Influence of AI and Human Feedback in Moral Decision-Making: a Mouse-Tracking Perspective

Alex Presa Hughes (202409807@post.au.dk)

School of Communication and Cognition, University of Aarhus,

Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Abstract

The effects of generative AI on human moral decision-making (Moral-DM) is becoming an increasingly hot topic with the recent advent and wide distribution of AI models over the last few years – generative AI is being adopted on an unprecedented scale ('Artificial Intelligence (AI) Usage Statistics 2025 | Global AI Users', 2025). In the present study, the influence on Moral-DM in a simulated war-like battlefield scenario was analyzed, with three trial types (Control vs. Aggressive AI vs. Conservative AI), participants were shown a moral dilemma, asked to decide whether to attack, followed by the AI input depending on the trial type, and finally they had to decide whether to attack again. It was found that participants' decisions were significantly influenced by the AI in terms of attack rates, increased reaction time (RT) when deciding whether to

Alex Presa Hughes
Aarhus University

attack the second time (possibly representing a reconciliation of their own views with the AI's) and an effect of AI on mind changes was found when AI conflicted with the participants' decision. Particularly, significance was found for aggressive AI, meaning that participants specifically switched from not attacking to attacking due to AI recommendations. This result is in accordance with previous findings (Salatino et al., 2025). No significance was achieved with mouse-tracking (MT) measurements, suggesting it is complicated to make MT reflect live moral decisions. These findings have implications for the design and implementation of AI-assisted decision-making processes in moral contexts, including the use of MT in moral decision-making and may help understand the role AI might play in a world with increasing AI presence.

Introduction

Increasing bulks of research are investigating the topic of the influence of Artificial Intelligence on human decision making in many contexts, including decision making in the field of morality and ethics, namely Moral-DM. A clear example often cited is that of AI automated cars, where an algorithm independently decides how to maneuver a vehicle, potentially having to decide whether to save car passengers versus pedestrians, a tough moral scenario (Bonnefon et al., 2016). Another example is that of LAWs (lethal automated weapons), which can identify and attack targets solely following algorithms (*Army of None | Paul Scharre*, 2018). For this reason, the

Alex Presa Hughes
Aarhus University

intersection between AI and moral decision-making is an increasingly important topic. Whilst some research has focused on peoples' decisions with AI as an advisor, there is no available research directly investigating the influence of AI in decision-making with the use of mouse-tracking to reveal new insights of how AI can modulate peoples' choices, beliefs and opinions, resulting in a noticeable gap in the literature. Particularly, mouse-tracking can offer a new way to analyze live mental processes and hesitations (Maldonado et al., 2019) that people experience when presented with moral dilemmas, and particularly to see how people change their mind.

One relevant paper, which the current experiment draws a lot from, investigated the influence of an AI advisor on human decision-making in hard situations, specifically looking into RTs, sense of agency (SoA) and responsibility in the case of moral choices (Salatino et al., 2025). The experiment consisted of a simulated battlefield with the participants operating drones and deciding whether to attack with varying levels of civil victims, infrastructure damage, and enemy damage. There were three types of trials: moral decision (a hard moral situation), attack (favouring attack) and no attack (favouring no attack), and three conditions of AI: no AI use, aggressive AI (to encourage attack) and conservative AI (to encourage no attack). The results indicate that people found it hard to answer in the hard moral condition, represented in longer RTs. This suggests that participants were in a moral dilemma. Participants were influenced by the AI in the hard moral conditions, which also resulted in lower levels of SoA, whilst having high levels of Temporal Binding (TB), meaning that they did feel like the cause of the

Alex Presa Hughes
Aarhus University

action but did not feel as responsible. Furthermore, the specific programming of the AI model influenced peoples' decisions, as people were more prone to attack with the aggressive AI and less prone with the conservative AI, indicating that AI may have a substantial influence on human decisions. The current experiment draws on this by implementing war-like scenarios and having participants deciding whether to attack, with the hypothesis that attack rates will be higher for aggressive AI, followed by control and then conservative AI – the same as in the referenced experiment. Furthermore, the three AI conditions (control, aggressive, and conservative) will also be implemented in a within-subject fashion. To date, no study has examined how AI advice modulates *real-time decision dynamics* in moral dilemmas using mouse-tracking, nor how such dynamics predict post-decision changes of mind.

