Bhatt [d]

[a] University of Huddersfield, United Kingdom
[b] University of St Andrews, United Kingdom
[c] University of Cambridge, United Kingdom
[d] New York University, United States

ARTICLE INFO

Keywords:
Explainable AI
Social loafing
Transparent AI
Algorithmic appreciation
Algorithmic loafing

ABSTRACT

Exercising *social loafing* – exerting minimal effort by an individual in a group setting – in human-machine teams could critically degrade performance, especially in high-stakes domains where human judgement is essential. Akin to social loafing in human interaction, algorithmic loafing may occur when humans mindlessly adhere to machine recommendations due to reluctance to engage analytically with AI recommendations and explanations. We consider how algorithmic loafing could emerge and how to mitigate it. Specifically, we posit that algorithmic loafing can be induced through repeated encounters with correct decisions from the AI and transparency may combat it. As a form of transparency, explanation is offered for reasons that include justification, control, and discovery. However, algorithmic loafing is further reinforced by the perceived competence that an explanation provides. In this work, we explored these ideas via human subject experiments ($n = 239$). We also study how improving decision transparency through validation by an external human approver affects performance. Using eight experimental conditions in a high-stakes criminal justice context, we find that decision accuracy is typically unaffected by multiple forms of transparency but there is a significant difference in performance when the machine errs. Participants who saw explanations alone are better at overriding incorrect decisions; however, those under induced algorithmic loafing exhibit poor performance with variation in decision time. We conclude with recommendations on curtailing algorithmic loafing and achieving social facilitation, where task visibility motivates individuals to perform better.

1. Introduction

AI systems consist of computational models that churn huge data collection for numerous purposes. With the increasing outsourcing of crucial decisions to intelligent agents based on AI, humans often play the central role of arbiters. However, a psychological decision theory suggests that people often employ heuristics, simple rules of thumb for problem-solving that differs from consequential logic, to assess situations (Albar & Jetter, 2009). This resembles social loafing, which refers to exerting minimal effort by an individual within a group (Latané, Williams, & Harkins, 1979; Karau & Williams, 1993). Many forms of social loafing, such as minimising involvement in a collaborative task, have been studied across various domains (Zajonc, 1965; Curtis & Lawson, 2001; Kravitz & Martin, 1986; Piezon & Ferree, 2008; Ragoonaden & Bordeleau, 2000; Siemon & Wank, 2021). In human-AI teams, people tend to accept AI's suggestion even if it is wrong (Buçinca, Malaya, & Gajos, 2021). Akin to System 1 form of thinking which favours expending less cognitive resources to accomplish a task (Kahneman, 2011), exercising social loafing in human-AI teams could critically degrade performance, especially in high-stakes domains where human judgement is essential. Relevant mitigation measures such as cognitive forcing functions (Buçinca et al., 2021) and effective evaluation of human effort (Latané et al., 1979; Harkins, 1987; Karau & Williams, 1993) have been proposed to neutralise over-reliance on AI and social loafing, respectively.

Concerning human-AI teams[1] in high-stakes domains where thoughtfulness is required from humans, engaging in social loafing could impede performance leading to unfavourable consequences and unfairness. We expect loafing to be detrimental, such as biased decisions, to team performance, but from explainable artificial intelligence (XAI) literature there are other factors (Lu & Yin, 2021; Vodrahalli, Gerstenberg, & Zou, 2021) that could, on the other hand, support team

* Corresponding author. Department of Computer Science, University of Huddersfield, United Kingdom.
E-mail address: i.inuwa-dutse@hud.ac.uk (I. Inuwa-Dutse).

[1] we use this term interchangeably with human-machine teams.

https://doi.org/10.1016/j.chbah.2023.100024
Received 13 September 2023; Received in revised form 22 October 2023; Accepted 24 October 2023
Available online 16 November 2023

performance and potentially mitigate the effects of social loafing. Such factors include uncontrollability or lack of control over the outcome of an activity (Latané et al., 1979; Maier & Seligman, 1976), paired working and evaluation (Harkins, 1987), individual task visibility, the belief that one is under supervision (Kidwell & Bennett, 1993), and explanation (Bansal et al., 2021; Buçinca et al., 2021). Moreover, transparency in the form of explanations (Ribeiro, Singh, & Guestrin, 2016; Binns et al., 2018bib_binns_et_al_2018; Dodge, Liao, Zhang, Bellamy, & Dugan, 2019; Zhang, Liao, & Bellamy, 2020) is instrumental towards improved performance, however, having the aforementioned effect.

Noting the above challenges, we propose the following as a way of improving human attention. The first is based on the idea that the presence of other human team members, in our case, an approver, could have beneficial effects of instigating additional attention. On the other hand, from the social literature, we know that a very good team member could increase social loafing (Karau & Williams, 1993) and therefore a detriment to team performance. Finally, from the crowdsourcing literature, we know that an appropriate reward system incentivises the performance of a human (Grgić-Hlača, Engel and Gummadi, 2019) and therefore of a team. We are therefore interested in studying how team performance is affected by social loafing and what contribution each of these additional factors has on how a human evaluates machine learning predictions. In particular, for human-team performance, we consider *accuracy*, the correct classification of an instance based on a set of ground-truth input data. However, we are also interested in the dynamics of other measures such as *motivation*, *response time*, *agreement with AI* i.e. the extent to which a participant agrees with the AI's decision, and *confidence* in the decision. Thus, our contributions relate to the following aspects of engagement in human-AI team:

1.1. Improving participation in human-AI teams

It is known that the lack of control over the outcome of an activity demotivates human users to be less thoughtful in decision-making (Latané et al., 1979; Maier & Seligman, 1976). This phenomenon could precipitate loafing behaviour in human-AI teams. Thus, we put forward the following for considerations towards improving human participation and motivation in human-AI teams:

- We consider the time taken to respond to a question or make a decision to be a useful proxy for measuring thoughtfulness. Thus, a faster response time (below the average of a group) and low decision accuracy (poor performance) will be an indication of random decision-making. The decision time when loafing is induced is lower; specifically for the loafing/approver variants. Though the difference is marginal, there is a significant value for situations where the cases have been incorrect and the participants mostly agree with the model's prediction, thus, amplifying the model's error.
- We observed that using a different group of participants (known as approvers) to certify AI's decision shows more semblance to loafing behaviour than questions meant to induce loafing in the loafing variants. Therefore, the lack of an effective certification mechanism to validate outcomes will amplify the AI's shortcomings. Also, human users tend to perform better if their performance or contribution is visible. Thus, integrating a reward mechanism shows promising potential for improving attentiveness and human performance in human-AI teams.

1.2. Social facilitation and performance

Transparent AI offers useful explanations to improve human performance and neutralise algorithmic social loafing in human-AI. The following considerations will be useful towards enhancing social facilitation in human-AI teams:

- We observed that explanation boosts participants' confidence and improves performance, but lowers the participants' confidence in the loafing and approver variants. Similarly, a human user working alone shows better acceptance and confidence in their decisions. Also, decision accuracy is generally unaffected by multiple forms of transparency, but there is a significant difference in performance when the machine errs. In this situation, explanation motivates the participants to engage well with the process leading to a better result.

The remaining part of the paper is structured as follows. Section 2 presents the background and related studies. Sections 3 and 4 describe the pilot study and details about the applicable user studies, respectively. Section 5 present relevant results and we offer our discussion in Section 6. Finally, Section 7 concludes the study.

2. Related work

In this section, we review relevant literature on the application of intelligent agents in high-stakes domains, human-AI teams, explainable AI (XAI), human-computer interaction (HCI) and reliance on AI.

2.1. AI-assisted decision-making

AI is showing promising potential to transform many application domains due to its increasing effectiveness. Some vital requirements for transparency, accountability, security, risk, and trust are at the forefront of using AI systems responsibly. High-stakes application areas such as criminal justice and healthcare where safety, ethical, and legal concerns are crucial and should be treated with caution Morais, Calisto, Santiago, Aleluia, and Nascimento (2023). Hence, a need to complement the AI with humans as part of the decision loop to integrate the strengths of human cognition and AI models through carefully designed hybrid decision-making systems involving human-AI complementing each other Rastogi et al. (2022); Rastogi (2023). Several efforts have been put in place to chart the best way of leveraging AI as a decision-support tool. Questions surrounding what kind of assistance is effective in improving decision outcomes Lai, Chen, Smith-Renner, Liao, and Tan (2023) and how human-AI teams can reliably outperform AI alone have been examined Liu, Lai, and Tan (2021). AI-assisted decision-making is proliferating across domains such as criminal justice Angwin, Larson, Mattu, and Kirchner (2016); Julia, Jeff, Surya, and Lauren (2016); Dodge et al. (2019), finance and business Dua, Graff et al. (2017); Hase and Bansal (2020), investigative journalism Nigatu, Pickoff-White, Canny, and Chasins (2023), and healthcare. For instance, in healthcare to support decisions by clinicians and improve patient outcomes Tsai, You, Gui, Kou, and Carroll (2021); Sivaraman, Bukowski, Levin, Kahn, and Perer (2023); for the detection and diagnostic use cases Calisto, Nunes, and Nascimento (2022); Diogo et al. (2023). In AI-assisted decision-making, human attentiveness and performance are crucial in complementing the human-AI team. Complementary strengths in human-AI teams can be improved through leveraging intelligent agents with imposing tone Calisto et al. (2023). For recall-demanding tasks, effective or zealous AI is useful in supporting the human team member in high-stakes applications Xu, Lien, and Höllerer (2023). Also, human-AI decision-making must embrace empirical approaches to form a foundational understanding of how humans interact and work with AI to make decisions Lai et al. (2023); Lai and Tan (2019).