Another curious question is whether humans trust AI responses. One vein of research has focused on the so-called "algorithm aversion", whereby people tend to penalize errors by AI model more than human errors, leading to increased distrust when AI models err. However, a different general trend is so-called "algorithm appreciation", whereby people adhere to and heed algorithmic advice (Logg et al., 2019). However, instead of formulating this question as binary, it may be that trust in AI is more nuanced as Chan et al. (2020) suggests. In this experiment, in kidney intervention situations participants trusted domain experts ("expert psychologists") more than AI models, depending on the available evidence and on the information available about the training and programming of the algorithm itself. When authorship is known, people show equal

Alex Presa Hughes
Aarhus University

skepticism toward human and AI-generated content (Huschens et al., 2023). Thus, as participants will know the content is AI-generated, there should be no problems with baseline levels of trust.

ChatGPT will be used specifically for this study because of its widespread use with 5.8 billion visits in September 2025 (Singh, 2025) - with most people being familiar or active users of this AI model. In addition, one paper examined the use of ChatGPT in higher education specifically, finding that trust in ChatGPT mediates the perceived ease of use, usefulness and intelligence help explain its large adoption (Shahzad et al., 2024). This aligns well with the present experiment design, as all participants are undergraduate students. The above paper is in accordance with Noh et al. (2025) too, where it was found that ChatGPT is widely trusted due to its warmth, general high performance and human-like interaction features. Therefore, the AI agents are expected to influence participant behavior consistently across all hypotheses, affecting attack rates, RTs, MT measures, and the likelihood of changing initial decisions.

AI recommendations in the experiment will be accompanied by explanations. This is supported by Vasconcelos et al. (2023), where the researchers found that people tend to rely on AI models' judgements using some form of a cost-benefit analysis – relying less when AI make mistakes and relying more when a difficult is presented (high cognitive load) or when the explanations provided by the AI are easy to understand. It is hypothesized that in the current experiment participants will often switch to the AI's decision when conflicting, as the explanations are kept short, the AI decisions are

Alex Presa Hughes
Aarhus University

programmed to look realistic and error-free – and most importantly it reduces SoA as suggested by Salatino et al., 2025. Moreover, it is expected that participants will show faster RTs and straighter MT trajectories when AI confirms their decisions. This aligns well with Bashkirova & Krpan (2024), where healthcare professionals in diagnostic mental health were given AI recommendations to assist their judgements, and it was found that they accepted AI judgements more often when they aligned with their own, an example of "confirmation bias".

Furthermore, it is hypothesized that participants will not change to their judgements if they are reinforced by the AI and higher rates of mind change are expected be found when AI challenges participants' judgements, because ChatGPT is expected to be perceived as a reliable and knowledge-laden authority, as suggested by Shahzad et al. (2024). Finally, when people hesitate more making their first decision, represented in longer RTs and greater AUCs, will be more likely to be swayed by the AI recommendations, as they are expected to be more unsure of their decisions. As can be seen, AUC will be used as the MT measure, as it is a good index of attraction to the unselected option, representing hesitation until the final choice was made, over time steps (Freeman & Ambady, 2010).

Hypotheses

- H.1.0: There will be no difference in attack responses among trial types.

Alex Presa Hughes
Aarhus University

- H.1.1: There will be a difference in responses in control, aggressive and conservative AI trials. Specifically, aggressive AI trials will elicit higher attack rates, followed by control, and finally conservative AI.

- H.2a.0: In trials where AI confirms an participants' choice, participants will not show significantly different RTs from the first to the second choice.

- H.2a.1: In trials where AI confirms an participants' choice, participants will show significantly different RTs from the first to the second choice. Specifically, the second RTs will be faster.

- H.2b.0: In trials where AI confirms an participants' choice, participants will not show significantly different MTs from the first to the second choice.

- H.2b.1: In trials where AI confirms an participants' choice, participants will show significantly different MTs from the first to the second choice. Specifically, the MTs show less hesitation (lower AUC).

- H.3a.0: There will be no difference in mind change rates between AI-confirmation and AI-challenge trials.

- H.3a.1: Participants will change their minds more frequently when AI challenges their choice than when AI confirms it.

- H.3b.0: Initial decision hesitation (RTs and AUCs) will not predict mind changes when AI challenges choices.

Alex Presa Hughes
Aarhus University

- H.3b.1: Greater hesitation in the first decision (longer RTs, greater AUCs) will increase the likelihood of switching to AI's recommendation when it challenges the initial choice.

Methods

Participants

Twenty participants took part in the study (mean age = 22.5, SD = 2.46, 10 women, 10 men). No participants were excluded from the data analysis. However, trials considered timeouts were excluded – these are trials where participants did not answer within the 5 second time limit. To participate in the study, participants were not required to have any special knowledge. Participants were recruited at Aarhus University following an adequate gender balance (male-to-female ratio) and are all within the range of 19 to 26 years of age. Participants were recruited with no external help – only the experimenter himself was involved. None of the participants were screened beforehand for previous knowledge of combat or war-like situations either, as it was not deemed relevant to the question of interest. The only personal data collected was the participants' age and gender

Alex Presa Hughes
Aarhus University

Ethics information

The study was conducted in line with the principles of the Declaration of Helsinki. Before participation, participants were informed as to the nature of the experiment, including what they would be shown and what situation they would be in. Specifically, they were told they would be in a soldier's position in a war zone and this soldier would come across different situations with the possibility of attacking or not attacking according to the participant's own judgement. Furthermore, they were shown the purpose, procedure and conditions of the experiment, potential risks (of which there were none), the terms of data confidentiality (anonymized data used strictly for analysis purposes and to be deleted afterwards) and the right to withdrawal (without justification and without consequences). Participation was voluntary and to proceed with the study, it was necessary to accept the outlined conditions on the computer screen. Participants were naturally given the option to either accept or reject the experiment conditions. Therefore, it can be said that informed consent was given. The study was conducted at Aarhus University as part of the Cognitive Science bachelor's program. The study does not qualify as a study with medical intervention. As a result, approval from a local ethical committee was not necessary, as per Danish regulations (Lov Om Videnskabsetisk Behandling Af Sundhedsvidenskabelige Forskningsprojekter, 2011).

Alex Presa Hughes
Aarhus University

Stimuli and procedure

The datasets generated and analyzed for the current study are available online on a public GitHub repository below in this paper. The experiment was programmed and presented using PsychoPy-2025.1.1, an open-source software package suited to performing psychology experiments (Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J., 2019). The experiment was conducted on a laptop (screen resolution: 1920*1080 pixels) and responses were recorded using a QWERTY keyboard.

At the beginning of the experiment procedure, basic personal information, i.e. age and gender, was collected in a tab on the computer screen, and participants were assigned a unique 6-digit ID. Participants were welcomed to the experiment and were briefly told about the decision-making experiment with mouse-tracking they were about to partake in. They were then presented with the consent form, outlining the purpose, procedure, conditions and conditions of the experiment, with the choice to either accept or reject the conditions. If participants pressed reject (the key ‘N’) they would exit the experiment, and if they accepted (the key ‘Y’), they would proceed. Due to the importance of the 5 second time limit to capture meaningful mouse-tracking data, participants were once more informed they would have to act within the allotted time frame when making a decision.

The experiment loop contained fifteen trials - each one being a different war-related dilemma with possible civilian casualties and enemy troop presence. Some examples of

Alex Presa Hughes
Aarhus University

the dilemmas shown to the participants are included in the table figure below. All the information displayed in the experiment was contained within a csv file that was looped through, including the display of AI recommendations and explanations. Therefore, no AI API was embedded in the experiment code. Furthermore, the conditions (control, aggressive AI and conservative AI) and the war situations shown were randomized. For each participant, 15 rows out of the 30 rows in the csv file were selected, each one with a different situation. This was done to have a greater variety of war situations. The situations and AI explanations used in the experiment were AI-generated, specifically by the latest ChatGPT AI model as of December 2025 (ChatGPT) (Open AI, 2025).

Situation	Option 1	Option 2	Explanation conservative	Explanation aggressive
Enemy mortar team is firing from positions 200m from a village. 5-15 civilians likely remain in their homes.	Attack	No Attack	When civilians are likely present military action risks unacceptable harm	Mortar teams cause continuous friendly casualties and must be neutralized
Enemy command post is confirmed in a residential building. 3-8 civilians may still be sheltering inside.	Attack	No Attack	Occupied residential buildings cannot be targeted without confirmed evacuation	Command posts coordinate enemy operations and represent high-value targets
Enemy supply convoy is moving along a road. 2-10 civilian vehicles are mixed with military trucks.	Attack	No Attack	Civilian vehicles intermixed create unacceptable collateral risk	Supply convoys sustain enemy combat capability and are time-critical targets

Alex Presa Hughes
Aarhus University

Enemy snipers are positioned in an apartment building. Some apartments appear occupied others empty.	Attack	No Attack	Occupied residential units prohibit building-wide military action	Sniper positions inflict disproportionate casualties and create tactical problems
Enemy artillery battery is near a school. School status is unclear - could be empty or have 10-30 children.	Attack	No Attack	Schools must be presumed occupied during daytime hours	Artillery positions shelling friendly forces require immediate response

Figure 1 – Table representing the first five rows of the csv file used in the experiment set-up.