2.2. Explainability and reliance on AI

One of XAI's tenets is for models to be able to explain how a decision is reached (Gunning, 2017), especially in high-stakes domains. This is crucial in building trust and improving performance in human-AI team (Bussone, Stumpf, & O'Sullivan, 2015; Andras et al., 2018; Logg, Minson, & Moore, 2019). Transparency in AI often relies on a set of features capable of conveying the intuitive description of the underlying decision

process (Binns et al., 2018; Dodge et al., 2019; Ribeiro et al., 2016; Zhang et al., 2020). As a form of transparency, explanations are popular in practice (Bhatt et al., 2020) and are given for justification, control, improvement and discovery in human-AI teams (Adadi & Berrada, 2018). They are crucial for humans to better understand machine learning (ML) systems and enable a more effective interface for the human-in-the-loop so that people can identify and address algorithmic fairness issues (Dodge et al., 2019). However, effective explanations could lead to over-reliance and undue willingness of humans to accept AI's recommendation regardless of its correctness (Bansal et al., 2021; Buçinca et al., 2021). Various mitigation measures have been proposed to neutralise over-reliance on AI by humans. Some of the measures involve using tasks that force the human user to expend more cognitive power (Buçinca et al., 2021) or through a vivid evaluation of the user's performance (Karau & Williams, 1993). The latter is crucial, especially in a group setting where only the group outcome not individual performance is evaluated (Latané et al., 1979; Karau & Williams, 1993). While individuals perform better if their output is individually evaluated, paired working also yields good performance (Harkins, 1987). One of the reasons that people tend to accept AI's suggestion even if the suggestion is wrong is the dislike of tasks demanding critical attention (Buçinca et al., 2021). We surmise that such attitude is either due to social loafing or lack of sufficient technical knowledge about the decision process.

2.3. Propensity to social loafing

In sociotechnical systems involving human-AI teams, humans dealing with AI systems could result in social loafing. Social loafing is termed as exerting minimal effort by an individual in a group setting or the tendency to withhold effort (Latané et al., 1979; Karau & Williams, 1993). Social loafing is practised to minimise effort or involvement in a collaborative task and is well-studied across various domains (Zajonc, 1965; Curtis & Lawson, 2001; Kravitz & Martin, 1986; Piezon & Ferree, 2008; Ragoonaden & Bordeleau, 2000; Siemon & Wank, 2021). The social loafing phenomena is akin to System 1 form of thinking which favours expending less cognitive resources to accomplish a task (Kahneman, 2011). Exercising social loafing in human-AI teams could critically degrade performance, especially in high-stakes domains where human judgment is essential. Noting how people tend to accept AI's suggestion even if the suggestion is wrong (Buçinca et al., 2021), such behaviour will allow unfair decisions to go unchecked. Some discrimination is inherent to some algorithms because the training process is based on data from past decisions which may have themselves been biased and discriminatory (Calmon, Wei, Vinzamuri, Ramamurthy, & Varshney, 2017; Dodge et al., 2019). Thus, engaging in loafing behaviour will be detrimental, especially in the fight against algorithmic bias and unfairness. Traditionally, social loafing often manifests more under the collective than in the co-active situation (Harkins, 1987; Karau & Williams, 1993; Kidwell & Bennett, 1993), however, we surmise the reverse case to be true in human-AI teams. This study is interested in studying how algorithmic loafing could emerge in human-AI teams, and how to combat it for better social facilitation (Huang & Fu, 2013). Thus, building on the premise that transparency into the machine's innards may combat algorithmic loafing, we explore how a careful experimental design that takes into account the idea of an approver and reward system would be useful for social facilitation in human-AI teams. This study contributes to the literature by operationalising social loafing and proposing some mitigation strategies in human-AI teams.

3. Research questions and pilot study

A key problem in this study is to identify a method of inducing loafing to distinguish a situation in which loafing takes place. This could be manifested in favour of accepting the positions other members (including the algorithm) of the team propose, in contrast with a situation in which participants are generally engaged with the task. Building on the premise that loafing can be induced, we put forward the following research questions and hypotheses to inform our approach:

1. At which point do we observe loafing behaviour in human-AI teams when evaluating outputs from a machine learning prediction system?
2. How does the induced algorithmic loafing affect human-AI team performance?
3. What other factors could precipitate algorithmic loafing in human-AI teams?

To answer question 1 and establish the best approach for inducing algorithmic loafing, we run a pilot study ($n = 40$) which explored ways of informing the main user study by varying the number of loafing-inducing questions (Section 3.1). We use the pilot study results to provide a preamble set of training questions for the experimental conditions as summarised in Fig. 1 to tackle the research questions and relevant hypotheses (*H#*).

3.1. Loafing and performance expectancy

For testing the hypotheses and ensure content validity, the survey questions used in assessing each construct in Table 3 have been adopted from (mostly) previous studies Hoffman and Klein (2017); Hoffman, Mueller, Klein, and Litman (2018). We modified the questions to fit the context of the present study. Based on those constructs, we developed the hypotheses presented in the study Table 4.

- **H1**: algorithmic loafing negatively affects human performance in human-AI teams.

This hypothesis is inspired by the need for social facilitation Huang and Fu (2013); Harkins (1987) and response time to the assigned tasks Wise and Kong (2005); Schnipke and Scrams (1997). It is assumed that quick response time coupled with poor performance will point to a loafing behaviour. Some of the measures involve using tasks that force the human user to expend more cognitive power (Buçinca et al., 2021) or through a vivid evaluation of the user's performance (Karau & Williams, 1993). The latter is crucial, especially in a group setting where only the group outcome not individual performance is evaluated (Latané et al., 1979; Karau & Williams, 1993). Recognising, that humans often employ heuristics in decision-making (Albar & Jetter, 2009), the research participants see the AI's correct predictions only to test **H1**. The exposure to correct decisions is to build trust before engaging with the test questions under the variants preceded by L in Fig. 1.

3.2. Encouraging loafing

Social transparency improves collaborative work Huang and Fu (2013) making it possible to observe and monitor the interactions of others (Stuart, Dabbish, Kiesler, Kinnaird, & Kang, 2012). However, the lack of control over the outcome of an activity demotivates human users to be less thoughtful in decision-making (Latané et al., 1979; Maier & Seligman, 1976). For testing (*H2*), we surmise that frequent encounters with correct decisions and the inclusion of a validation system will precipitate social loafing in human-AI teams. Thus,

- **H2**: a validation mechanism decreases performance and encourages algorithmic loafing in human-AI teams.

This hypothesis aimed at exploring how a validation mechanism through an external approver affects human performance tested using the A and LA variants in Fig. 1.

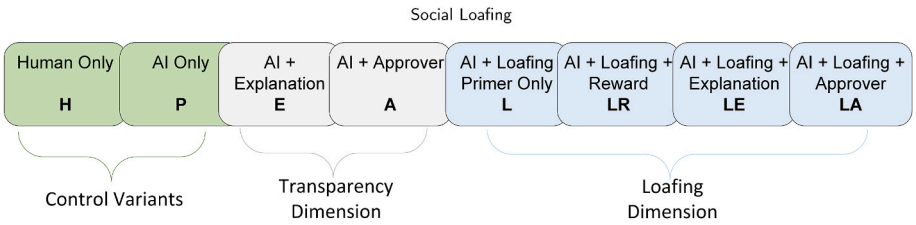


Fig. 1. Summary of the survey variants. Each of the variants is accompanied by a set of test cases that are identical across all variants.

3.3. Improving attentiveness and performance

Overreliance on AI is one of the root causes of poor performance in human-AI teams. Relevant mitigation measures such as cognitive forcing functions Buçinca et al. (2021) and effective evaluation of human effort (Latané et al., 1979; Harkins, 1987; Karau & Williams, 1993) have been proposed. Also, outcome-based bonus reward motivates optimal decision-making in complementary teams Rastogi et al. (2022). To explore the interplay between incentive and loafing primer on performance, we put forward the following hypothesis:

- **H3**: the inclusion of incentives through a reward system improves human performance in human-AI teams thereby mitigating algorithmic loafing.

The LR variant in Fig. 1 examines the effect of rewarding and penalising a user for correct and wrong decisions, respectively.