In each case, a three-second countdown to commence the round was presented to make sure the participants were active and ready. After that, the dilemma was shown and participants would have to read it and press any computer key to continue to the choice selection. After another three-second countdown, participants were shown two options equally distant from the starting point, with one being “attack” and the other “no attack” across all trials. They were given a five-second window to hover over to their desired choice, otherwise the trial would be counted as a timeout and would later on be excluded from data analysis. The purpose of implementing a time window to collect mouse-tracking responses is to make the mouse trajectories reflect real-time decision processes, rather than the participants having a long time while leaving the mouse idle.

Alex Presa Hughes
Aarhus University

After their first choice (attack or no attack), participants were shown one out of three possibilities on screen, depending on the condition assigned to the specific trial, as this was a within-subject experiment. For the control condition, they would see a message saying “No ChatGPT recommendation available. Continue to final choice.”. For both AI conditions, they would see a loading message animation saying “ChatGPT is analyzing…” and “Please wait until the response has loaded, then press any key to continue.”. After a simulated time interval, the response would load following the following structure “ChatGPT recommends (option) because (explanation)”. The difference between both AI conditions is that the conservative AI would always recommend “no attack” and provide an explanation for not attacking, whereas the aggressive AI would always recommend “attack” and provide an explanation for attacking. Participants would then have to press any key to continue.

Another three-second countdown was shown before the second-choice selection, where participants would see “attack” and “no attack” again. They had to choose one of these two options again, with the aim of seeing if they change their minds, show different mouse movements or reaction times, among other measures. The five-second time limit also applied in this round, and any timeouts in the second round were also to be excluded.

Alex Presa Hughes
Aarhus University

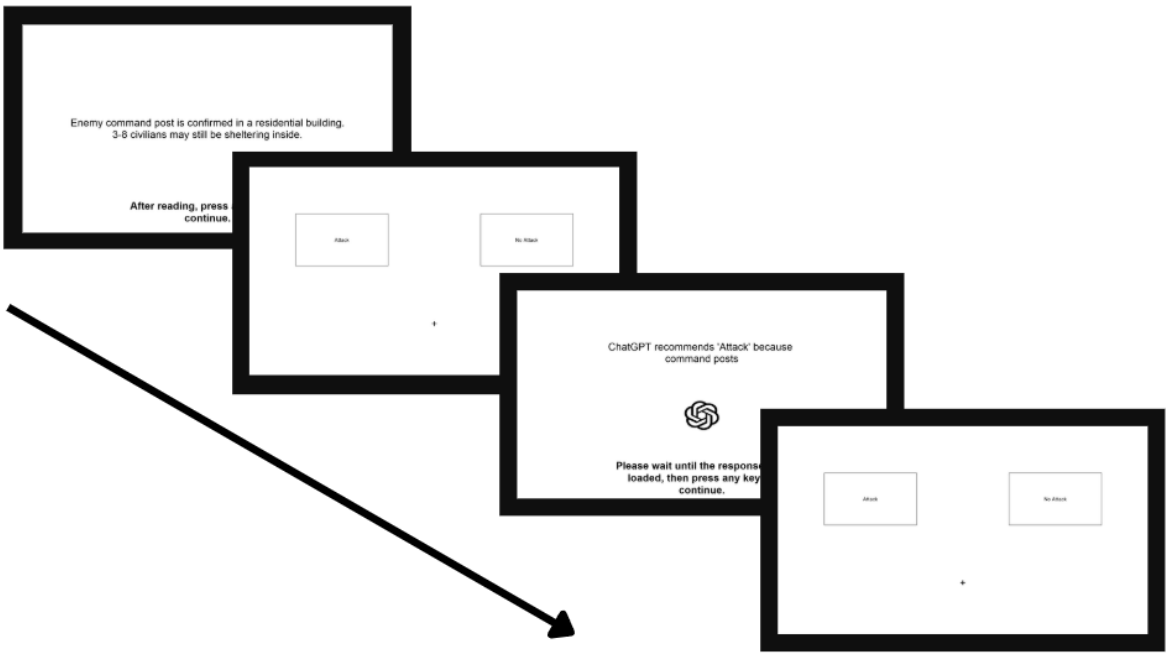


Figure 2 – Experiment set-up (chronological sequence of stimuli in each trial). Each trial consisted of a written situation, with different civilian damage, enemy damage and strategic importance. Participants were asked to decide whether to attack or not the described target in the description of the situation. Then they were either shown no AI recommendation (control), an attack recommendation by AI (aggressive AI shown on screen) or a no-attack recommendation by AI (conservative AI). Participants were instructed to decide whether to attack or not again. In both cases, when deciding whether to attack, a five-second time limit was imposed to collect reliable and meaningful mouse-tracking data.