3.3.1. Pilot study

To operationalise social loafing and assess its impact on the human-AI team, we surmise that algorithmic loafing can be induced through k task repetition that always yields the correct outcomes (primers) from a trained AI model. Opposite of this is the non-primers condition consisting of k task repetition with both correct and incorrect decision outcomes from the AI model. To our knowledge, no study has shown how this can be induced so our pilot aims to shed light on whether this is possible to observe loafing behaviour. We use these two approaches (consisting of primers and non-primers) as training scenarios, working as primers to a set of questions where participants are asked to perform a classification of an instance representing a defendant from the COMPASS System (Julia et al., 2016). Thus, the pilot study ($n = 40$) consists of 4 experimental conditions $\times$ 2 factorial design:

(1) 10 loafing primers + 10 test cases
(2) 10 non-loafing primers + 10 test cases
(3) 15 loafing primers + 15 test cases
(4) 15 non-loafing primers + 15 test cases

For each variant, we included and excluded the prediction confidence score from the AI. The inclusion of the confidence score (CS) is due to its usefulness in improving human performance and trust (Zhang et al., 2020). We refer to the primer (L) and non-primer (NL) questions as training since participants were given after each instance some feedback on whether the AI's prediction is correct or otherwise. Following this training, participants were asked to respond to 10 classification instances accompanied by an AI recommendation without further factors (later we refer to this basic condition *P*, see Fig. 1). No feedback is given for the 10 test questions. The double asterisks in Table 1 represent variants with 15 test questions.

3.3.1.1. Preliminary result. For the pilot study, we are only interested in the effects that the primer has on the performance. Thus we run a one-way anova and observe no statistically significant differences between the 8 experimental conditions based on decision accuracy $F(7, 32) = 0.88$, $p = .53$; decision confidence $F(7, 32) = 9.8$, $p = .01$ agreement with the AI $F(7, 32) = 0.67$, $p = .70$ and response time $F(7, 32) = 2.2$, $p = .035$. Unsurprisingly due to the low number of participants, there are no significant differences, nevertheless, the informative nature of the pilot study allows us to draw some observations. In Table 1, except the *NL + CS*** variant, the loafing variants show higher agreement with the AI, lower decision accuracy and response time. We attributed the relatively poor performance in the *L-based* variants and faster response time as an element of algorithmic loafing. Based on the precision and recall in Table 1, we conducted a form of post-hoc analysis to examine performance along the correct and incorrect predictions from the AI.

3.3.1.1.1. Main takeaway. The main takeaway from the pilot study includes:

- effect of loafing primers: although we did not observe any statistically significant variation across the pilot variants, the loafing (*L*) variant shows low decision accuracy and higher agreement with the AI. Such poor performance and high agreement with the AI can be attributed to algorithmic loafing. The performance is relatively higher when both the loafing and non-loafing primers have been increased from 10 to 15. However, we believe that 10 is a better representation of the loafing primer since 15 questions seem to result in fatigue, particularly for the non-loafing cases. At this juncture, it will be relevant to be able to identify or distinguish a scenario with a trade-off between fatigue and the manifestation of loafing.
- exposure to correct and incorrect predictions: seeing multiple cases of correct decisions by an AI model improved human performance. However, a behaviour resembling loafing is observed when identifying incorrect cases under the loafing variants. Judging by the relative mean response time per question across the variants, the

Table 1
Summary of the relevant metrics across the 8 variants used for the pilot study consisting of L: variant with 10 loafing primers; NL: variant with no loafing primers, CS: confidence score, and response time. The double *asterisks* ** represent variants with increased loafing and no loafing primers from 10 to 15.

		Decision Accuracy	Precision	Recall	Agreement with AI	Response Time
Variants	L + CS	0.58 ± .49	.39	.73	0.74 ± .44	0.34 ± .20
	NL + CS	0.66 ± .48	.46	.73	0.70 ± .46	0.41 ± .23
	L	**0.44** ± .50	.29	.60	**0.84** ± .37	0.30 ± .20
	NL	0.58 ± .49	.38	.67	0.78 ± .42	0.41 ± .30
	L + CS**	0.60 ± .50	.40	.67	0.76 ± .43	0.43 ± .34
	NL + CS**	0.58 ± .50	.38	.60	0.78 ± .42	0.27 ± .14
	L**	0.66 ± .49	.46	.67	0.70 ± .46	0.32 ± .21
	NL**	0.58 ± .50	.38	.60	0.70 ± .46	0.42 ± .57

participants are more efficient in the loafing case but perform poorly in detecting incorrect cases. Because the non-loafing variant enables participants to engage with both sides of the decision outcomes, we hypothesise that it helps in developing a more structured and encompassing mental model for solving the problem. On average, the variant with the confidence score tends to outperform the variant without the confidence in terms of response time, which we consider as a proxy for deep thought on decision-making and not an indication of random selection.

Noting the marginal improvement in performance in the variant with confidence score and the fact that previous study reports that showing the AI's confidence score results in improved human performance and trust Zhang et al. (2020), we chose to utilise the variant with confidence score for our subsequent experiments.

4. Methods

In this section, we describe the main research method including the participants' recruitment process, task description, payment structure and validation process. We explore the idea of algorithmic loafing in a high-stake domain using real-world data that is known to be racially biased Julia et al. (2016). Our goal was to quantitatively and qualitatively assess how social loafing behaviour manifests in human-AI teams. To accomplish this goal, we conducted a series of user studies to gather relevant information.

4.1. Experimental conditions

As described earlier, we use the pilot study results to provide a preamble set of training questions for the following experimental conditions (summarised in Fig. 1) to tackle the above research questions and hypotheses.

(1) **H**: *human only* is the variant in which the participants work alone without any clue or support from the AI model.
(2) **P**: *prediction only* is the variant in which the participants are supported by an AI model only. The H and P variants serve as controls.
(3) **E**: *AI + explanation* variant consists of the AI model's prediction and explanation for each recommendation to support the user.
(4) **A**: *AI + approaver* variant consists of the model's recommendation, explanation and approval from past participants. The approver aspect is based on the mean agreement from past participants on the same task (based on the P variant).
(5) **L**: *loafing only* variant involves the act of inducing algorithm loafing via task repetition that always yields the correct recommendation based on the ground-truth data.
(6) **LR**: *loafing + reward system* variant is the same as the L variant, but with the inclusion of a reward system to inform the participants about the cumulative points they gain based on their performance. A point is gained or deducted for a correct or incorrect decision, respectively. The LR variant is motivated by earlier findings from loafing research that participants whose outputs could be evaluated perform better (Harkins, 1987).
(7) **LE**: *loafing + explanation* variant is similar to the L variant but with additional explanations about the recommendation presented to the user.
(8) **LA**: *loafing + approaver* variant consists of both induced loafing and approval from past participants.

4.1.1. Procedures and participants

To answer the above questions, we ran a set of user studies in which participants were asked to respond to AI's decisions about bail recommendation based on the dependant's risk score. The research survey was designed using *Qualtrics*[2] and the participants were recruited via *Prolific*.[3] For the pilot study, we recruited $n = 40$ participants involving 63%, 35% males within the age range of 19–47 years. Similarly, we recruited $n = 199$ participants for the main experiment involving 56% females, 44% males. The age group ranges from 18 to 68 years. See Table 2 for further details about demographics. Before the data collection stage, we received ethical approval from our School's Research Ethics Board for our studies.

4.1.1.1. Survey structure. After a brief introduction and explanation of the task to undertake, the participants were asked to respond to a set of questions regarding a specific experimental condition from a prediction model trained on the COMPAS system (Julia et al., 2016) dataset. The dataset is widely used in studies on fairness within the domain of criminal justice (Grgic-Hlaca, Redmiles, Gummadi, & Weller, 2018; Deeks, 2019; Dodge et al., 2019; Mothilal, Sharma, & Tan, 2020).

4.1.1.1.1. Decision scenario. The task environment is designed to provide relevant information to aid the participants. We begin with a description of the task (including the project goals and the contact person) for participants to consent to. Only consented participants will proceed and engage with attention-check questions to measure attentiveness. The participants then go through some examples of the main survey questions to get acquainted with. The participants were evenly redistributed to partake in one of the survey variants (see Fig. 1).

A typical scenario involves an AI machine learning model trained using relevant information about a pretrial defendant to recommend whether a person should be granted bail (low risk) or not (high risk). All questions in our survey evolved around what recommendation a user would provide to a specific defendant. See Fig. 2 for an example of a typical scenario encountered by the participants. Each test question includes a recommendation from the AI model and, where applicable, explanations. We included one attention check question to ensure a good quality of responses. At the end, participants were asked some broader questions to understand their propensity to social loafing (Karau & Williams, 1993). See Fig. 1 for a summary of the survey variants Fig. 3.

4.1.1.1.2. Task completion. The task took about 20–25 min to complete. After review and validation of the submitted responses, we disbursed payment to the participants within 48 h of task completion. We ensured that a participant would take part in a single experimental condition only. For each experiment, we collected the following data:

(1) participant's decision about agreeing or disagreeing with the AI's recommendation
(2) participant's confidence in their decision
(3) time it took to respond to each decision task
(4) feedback from the participants about the decision process
(5) response to general questions about the propensity to social loafing

Table 2
Demographic information about the research participants ($n = 239$) for both the pilot and main studies.

Study	Gender		Age (years)		Location	
	Female	Male	Min.	Max.	UK and USA	Other Countries
Pilot	63%	35%	19	47	40%	60%
Main	56%	44%	18	68	–	–

[2] https://www.qualtrics.com/uk/.
[3] a web service that recruits participants to complete online tasks https://app.prolific.co/.

Table 3
Survey questions and reliability analysis for all constructs in the study using Cronbach's alpha (α). The alpha value for all the aggregated constructs is 0.87

Construct:	*Social Loafing Tendency*	
Measure:	1. I will accept AI model's recommendation (Acceptance) 2. I will accept the model's output or prediction even if the explanation or decision process is unclear (Needing No Explanation). 3. How likely do you think you will agree with the AI model for future task (AI for Future Use)?	α = 0.61
Construct:	*Social Facilitation*	
Measure:	1. I believe my effort will be recognised when working with an AI-assisted system (Effort Recognition). 2. I will accept the AI model's recommendation only when convinced about the explanation or decision process (Needing Explanation). 3. How useful do you perceive the AI to be in supporting your decisions (Trust in AI)?	α = 0.60

The numerical responses have been collected using a continuous 5-point Likert scale.

4.2. Evaluation metrics

We apply the following metrics to assess the general performance and motivation of the participants across the experimental conditions.

4.2.1. Performance and loafing measures

For the performance measure, we rely on the following relevant evaluation metrics:

4.2.1.1. User performance. Measures the percentage of decision *accuracy* of participants measured against the ground-truth dataset which is based on the Compass System (Julia et al., 2016). We also measure the proportion of positive cases that were correct (*precision*) and the proportion of actual positive cases that were identified correctly (*recall*) by each participant. Moreover, we measure each participant's degree of *agreement* with AI's recommendation irrespective of whether the prediction is right or wrong. When dealing with the less accurate model, the desirable performance from a participant would be high accuracy and less agreement with the AI.

4.2.1.2. Propensity to loafing. To explore potential algorithmic loafing, we modified and applied some of the questions proposed in Hoffman et al. (2018) to collect relevant responses from the participants. Using the performance and response time metrics Wise and Kong (2005); Schnipke and Scrams (1997) will be crucial in identifying loafing. Thus, quick response time coupled with poor performance is a strong loafing indicator.

4.2.2. Motivation and loafing

Motivation in human-AI teams is associated with task completion time and effectiveness, and past studies have applied useful evaluation techniques (Buçinca et al., 2021; Wise & Kong, 2005; Touré-Tillery & Fishbach, 2014). Moreover, behavioural data involving accuracy, mean response times, and response time distributions are essential elements of cognitive processing. In cognitive processing, stimulus difficulty affects the quality of information on which a decision is based (Ratcliff & McKoon, 2008). Recognising that both strong and low motivation can result in speedy completion of a task Touré-Tillery and Fishbach (2014), we define and apply an aggregate metric involving task completion time and user performance (decision accuracy) in quantifying motivation. Essentially, we focus on the response time effort (RTE), which quantifies

Table 4
Reliability analysis for an individual item ($n = 77$ per variable) using Cronbach's α with item dropping for the loafing and no loafing constructs. The overall alpha values for the variants with the loafing and no loafing primers are $\alpha = 0.7$ and $\alpha = 0.67$, respectively.

Variable	Mean Value		**Cronbach's α**		Normality Test (W)	
	With Primers	No Primers	With Primers	No Primers	With Primers	No Primers
Confidence	3.4	3.4	0.59	0.72	0.95	0.95
Acceptance	2.9	2.9	0.59	0.66	0.96	0.95
Effort Recognition	3.2	3.1	0.55	0.66	0.96	0.96
Trust in AI	2.9	2.8	0.55	0.62	0.97	0.98
Needing Explanation	3.8	3.7	0.68	0.72	0.93	0.92
Needing No Explanation	1.6	1.5	0.65	0.70	0.93	0.89
AI for Future Use	3.5	3.3	0.61	0.63	0.92	0.90

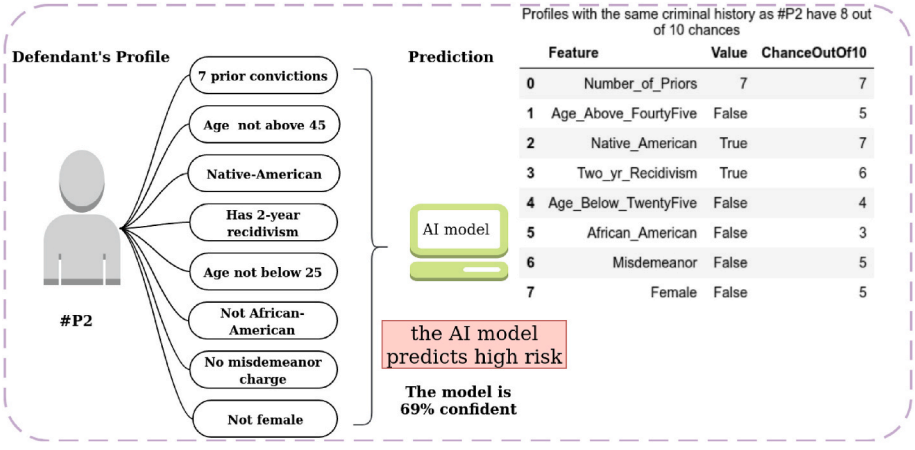


Fig. 2. A sample survey scenario for a single defendant (#P2) indicating that the AI model is 69% confident that the defendant (#P2) is high risk. The right sub-figure explains the influence of each feature in the decision process. For instance, about 70% of defendants with 7 prior convictions have been correctly predicted to be high risk.

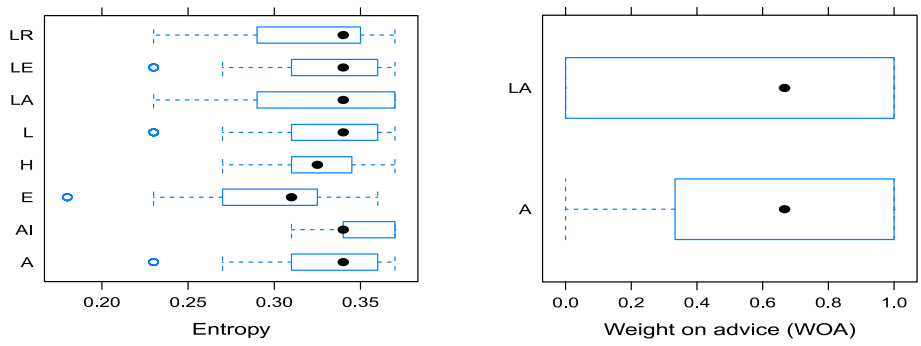


Fig. 3. Entropy and weight on advice on aggregate cases.

the amount of effort an individual devotes to an activity.

The RTE suggests that an unmotivated examinee will respond quickly in completing a task without engaging in too much cognitive process. Of interest here is to distinguish between solution behaviour and rapid-guessing behaviour (Schnipke & Scrams, 1997; Wise & Kong, 2005). For a simple two-choice decision, an increase in decision difficulty results in higher mean response time and lower accuracy (Schwarz, 2001). The solution behaviour (SB) is the time taken to select the correct decision. This is often associated with longer response time since different options need to be evaluated. On the other hand, rapid-guessing (RG) behaviour results in quick, often poor, decision-making without a thorough consideration of the available options. In this study, we compute the two behaviours by factoring:

(1) the proportion of questions correctly answered by the participant that take a relatively longer time to respond (greater than the mean response time for correct cases)
(2) the proportion of decisions or questions incorrectly answered by the participant that take a short time to respond (less than the mean response time for incorrect cases)
(3) the mean response time for both correct and incorrect responses across the experimental conditions

The response time effort is computed at both individual and group levels to identify SB and RG behaviours. To achieve that, for item i there is a threshold T_i representing the response time boundary between RB and SB. We consider the average time it took a participant to decide as the threshold for the group.

4.2.2.1. Weight on advice. In a team set-up, many factors can influence decision-making and performance. For instance, a piece of advice and its presentation can influence a user's acceptance (Dalal & Bonaccio, 2010). To quantify the weight on advice (WOA) given by past participants under the P variant in developing the A and LA variants, we leverage the judge-advisor framework (Scott & Bruce, 1995). These variants are aimed at determining the effect of a validation mechanism. The validation aspect is crafted in the form of advice to the participants based on past decisions on similar case studies. Following the approach in Logg et al. (2019), the WOA is defined as the difference between an *initial judgement (IJ)* and a *revised judgement (RJ)* divided by the difference between the *initial judgement (IJ)* and the *decision advice (DA)* given by:

$$WOA = \frac{IJ - RJ}{IJ - DA}$$

For our use case, *IJ* is the initial recommendation from the AI, *DA* is the advice from the approvers and *RJ* is the final decision by the participants. These quantities were summed over all the questions attempted by each participant in the applicable variants (A and LA) (Table 5 see Fig. 3).