Within the PsychoPy software, different packages were used; specifically, *core* (to set clocks and exit procedures), *visual* (to create windows, text and image stimuli), *event* (to wait for key presses and mouse-tracking) and *gui* (to collect participant data).

Alex Presa Hughes
Aarhus University

Furthermore, the *random* package was used to generate random digits for the participant IDs, and to randomize condition and trial order. The *string* package was used to generate both letters and numbers for the IDs. The *pandas* package was imported to handle data using data frames (Version 2.2.3; The pandas development team, 2025). Finally, the *os* package was used to save the collected data.

Analysis

Two dependent variables were used in this study, which are RTs and MT measurements, specifically encoded as AUC. There were two RT and AUC measures for when participants made their first and second attack choices, namely before and after the AI recommendations. The AUC in the study was computed by

The RStudio software was used to conduct the analysis (Posit team, 2025). The relevant csv file with the data collected from the experiment performed in the PsychoPy software. The data preparation was the first thing to be done. In this regard, any RTs classified as timeouts (5 seconds) were excluded. In total, out of the 300 columns with observations 5 were classified as timeouts – 2 during the first choice and 3 during the second choice, resulting in 295 valid rows with observations. Afterwards, the RT variables were analyzed to see if they were normally distributed, which they were not ($p < .05$). Then, as the experiment follows a within-subject design, the RTs were standardized within participants, so that all participants start from the same baseline

Alex Presa Hughes
Aarhus University

and individual differences do not come into play. The “Choice_1” and “Choice_2” columns were converted into binary numeric values – 1 standing for attack and 0 for no attack. The significance threshold for all statistical tests was 0.05. For logistic regression in R, the glmer function from the lme4 (v.1.1-35.3; Bates et al., 2024). package was used.

To test H.1, namely if the attack responses vary by trial type, a logistic regression model was run with the second choice as the outcome and condition as the predictor, with a random intercept for each participant due to repeated measures. In addition, a more complex model was explored, adding the standardized reaction time for the first round (continuous), the first choice (categorical), gender (categorical) and a random intercept for each war situation. Additionally, the *bobyqa* optimizer was used to help the model converge due to the many predictors. This was done to get a more nuanced picture of potentially influencing factors.

To test H.2a, only so-called confirmation trials were kept, namely AI conditions where the AI reinforces the participant’s first choice. The differences in RTs before and after AI input were tested for normality – they ended up not being normal (W = 0.92, p = 3.12x10^-11). First, a paired t-test to compare the RTs was run, but due to the violation of normality a Wilcoxon signed rank test was run. With regards to H.2b, movement trajectories were analyzed by calculating the deviation from a straight-line path (AUC), excluding the initial stationary period, i.e. only measurements after movement was detected were included. The difference in AUC was not normal (W = 0.80844, p =

Alex Presa Hughes
Aarhus University

2.633e-10), so a Wilcoxon test as run. Delving into mouse latencies, the time of the start of movement was calculated and was compared across AI conditions using a non-parametric test, the Kruskal-Wallis test.

For H.3a, the trials where participants changed their mind both when AI challenged and confirmed their decision were extracted, and the mind change ratios were calculated. After, a Chi-square test was run, as all assumptions were met. For H.3b, challenge trials where participants changed their mind were isolated, and the mouse trajectories were analyzed using AUC, removing data points before movement occurred. A logistic regression model was run predicting mind changes using RT, AUC, AI condition and a random intercept for participants.