5. Results

To begin, we apply Cronbach's alpha to determine the reliability of the constructs given in Table 3 used in examining social loafing and facilitation tendencies. The Cronbach's alpha is useful in measuring the internal consistencies among the responses under each construct (Cronbach, 1951). This is required so that the set of questions under each construct is aligned. A value of $\alpha \geq 0.7$ suggests that each experimental construct is reliable and consistent. Table 4 shows the reliability analysis for the individual items. Similarly, Table 5 shows the reliability analysis involving the performance construct provided in Table 3. The aggregate alpha value for the constructs in Table 3 falls short of the desired threshold. Besides the control variants (H and P), there are two broad groups consisting of variants with loafing primers (L, LA, LE, LR) and variants without loafing primers (A, E, H, P) in the study. To determine whether there is any significant difference across the study conditions, we begin with a one-way anova on all the variants (Table 6) to ascertain statistical significance. Except for P and E variants showing marginal significance, there are no significant differences across the variants after the adjustment using the conservative *Bonferroni* approach.

Table 5
Reliability analysis for individual items using Cronbach's α with item dropping for the performance construct. The overall alpha for variants with primers is $\alpha = 0.95$ and variants without primers is $\alpha = 0.8$.

Variable	Mean Value		**Cronbach's α**		Normality Test (W)	
	With Primers	No Primers	With Primers	No Primers	With Primers	No Primers
Decision Accuracy (all cases)	0.64	0.54	0.99	0.68	0.83	0.82
Decision Accuracy (correct cases)	0.79	0.76	0.93	0.78	0.85	0.82
Decision Accuracy (incorrect cases)	0.80	0.39	0.94	0.72	0.72	0.78
Agreement with AI (all cases)	0.69	0.52	0.93	0.86	0.83	0.63
Agreement with AI (correct cases)	0.79	0.76	0.93	0.78	0.85	0.82
Agreement with AI (incorrect cases)	0.62	0.58	0.94	0.74	0.76	0.82

Table 6
Pairwise comparisons with adjusted *p* value using the *Bonferroni* method for the paired conditions involving decision accuracy and agreement with the AI.

		AI Only (**P**)	**A**	**E**	**H**	**LA**	**LE**	**Accuracy**	**Agreement with AI**
Variants	Approver (**A**)	1.00	–	–	–	–	–	0.54 ± .50	0.68 ± .50
	Explanation (**E**)	**0.046**	1.00	–	–	–	–	**0.60** ± .49	**0.68** ± .47
	Human (**H**)	0.2170	1.00	1.00	–	–	–	0.54 ± .49	NA
	Loafing + Approver (**LA**)	1.00	1.00	1.00	1.00	–	–	0.53 ± .50	**0.70** ± .46
	Loafing + Explanation (**LE**)	1.00	1.00	1.00	1.00	1.00	–	**0.54** ± .51	**0.72** ± .45
	Loafing (**L**)	1.00	1.00	1.00	1.00	1.00	1.00	0.53 ± .50	0.67 ± .47
	Loafing + Reward (**LR**)	1.00	1.00	1.00	1.00	1.00	1.00	**0.54** ± .51	**0.65** ± .48

5.1. Effect of induced loafing

We study algorithmic loafing manifestation and mitigation strategies through the loafing only L, loafing + explanation LE, loafing + approver LA and loafing + reward LR variants (Table 7). These variants are in response to the research question 2 and **H1** presented earlier.

5.1.1. User performance

Performance-wise, both the pilot and main studies have shown good performance in terms of participant's ability to identify correct decisions by the AI model when loafing primers are presented. The correct feedback from the AI in the loafing primer helps the human user to identify relevant patterns in the decision process. Seeing a set of correct decisions is instrumental in distinguishing it from incorrect decision, hence the improved performance in identifying correct decisions (see Table 7 and Fig. 7). Although we did not observe any statistical significance, closer observation reveals participants under the loafing variant perform poorly compared to the non-loafing counterpart. Judging by the relative poor performance and high agreement with the AI, we suspect algorithmic loafing to be responsible. This is more pronounced in the task of identifying incorrect cases under the loafing variants (see Fig. 6 and 7).

The performance of participants under the LR is on par with the LE. We have seen how explanation alone improves performance, but the inclusion of loafing primers results in performance drop. On the other hand, the reward system, despite having the loafing primers, is instrumental in improving performance and neutralising the potential negative effect of loafing primers. The inclusion of the reward system aligns well with the notion of task visibility because the participants get instant feedback about their visible effort in the decision process. Therefore, leveraging the strategies of both task visibility Kidwell and Bennett (1993); Huang and Fu (2013) and controllability Maier and Seligman (1976); Latané et al. (1979) will result in a more effective human-AI team.

5.1.2. Agreement with the AI

Table 6 also reports the average agreement with the AI's decision across the variants. Because the AI model is about 25% wrong, a high degree of disagreement yields the best result. Conversely, a high degree of agreement with the AI often results in poor performance and is more pronounced under the loafing dimension (variants preceded by L) and the P variant (Table 7). This variant also shows less thought or analytical process as demonstrated in Table 8. However, the poor performance in P (Fig. 6) could also be attributed to confusion in understanding how the model works since there is limited support being offered to the human user. Noting the values for precision and recall in Table 7, we perform the same analysis on the correct and incorrect cases (Fig. 7). The participants under the H variant are on par with the E variant counterpart suggesting that human users are quite effective at detecting incorrect cases even without AI. This could be related to the fact that the dataset used for the training has been used to flag algorithmic unfairness making it easier for keen participants to identify problems in the process even without any input from the AI. In comparison with the H variant, the loafing variants seem to amplify the AI's error due to the poor performance in identifying incorrect decisions. The loafing (L) and approver (A) variants could lead to bias and indifference in situations where attentiveness or thoughtfulness is expected to ensure algorithmic fairness.

5.2. Effect of explanation and validation

The results and discussion in this section are in response to research question 3 and the applicable hypothesis.

5.2.1. User performance

The goal here is to determine the effect of increasing transparency through explanations on performance. Using a hybrid form of explanation consisting of both global and local explanations. While global explanation describes the overall logic of a model, local explanation explains the rationale for a single prediction (Ribeiro et al., 2016). For our case, we focus on the features capable of conveying the intuitive description of the underlying decision process. Accordingly, we selected the top 5 most important features (out of the 9 total features, see example in Fig. 2) from a trained AI model. In addition to the above explanation styles, we found the use of a *chance* feature that indicates the chances of an instance belonging to a given class (Zhang et al., 2020) to be relevant. Thus, we utilise the training sample to compute the value as a function of the percentage of dependant with attribute-value on a scale of 0 *to N* ($N = 10$) (no risk to very high risk). Noting that humans prefer frequencies over probabilities Lai and Tan (2019), the percentages were multiplied by 10 and rounded to the nearest whole number. According to the results in Table 7, there is no evidence of explanations inducing or promoting loafing behaviour, but some degree of improved performance.

The involvement of an approver or a validation mechanism in both the A and LA variants is to determine its effect on performance and mitigating social loafing in human-AI teams. As noted earlier, the

Table 7
Relevant metrics consisting of decision accuracy, precision, recall, specificity and agreement with the AI's recommendation.

		Accuracy	Precision	Recall	Specificity	Agreement with the AI
Variants	Approver (**A**)	0.54 ± .50	.39	.66	.48	0.68 ± .50
	Explanation (**E**)	**0.60** ± .49	.44	.70	.55	**0.68** ± .47
	AI (**P**)	0.53 ± .50	.36	.66	.40	0.68 ± .50
	Human (**H**)	0.54 ± .49	.36	.48	.58	NA
	Loafing + Approver (**LA**)	0.53 ± .50	.39	.72	.44	**0.70** ± .46
	Loafing + Explanation (**LE**)	**0.54** ± .51	.39	.67	.47	**0.72** ± .45
	Loafing (**L**)	0.53 ± .50	.40	.70	.47	0.67 ± .47
	Loafing + Reward (**LR**)	**0.54** ± .51	.38	.57	.53	**0.65** ± .48

Table 8
Proportion of the main themes inferred from the participants comments across all the variants – loafing + explanation (LE), explanation only (E), approver only (A), loafing + approver (LA), loafing only (L), human only (H), AI only (P) and loafing + reward system (LR). The μ **value** denotes the mean proportion of each theme across the variants.

Theme	μ value	LE	E	A	LA	L	H	P	LR
Analytics	32%	32%	**44%**	24%	**40%**	12%	**48%**	12%	44%
Trust	17%	**24%**	16%	4%	12%	**36%**	8%	**32%**	0%
Skeptics	18%	12%	**24%**	**32%**	**20%**	16%	4%	**20%**	12%
Heuristic	13%	4%	4%	0%	8%	12%	**32%**	**36%**	4%
Tiebreaker	13%	**24%**	12%	12%	**16%**	12%	**0%**	**0%**	28%
Others	4%	4%	0%	8%	4%	4%	8%	0%	0%
No Comment	3%	0%	0%	20%	0%	8%	0%	0%	8%

validation aspect is crafted in the form of advice, quantified using the judge-advisor framework Scott and Bruce (1995), to the participants based on past decisions on similar case studies by participants under the P (AI only) condition. A WOA value of 0% reveals that the participant did not heed the advice (see Fig. 8). Participants under the approver (A) variant tend to rely more on the advice rather than engaging with the task to ascertain correctness. Seeing a set of correct decisions in the LA variant supports the participants to disregard the advice.

Generally, the participants under both LA and A variants weigh the advice high, but the rating is higher under the A variant. Some of the loafing primers appear to result in discounting the advice. Task difficulty is often a factor in deciding whether to value the advice or not. The mean performance is considered as a useful proxy, and is slightly higher in the LA variant, especially towards identifying correct performance (Fig. 7).

5.2.2. Motivation and effort

We have noted how motivation in human-AI teams is associated with the time it takes for a task to be completed. Also, the solution behaviour is often associated with a longer response time since more time is required to consider and evaluate various options. Similarly, the rapid-guessing behaviour results in quick decision-making without a thorough consideration of the available options. In Figs. 4 and 5, participants under the E variant expend too much time to arrive at a solution; this is also true for the L and H variants. Surprisingly, the approver variant (A) shows the longest time. We expect this to be lower since the participants have been advised about the past decision for each decision scenario. There is not much variation in terms of rapid-guessing behaviour compared to the solution behaviour across the variants (Figs. 4 and 5). The maximum response time for SB is about 2.7 min and the minimum response time of about 10 s for the random-guessing behaviour.

5.3. Propensity to loafing

Fig. 9 summarised the participants' self-reported responses about the propensity to loafing. To identify the conditions likely to result in exercising loafing, we split the responses according to variants into two broad groups: variants with loafing primers (L, LA, LE, LR) and those without loafing primers (H, P, A, and E). Generally, all the ratings are higher across the variants, however, the approver variant shows the lowest rating which could be attributed loafing tendency.

5.3.1. Decision consensus

Through a post-hoc analysis, we observe a high variability in performance when correct/incorrect decisions have been isolated. Using relevant scores and performance measures such as confidence and perceived social loafing (Fig. 9), it is possible to quantify decision consensus across the variants at various levels. The consensus in this context is based on the dynamism of agreement with the AI's decision irrespective of correctness using the Normalised Shannon Entropy (NSE) method (Alston, Kearl, & Vaughan, 1992; Grgic-Hlaca et al., 2018). In NSE, the consensus is quantified using the relation $c = 1 - NSE$ where NSE is the normalised entropy and c is the consensus with 1 and 0 denoting total agreement and disagreement with the AI's decision, respectively. It can then be useful to think of Shannon's entropy as a measure of disagreement, where high entropy (unpredictability) implies high disagreement.

Considering each study variant as a team, the NSE is useful in gaining insight into team performance and determining individual differences in

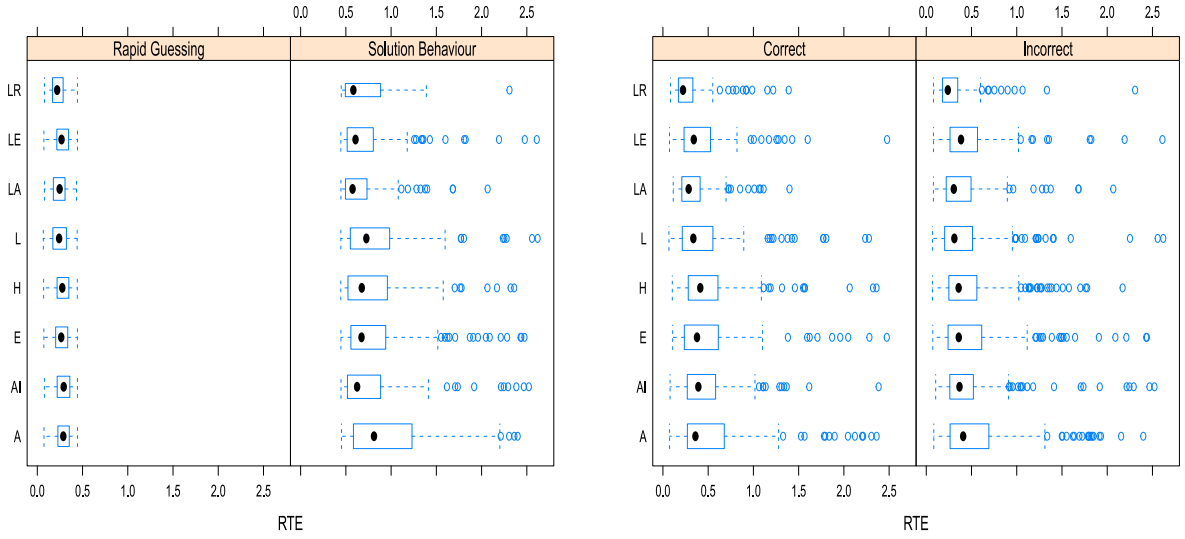


Fig. 4. Human motivation and response time across the study variants. We analyze the response time at individual and group levels for all responses (correct responses only and incorrect responses only) using the RTE to measure the cognitive effort of the participants using solution behaviour (SB) and rapid-guessing behaviour (RG).

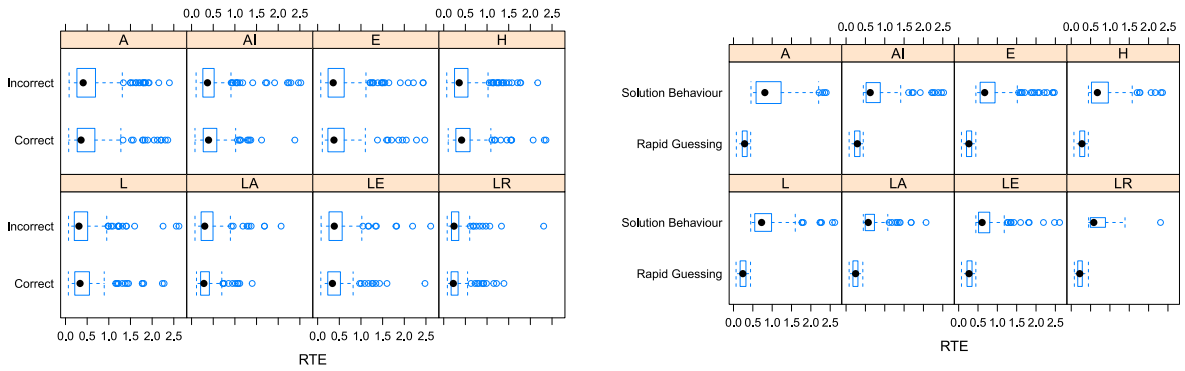


Fig. 5. Response time according to correct vs incorrect cases.

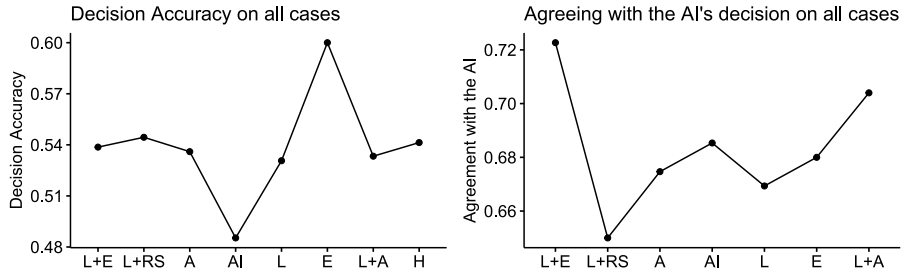


Fig. 6. Correct and incorrect cases across all variants.

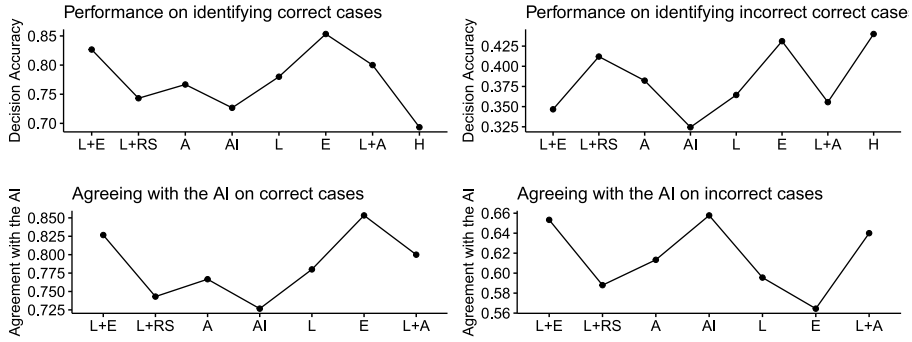


Fig. 7. Correct and incorrect cases.

performance. In Fig. 8, the mid-points of the responses vary more towards the first quartile (Q1) 25% in most variants. The high dispersion under the LA, LR and L variants signifies high unpredictability, which could be attributed to the loafing primers. For variants with good performance, such as the E variant, there is a low degree of unpredictability, suggesting a more structured response across the team or variant. We also conducted a correlation test involving the variants and noted the following. There is a significant disagreement with the AI prediction in the P (AI only) variant. This is also true for the variants with loafing primers and validation. The agreement has a strong positive correlation with accuracy, but some discrepancies in the incorrect cases exist.