Results

The first logistic regression model for H.1 is the following:

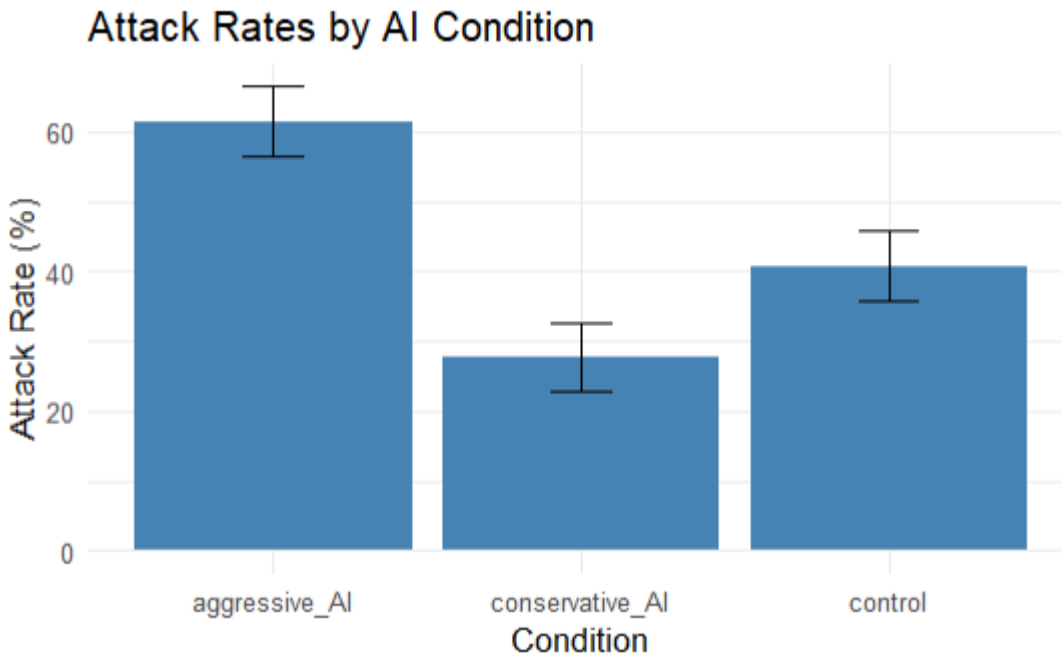
$$Choice\ 2 \sim Condition + (1 \mid ID)$$

This model revealed a significant effect of conservative AI (β = -0.66, SE = .31, z = -2.076, p = .037), aggressive AI (β = .83, SE = .3, z = 2.74, p = .006), but not of the intercept (β = -0.35, SE = .23, z = -1.5, p = .13). The second logistic regression model for H.1 is this:

$$Choice\ 2 \sim Condition + std\ RT\ 1 + Choice\ 1 + Gender + (1 \mid ID)\ +\ (1 \mid Situation)$$

Alex Presa Hughes
Aarhus University

This model shows a significant effect of the intercept (β = -0.66, SE = .31, z = -2.076, p = .037), conservative AI (β = -0.66, SE = .31, z = -2.076, p = .037), aggressive AI (β = -0.66, SE = .31, z = -2.076, p = .037) and the first choice (β = -0.66, SE = .31, z = -2.076, p = .037). However, no significance was found for standardized reaction times and gender, suggesting that longer or shorter time latencies making their first decision and gender differences are not predictive of participants' decisions. Rather, what seems to modulate participants' decisions is their first choice and the AI manipulation, with the attack rates in the aggressive AI condition being 61.45%, in the control condition being 40.81% and in the conservative AI condition being 27.72%. These findings support H.1.1, with the AI recommending attack and the AI recommending no attack resulting in higher and lower attack rates, respectively.



Alex Presa Hughes
Aarhus University

Figure 3 – Plot run in RStudio showing the difference in attack rates per AI condition with error bars.

In regard to H.2a, the paired t-test comparing RTs for the first and second choice, suggests that RT when participants were making their second decision (after AI input) were significantly longer t(103) = 2.59, p = .011, 95% CI [0.03, 0.24], with a mean increase of 0.14 seconds. However, the differences in RT are not normally distributed, so a non-parametric alternative was used, a Wilcoxon signed rank test with continuity correction. This test also suggested a significant difference in RTs (*V* = 3670, p = .002, pseudo-median difference = 0.114 s, 95% CI [0.041, 0.191]), meaning that on average second responses were 114 milliseconds longer. This rejects H.2a.0 and also H.2a.1 because a significant difference was found but it is in the opposite direction. Concerning H.2b, the Wilcoxon signed rank test with continuity correction shows no significant difference (V = 2852, p = .69), meaning that mouse trajectories and deviations are homogenous before and after AI input, supporting H.2b.0. A Kruskal-Wallis test was conducted comparing MT initiation across AI conditions. The results suggest there was also no significant difference as follows: H(2) = .58, p = .75.