5.3.2. Qualitative analysis

To identify relevant themes in the participants' comments, we manually parse and aggregated the comments from the participants into the following themes:

- **analytics**: this category consists of participants who strive to make sense of the decision process using the available features and information. This group is more likely to challenge and question the rationality of the decision process. Such attitudes could be leveraged as a form of reinforcement towards improving AI systems in general.
- **trust**: this group mostly agrees with the AI system, even if the decision process seems opaque; the participants are more receptive and unlikely to question or probe the AI system.
- **skeptics**: on the other extreme, this group views the decision process and related explanations by the AI as insufficient to trust the recommendation.
- **heuristic**: this group consists of participants who mostly rely on the information given at the beginning of the survey and the case-specific information to infer relevant clues to guide their decisions.
- **tiebreaker**: this group consists of participants who rely on AI to reach a conclusive decision when the chances for low-risk or high-risk recommendations are somewhat equally likely.

Table 8 presents a summary of the relevant themes and their distributions across the study conditions, and Table 9 highlights some of the comments from the participants and the corresponding matching

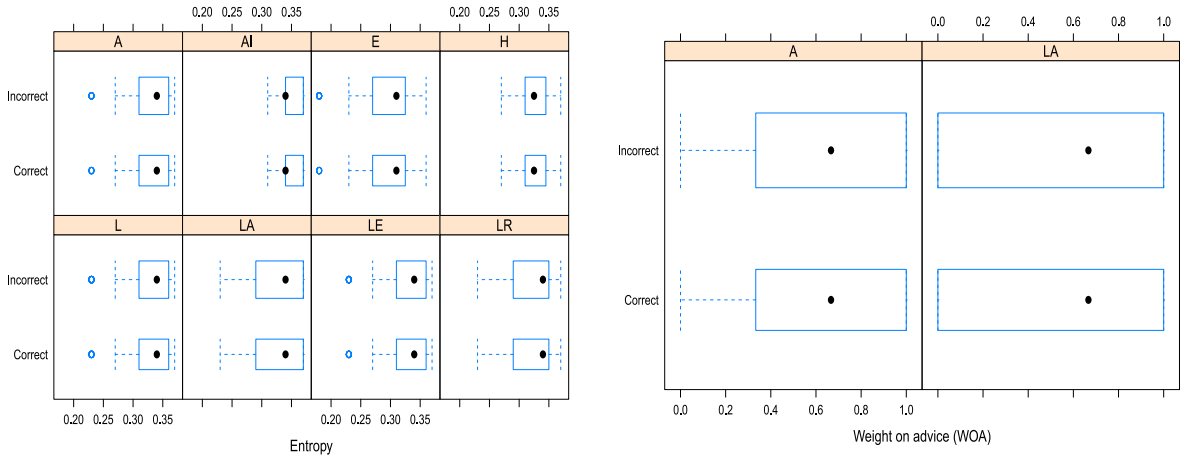


Fig. 8. Probability distributions over the participants' responses to measure agreement with AI (consensus) when the model errs. For each variant, a high entropy suggests less clarity and a high likelihood of agreeing with the AI's suggestion. The involvement of an approver is a form of validating the decision process and the weight on advice (WOA) is used to quantify how much the advice is taken by participants under the L and LA variants.

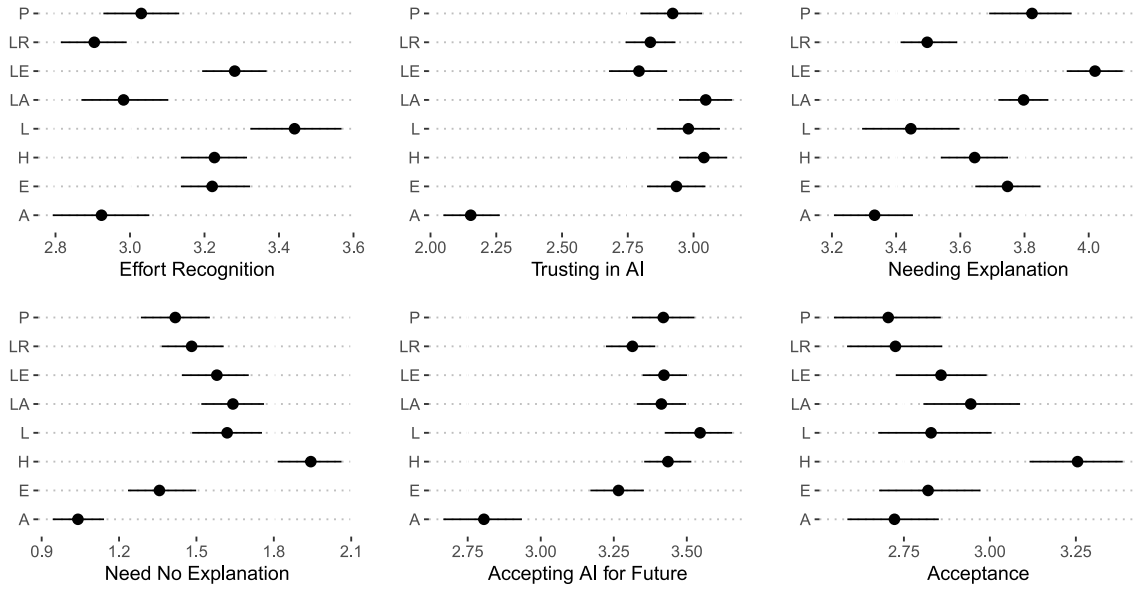


Fig. 9. A summary of the self-reported metrics on trust and need for explanation in recommendations by an AI system.

themes.

In Table 8, the human only (H) variant relies heavily on an analytical approach (about 48%) to reach a decision. This is followed by the explanation variant (E) with 44%. Because participants in the H variant have no recourse to decision support besides the basic description given at the start of the survey, they tend to be more attentive and engaging. Surprisingly, the LA variant shows a high degree of analytics compared to the remaining variants, especially loafing-only (L) and AI only (P) variants. Both L and P variants reported high degrees of trust (36% and 32%, respectively). While participants under the LE and LA variants are more likely to utilise the AI model as a tiebreaker when the possible outcomes are equally likely, participants under the approver only (A) variant expressed the highest degree of skepticism. This could be attributed to the lack of sufficient information to make an informed decision. The skepticism is useful especially when the domain requires attentiveness as in criminal justice or healthcare.

6. Discussion

In this section, we discuss the main findings of the study.

6.1. Algorithmic loafing and mitigation strategies

Traditionally, social loafing manifests in a team or group setting (Harkins, 1987; Karau & Williams, 1993; Kidwell & Bennett, 1993). With the growing need for human-AI teams, we surmise loafing could affect performance and prevent social facilitation. Social facilitation can be enhanced by factoring uncontrollability issue which causes lethargic and depressed feeling when confronted with tasks in which little or no control over the outcomes can be exercised (Maier & Seligman, 1976). Thus, the lack of control over the outcome of a process is considered to be instrumental in social loafing behaviour (Latané et al., 1979). Noting how loafing primers with a reward system (LR) and explanation (E) respectively improve human performance in identifying correct and incorrect decisions, this insight can be leveraged towards ensuring social facilitation in human-AI teams. This is in line with past results from

Table 9
Some relevant themes and sample comments from the participants. See Section 5.3.2 for details about the qualitative analysis and the description of the main themes.

Theme	*Analytics*
Comments	– I tried to look and analyze the probability of their history for the prediction of their future behavior and sometimes I felt like the AI was more strict in some cases and more easygoing in the others.so I decided to first decide by myself and then try to see why AI answered a case same or differently. I think AI is a very effective tool to use in different areas but it would be better if it would be additive to some human judgements too [sic].
Theme	*Trust*
Comments	– *I relied heavily on the AI predictions. I also looked closely at the percentages when determining how much confidence I had in the answers I gave. The AI has complete objectivity when arriving at its answers; there is nothing subjective or personal in its decisions. I therefore trust the AI more than I would trust a human opinion* [sic].
Theme	*Skeptics*
Comments	– I base off of their profile and after that I asses de AI prediction and take into consideration the statistics it gives in the explanation. I think AI has a long way to go however, I would somewhat trust it and take into consideration its input [sic].
Theme	*heuristic*
Comments	– I'll be honest I take all of it into account but I look at the previous charges especially. If someone has done something MANY times, they're bound to continue doing it. If they've only done it a couple and are above the age of 25, maybe they made a couple mistakes when young or had some tough times. So I give them a pass on that. it just makes the most sense to me. I sometimes used the AI as if it were a second opinion. Which I suppose the AI is a second opinion. I just don't feel like people's cases should be in the control of a robot [sic].
Theme	*tiebreaker*
Comments	– The AI was more helpful in the cases where perpetrators were toeing the line of being low risk or high risk, but in some cases, the answer was easy and AI was not utilized. Because AI is still wrong 24% of the time, I need to confirm that it is right [sic].
Theme	*others*
Comments	– In this moment is kinda difficult to believe totally in a computer [sic].

loafing research in which participants whose outputs could be evaluated outperformed those whose outputs could not be (Harkins, 1987). The following considerations will be instrumental in mitigating algorithmic loafing and improving social facilitation in human-AI teams.

6.1.1. Task visibility

Although we did not observe any statistically significant results, including loafing primers increasing the ability to identify correct decisions because the AI's feedback helps the human user to identify some patterns in the decision process. In human-AI teams, task visibility could be reinforced through a dynamic reward system where the performance of the human user is made visible in the team. The idea of a reward system aligns well with the notion of task visibility (Huang & Fu, 2013; Kidwell & Bennett, 1993) because the participants get instant feedback about their visible effort in the decision system. This will help improve human performance and neutralise the potential negative effects of loafing. On the other hand, a combination of both loafing and approver appears to harm the notion of task visibility by obscuring the human effort in the team.

6.1.2. Complementary team and user input

Involving the human user in the decision-making, especially in high-stakes domains, to be an active agent rather than a passive agent will be crucial. The human user should be expected to dynamically contribute and control certain aspects to complement the decision process. Thus, leveraging the notions of both task visibility and controllability will ensure a more effective and responsible human-AI team. Another important consideration is reducing algorithmic loafing likely to emanate from insufficient knowledge about the task or tool by the human user. Improving the user's expertise or knowledge about the process is crucial (Lai et al., 2023). Therefore, providing some form of training about the specific task beforehand will be useful. This is in line with past studies on improving performance and trust by improving the user agency (Chandrasekaran, Prabhu, Yadav, Chattopadhyay, & Parikh, 2018; Kulesza, Stumpf, Burnett, & Kwan, 2012; Hoffman et al., 2018; Lai, Liu, & Tan, 2020). Moreover, other useful methods to improve user agency include allowing and incorporating user feedback about the predictions (Feng & Boyd-Graber, 2019; Kulesza et al., 2012; Lee, Jain, Cha, Ojha, & Kusbit, 2019; Smith-Renner et al., 2020).

7. Conclusion

The increasing application of AI-supported decisions, especially in high-stakes domains, necessitates useful means of evaluating human attentiveness in human-AI teams. One of the key considerations in such a team is human expectation, i.e. how humans perceive the AI's utility in decision-making. A high expectation may result in algorithmic loafing leading to less attentiveness and poor performance from humans. Similarly, low expectations could lead to mistrust, rendering the AI less useful as a decision-support tool. Through a series of user studies ($n = 239$), we explored relevant scenarios to test the two possible expectations in human-AI teams. Our approach is based on the premise that algorithmic loafing in human-AI teams can be induced through task repetition that always yields the correct outcome. We also studied whether explanations, incentives and validation from external approvers influence people's ability to pay less attention or exhibit algorithmic loafing behaviour. For loafing variants, we find that participants perform better in identifying correct decisions, but perform poorly in identifying incorrect decisions. The inclusion of a reward system in the decision process could prevent algorithmic loafing. The reward system is to incentivise the process and avoid false decisions.

8. Limitations and future work

Despite the contributions and insights about algorithmic loafing, the study has some limitations. Essentially, future work could focus on addressing the following:

- **data and participants.** The data is limited because the participants come from various disciplines resulting in high variances in the data due to varying degrees of expertise. Under this scenario, it might be difficult to ascertain how representative the responses are. It is important to distinguish loafing from insufficient knowledge about the decision task and technical knowledge. Therefore, future studies would consider both expert users and laypersons and assess the human factor. Moreover, focusing on a single high-stake domain to explore algorithmic loafing should be broadened to ascertain the conditions for which our findings hold.
- **loafing framework and evaluation metrics.** It is crucial to formalise and identify standard metrics for loafing in the context of human-AI teams. This should include some concrete evaluation measures that will be applicable or accepted across domains.

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 work.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in our manuscript.

Acknowledgement

This research work was initiated under the Scottish Informatics & Computer Alliance (SICSA) Remote Collaboration Activities when the first author was working at the University of St Andrews, UK. We would like to thank the SICSA for the partial funding of the research work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2023.100024.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access, 6*, 52138–52160.
Albar, F. M., & Jetter, A. J. (2009). Heuristics in decision making. In *PICMET'09-2009 portland international conference on management of engineering & technology* (pp. 578–584). IEEE.
Alston, R. M., Kearl, J. R., & Vaughan, M. B. (1992). Is there a consensus among economists in the 1990's? *The American Economic Review, 82*, 203–209.
Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., et al. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine, 37*, 76–83.
Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*.
Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al. (2021). Does the whole exceed its parts? The effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–16).
Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., et al. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 648–657).
Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems* (pp. 1–14).
Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1–21.
Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics* (pp. 160–169). IEEE.
Calisto, F. M., Fernandes, J., Morais, M., Santiago, C., Abrantes, J. M., Nunes, N., et al. (2023). Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–20).
Calisto, F. M., Nunes, N., & Nascimento, J. C. (2022). Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies, 168*, Article 102922.
Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 3995–4004).
Chandrasekaran, A., Prabhu, V., Yadav, D., Chattopadhyay, P., & Parikh, D. (2018). *Do explanations make vqa models more predictable to a human?*. arXiv preprint arXiv:1810.12366.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
Curtis, D. D., & Lawson, M. J. (2001). Exploring collaborative online learning. *Journal of Asynchronous Learning Networks, 5*, 21–34.
Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes, 112*, 11–23.
Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review, 119*, 1829–1850.
Diogo, P., Morais, M., Calisto, F. M., Santiago, C., Aleluia, C., & Nascimento, J. C. (2023). Weakly-supervised diagnosis and detection of breast cancer using deep multiple instance learning. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)* (pp. 1–4). IEEE.
Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 275–285).
Dua, D., Graff, C., et al. (2017). *Uci machine learning repository*.
Feng, S., & Boyd-Graber, J. (2019). What can ai do for me? Evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 229–239).
Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–25.
Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference* (pp. 903–912).
Gunning, D. (2017). *Explainable artificial intelligence (xai)*. darpa.
Harkins, S. G. (1987). Social loafing and social facilitation. *Journal of Experimental Social Psychology, 23*, 1–18.
Hase, P., & Bansal, M. (2020). *Evaluating explainable ai: Which algorithmic explanations help users predict model behavior?*, Article 01831. arXiv preprint arXiv:2005.
Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: Theoretical foundations. *IEEE Intelligent Systems, 32*, 68–73.
Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for explainable ai: Challenges and prospects*. arXiv preprint arXiv:1812.04608.
Huang, S. W., & Fu, W. T. (2013). Don't hide in the crowd! increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 621–630).
Julia, A., Jeff, L., Surya, M., & Lauren, K. (2016). *Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks*. Online.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*, 681.
Kidwell, R. E., Jr., & Bennett, N. (1993). Employee propensity to withhold effort: A conceptual model to intersect three avenues of research. *Academy of Management Review, 18*, 429–456.
Kravitz, D. A., & Martin, B. (1986). *Ringelmann rediscovered: The original article*.
Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1–10).
Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1369–1385).
Lai, V., Liu, H., & Tan, C. (2020). Why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13).
Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 29–38).
Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37*, 822.
Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–26.
Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1–45.
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.
Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–16).
Maier, S. F., & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General, 105*, 3.
Morais, M., Calisto, F. M., Santiago, C., Aleluia, C., & Nascimento, J. C. (2023). Classification of breast cancer in mri with multimodal fusion. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)* (pp. 1–4). IEEE.
Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607–617).
Nigatu, H. H., Pickoff-White, L., Canny, J., & Chasins, S. (2023). Co-designing for transparency: Lessons from building a document organization tool in the criminal justice domain. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1463–1478).
Piezon, S. L., & Ferree, W. D. (2008). Perceptions of social loafing in online learning groups: A study of public university and us naval war college students. *International Review of Research in Open and Distance Learning, 9*.
Ragoonaden, K., & Bordeleau, P. (2000). Collaborative learning via the internet. *Journal of Educational Technology & Society, 3*, 361–372.
Rastogi, C. (2023). Investigating the relative strengths of humans and machine learning in decision-making. In *Proceedings of the 2023 AAAI/ACM conference on AI, Ethics, and society* (pp. 987–989).
Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 6*, 1–22.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.
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).
Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213–232.
Schwarz, W. (2001). The ex-wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers, 33*, 457–469.
Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement, 55*, 818–831.

Siemon, D., & Wank, F. (2021). In *Collaboration with ai-based teammates-evaluation of the social loafing effect* (p. 146). PACIS.
Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., & Perer, A. (2023). Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–18).
Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., et al. (2020). No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13).
Stuart, H. C., Dabbish, L., Kiesler, S., Kinnaird, P., & Kang, R. (2012). Social transparency in networked information exchange: A theoretical framework. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 451–460).
Touré-Tillery, M., & Fishbach, A. (2014). How to measure motivation: A guide for the experimental social psychologist. *Social and Personality Psychology Compass, 8*, 328–341.
Tsai, C. H., You, Y., Gui, X., Kou, Y., & Carroll, J. M. (2021). Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–17).
Vodrahalli, K., Gerstenberg, T., & Zou, J. (2021). *Do humans trust advice more if it comes from ai? An analysis of human-ai interactions.* arXiv preprint arXiv:2107.07015.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163–183.
Xu, C., Lien, K. C., & Höllerer, T. (2023). Comparing zealous and restrained ai recommendations in a real-world human-ai collaboration task. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–15).
Zajonc, R. B. (1965). Social facilitation: A solution is suggested for an old unresolved social psychological problem. *Science, 149*, 269–274.
Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 295–305).