Moving on to H.3a, the Chi-square test revealed a significant difference in mind changes between challenge and confirmation trials (X^2 = 32.43, p = 1.23e-08), supporting H.3a.1. Furthermore, in challenge and confirmation trials participants changed their mind 31.18% and 0.96% of the time, respectively. Regarding H.3b, the logistic regression model only revealed a significant intercept, the aggressive AI, (β = -

Alex Presa Hughes
Aarhus University

1.39, SE = .31, z = -1.96, p = .049), whilst RT (β = .15, SE = .63, z = .24, p = .80), AUC (β = .019, SE = .031, z = .62, p = .53) and conservative AI (β = .14, SE = .52, z = .27, p = .78) were not found to be significant predictors of mind change, supporting H.3b.0. But this suggests that the aggressive AI was the main driving force behind mind changes, with people switching from no attack as their first decision to attack as their second decision when seeing the attack recommendation.

Discussion

This study investigated the effects of AI as an advisor in moral decision-making, following previous research in the field. However, this study incorporated a novel MT approach in this field. Using a task in which participants were drone operators and had to decide whether to initiate an attack or not in simulated war-like scenarios, this study allowed us to explore the effect of AI behavior on the human agent in morally challenging situations. By analyzing differences between control, aggressive and conservative AI, our results showed that in moral situations decisions are significantly influenced by the AI.

According to the experiment, baseline attack behavior was more attack-averse (40.81%), but the effect of different AI types had the potential to lift the attack rates to make behavior significantly more attack-prone or even more attack-averse, with aggressive (61.45%) and conservative AI (27.72%), respectively. Furthermore, the

Alex Presa Hughes
Aarhus University

model. showed that these AI were highly significant predictors of second attack choices, alongside the intercept and the first choice in the more complex model. This supports H.1.1, suggesting a profound influence of AI modulating human decision-making. It was also found that the RTs when making their second decision after AI input were significantly longer by 144 ms, rejecting H.2a.0 as well. This is the opposite direction that was expected from H.2a.1, contradicting the initial "confirmation bias" explanation, meaning that more or other factors came into play. An exploratory analysis (Kruskall-Wallis test) showed that MT initiation was not significantly different across AI conditions. This suggests it may be an "agency gap", where people have reconciled AI input with their own moral authority – processing if they agree with the AI and if they will change their mind, because despite decreased SoA, high levels of TB are to be found - and this dissociation requires time, potentially explaining the extra RT. Furthermore, no significant differences was found in MT complexity (AUC) before and after AI input, suggesting that participants often remained idle and MT did not always reflect live decision-making. AI influence was also found in terms of mind changes through the Chi-square test, suggesting that participants changed their mind significantly more when AI contradicted their decision than when it confirmed it, namely this mind change behavior shows a clear direction, namely a change to adjust to AI behavior, further showcasing AI's potential influence and demonstrating that AI is followed, and is perceived as a reliable authority, as reviewed beforehand (Shahzad et al., 2024). The only significant predictor found of mind changes was aggressive AI, suggesting that participants

Alex Presa Hughes
Aarhus University

significantly switched from their “no attack” judgement to the AI’s recommendation of “attack”. This could be explained with the finding that in these hard moral scenarios, a decreased SoA as an intermediate mechanism could make people feel less responsible and make them more prone to the seen aggressive behavior, as this is supported by the reviewed literature, including Salatino et al., 2025.

Limitations and future directions

It must be noted that this study has limitations. First of all, this study does not allow inferences of causality – only correlational inferences through behavioral measures, so neural correlates and brain activations of causality could be established in the future. Moreover, only 20 participants were included, which could lead to a lack of power, so future studies could accumulate more power to get more powerful results.

Another issue is calibrating a time limit with MT measurements to make MT show peoples’ live decision processes – a 5 second limit was chosen for the present experiment, but other time limits could be considered, particularly focusing on making these limits shorter and making people more time-pressured. Despite this, spatial and temporal information were well recorded, and no problems were encountered when analyzing.

Another limitation is that one specific AI model was used in this experiment, ChatGPT, which may introduce its own biases in trust, responses and decisions in

Alex Presa Hughes
Aarhus University

moral scenarios. Therefore, future research could focus on using additional AI models (e.g.: Deepseek, Gemini) to see any potential differences and generalizability of results. Furthermore, more studies could be conducted using a shorter time limit to reflect live decisions. Alternatively, other MT measures could be used to run tests and models – either with new collected data or with the present data available publicly.

Another interesting direction of this research could be to analyze differences in age and the trust and influence of AI on difference populations by including a greater diversity of ages. This would be interesting, as older populations have been more neglected in the age of AI and it is hypothesized that they trust AI less generally. AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies | AI & SOCIETY

Data and code availability

The datasets and code generated and analyzed during the current study are available in a public repository in the GitHub platform under an MIT license, which can be accessed through the following link:

https://github.com/alexphughes/Influence-of-AI-and-Human-Feedback-in-Moral-Decision-Making-a-Mouse-Tracking-Perspective-

Alex Presa Hughes
Aarhus University

References

Bashkirova, A., & Krpan, D. (2024). Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. Computers in Human Behavior: Artificial Humans, 2(1), 100066. https://doi.org/10.1016/j.chbah.2024.100066

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models Usinglme4. Journal of Statistical Software, 67(1). https://doi.org/10.18637/jss.v067.i01

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Chan, L., Doyle, K., McElfresh, D., Conitzer, V., Dickerson, J. P., Schaich Borg, J., & Sinnott-Armstrong, W. (2020). Artificial Artificial Intelligence: Measuring Influence of AI 'Assessments' on Moral Decision-Making. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 214–220. https://doi.org/10.1145/3375627.3375870

Docherty, B. (2025). A Hazard to Human Rights. Human Rights Watch. https://www.hrw.org/report/2025/04/28/a-hazard-to-human-rights/autonomous-weapons-systems-and-digital-decision-making

Alex Presa Hughes
Aarhus University

He, Y., Gu, R., Deng, G., Lin, Y., Gan, T., Cui, F., Liu, C., & Luo, Y. (2024). Psychological and Brain Responses to Artificial Intelligence's Violation of Community Ethics. Cyberpsychology, Behavior, and Social Networking, 27(8), 562–570. https://doi.org/10.1089/cyber.2023.0524

Huschens, M., Briesch, M., Sobania, D., & Rothlauf, F. (2023). Do You Trust ChatGPT? -- Perceived Credibility of Human and AI-Generated Content (No. arXiv:2309.02524). arXiv. https://doi.org/10.48550/arXiv.2309.02524

Lov Om Videnskabsetisk Behandling Af Sundhedsvidenskabelige Forskningsprojekter, LOV nr 593 af 14/06/2011 (2011). https://www.retsinformation.dk/eli/lta/2011/593

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. https://doi.org/10.1016/j.obhdp.2018.12.005

Maldonado, M., Dunbar, E., & Chemla, E. (2019). Mouse tracking as a window into decision making. Behavior Research Methods, 51(3), 1085–1101. https://doi.org/10.3758/s13428-018-01194-x

McKinsey (2025). The State of AI: Global Survey 2025. Retrieved 15 December 2025, from https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Noh, H.-H., Rim, H. B., & Lee, B.-K. (2025). Exploring User Attitudes and Trust Toward ChatGPT as a Social Actor: A UTAUT-Based Analysis. Sage Open, 15(2), 21582440251345896. https://doi.org/10.1177/21582440251345896

OpenAI. (2025). ChatGPT (December 2025 version) [Large language model]. https://chat.openai.com/chat

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. Behavior Research Methods. Home — PsychoPy v2025.2.3

Posit team (2025). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

Alex Presa Hughes
Aarhus University

Python Software Foundation. (2025). Python (Version 3.14) [Computer software]. https://www.python.org/

Salatino, A., Prével, A., Caspar, E., & Lo Bue, S. (2025). Influence of AI behavior on human moral decisions, agency, and responsibility. Scientific Reports, 15(1), 12329. https://doi.org/10.1038/s41598-025-95587-6

Shahzad, M. F., Xu, S., & Javed, I. (2024). ChatGPT awareness, acceptance, and adoption in higher education: The role of trust as a cornerstone. International Journal of Educational Technology in Higher Education, 21(1), 46. https://doi.org/10.1186/s41239-024-00478-x

Singh, S. (2025, November 20). ChatGPT Users Statistics (January 2026) – Growth & Usage Data. DemandSage. https://www.demandsage.com/chatgpt-statistics/

Stypinska, J. (2023). AI ageism: A critical roadmap for studying age discrimination and exclusion in digitalized societies. AI & SOCIETY, 38(2), 665–677. https://doi.org/10.1007/s00146-022-01553-5

Alex Presa Hughes
Aarhus University

The Global Statistics (2025). Artificial Intelligence (AI) Usage Statistics 2025 | Global AI Users. (2025, October 24). https://www.theglobalstatistics.com/artificial-intelligence-ai-usage-statistics/

The pandas development team. (2025). pandas (Version 2.2.3) [Computer software]. https://pandas.pydata.org/

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–38. https://doi.org/10.1145/3579605