scientific reports

OPEN

Influence of AI behavior on human moral decisions, agency, and responsibility

Adriana Salatino[1]✉, Arthur Prével[2], Emilie Caspar[3] & Salvatore Lo Bue[1]

There is a growing interest in understanding the effects of human-machine interaction on moral decision-making (Moral-DM) and sense of agency (SoA). Here, we investigated whether the "moral behavior" of an AI may affect both moral-DM and SoA in a military population, by using a task in which cadets played the role of drone operators on a battlefield. Participants had to decide whether or not to initiate an attack based on the presence of enemies and the risk of collateral damage. By combining three different types of trials (Moral vs. two No-Morals) in three blocks with three type of intelligent system support (No-AI support vs. Aggressive-AI vs. Conservative-AI), we showed that participants' decisions in the morally challenging situations were influenced by the inputs provided by the autonomous system. Furthermore, by measuring implicit and explicit agency, we found a significant increase in the SoA at the implicit level in the morally challenging situations, and a decrease in the explicit responsibility during the interaction with both AIs. These results suggest that the AI behavior influences human moral decision-making and alters the sense of agency and responsibility in ethical scenarios. These findings have implications for the design of AI-assisted decision-making processes in moral contexts.

Keywords Human-autonomous systems interaction, Human performance, AI behavior, Moral decision-making, Sense of agency

In contemporary society, autonomous systems have permeated various aspects of human existence, revolutionizing industries, services, and even personal routines. From personal assistance and product recommendations to applications in healthcare, transportation, and defence, the spread of autonomous technologies is reshaping the landscape of human-machine interaction[1–5]. As the development and reach of Artificial Intelligence (AI) continues to advance, the integration of autonomous systems is leading to a profound examination of the dynamics of human-machine interaction. This advancement implies a departure from traditional modes of task execution, where humans rely on mechanical machines as tools to achieve goals, towards a future where highly autonomous intelligent systems undertake tasks directly. However, beyond efficiency and productivity, it is becoming increasingly important to understand the intricate interplay between humans and AI in terms of ethical decision-making and the attribution of responsibility, especially in today's world where AI agents are increasingly used as advisors and sometimes even delegates. Studies on Human-AI interaction have already attempted to shed light on the complex relationship between humans and machines. The pitfalls of this relationship include problems of complacency[6], loss of situational awareness[7], skill decay and increased workload, particularly in environments where automation failures may occur[8,9]. As human tend to react to AI as to human partners sometimes[10–12], those problems could be worse.

Research has already shown that the detrimental effects of cooperation with AI seem to correlate with the degree of autonomy of these systems[13,14], leading, for example, to a reduced number of correct decisions with a higher level of autonomy[13] and to a decrease in the sense of agency (SoA) and responsibility[14]. Of particular interest is the impact of autonomy on the human SoA, a fundamental aspect of consciousness that is closely linked to moral responsibility[14–16]. SoA, defined as the perception of causing changes in the external world through voluntary action[17,18], underpins intentional behavior and is associated with moral decision-making[19]. Recent research on this topic utilising the "Temporal Binding" (TB) effect, which refers to the subjective time compression between a voluntary action and its outcome[20,21], have provided insights into the implicit measurement of agency and have shown a decrease in implicit agency with increasing levels of automation[14,16]. The Temporal Binding (TB) effect is a well-known method for implicitly measuring the Sense of Agency (SoA) [see[22] for a review]. In classic tasks,

[1]Department of Life Sciences, Royal Military Academy, Brussels, Belgium. [2]University of Lille, CNRS, UMR 9193 – SCALab – Sciences Cognitives et Sciences Affectives, Lille, France. [3]The Moral & Social Brain Lab, Department of Experimental Psychology, Ghent University, Ghent, Belgium. ✉email: adriana.salatino@mil.be

participants have to estimate the time in milliseconds that elapsed between a keypress and outcome[23] or report the timing of the action and the outcome on a Libet clock[24]. Numerous studies have shown that time appears compressed in situations in which the person is active, while time appears stretched in situations in which the person is passive [e.g.[20–25]], suggesting it represents a proxy of the sense of agency. Explicit measurements also exist, and are usually obtained through a direct report of how people attribute the consequences of their own actions or how in control they feel[26], but these measurements are sensitive to social desirability and other biases [e.g.[27,28]]. In the present study, TB is used together with explicit self-reports to capture potentially distinct aspects of agency. Explicit measures are thought to rely on higher-level conceptual judgements of agency that are influenced by social and contextual cues[29]. In contrast, temporal binding reflects lower-level, automatic processes. Evidence suggests that these two systems operate via separate mechanisms [e.g.[30]] and that implicit and explicit measures of agency may not always align[31]. Since previous studies have shown a decrease in both implicit and explicit SoA during interaction with autonomous systems[14,32], here we aimed to explore the effects of human-autonomous systems interaction in morally challenging situations, extending the understanding of how agency is modulated under such conditions. However, despite the growing body of literature on the impact of interaction between human and autonomous systems, there are still gaps in understanding the implications for moral decision-making (i.e., decisions implying conflicting moral values) and responsibility, particularly in sensitive areas such as the military, health care or human resources management. For a military commander, a surgeon, or a responsible for the recruitment and selection of personnel, decisions can imply conflicting personal values. While previous research has primarily focused on the programming of ethical behavior of AI, little attention has been paid to the influence of human-agent interaction on moral decision-making[33–35]. Importantly, the existing evidence on the impact of autonomous systems on social and moral decision-making is mixed, with studies reporting both prosocial and unethical behavior[36–38]. Furthermore, according to common ethical and legal standards, human beings are generally deemed to remain responsible for the decisions they make and the actions they take. Therefore, it is crucial to understand the mechanisms behind the decision to accept or reject AI input[39], and how the AI behavior might influence these processes. This is especially problematic in high-stakes situations where decisions affect human lives such as self-driving cars and medical diagnosis[39–41].

Given these considerations and the importance of this topic, our study aims to investigate the effects of the interaction between humans and autonomous systems on moral decision-making and on the sense of agency and responsibility, and in particular how AI behavior might influence people's decisions. Specifically, we employed an experimental setup in which participants act as drone operators on a simulated battlefield. During the experiment, participants were confronted with morally challenging situations regarding the launch of attacks based on the presence of enemies, civilians, and infrastructure, as well as the risk that allies could also be harmed. Participants were exposed to three types of trials representing three types of uncertainty: Moral Decision-Making (Moral), Attack (A1), and No Attack (NA) Trials. The autonomy conditions included no system assistance (Level 0), an AI assistance with an "aggressive" approach, which suggest to always attack the enemy (Level 1) except during NAS trials, and an AI assistance with a "conservative" approach, which suggest never attacking the enemy (Level 2) except during AS trials. We thus manipulated the "moral behavior" of the machines to test how this influenced the moral decisions of our participants. Agency was measured both at the implicit level, using the TB paradigm, and at the explicit level through an assessment of responsibility (using an ad-hoc self-report scale)[19,42,43].

We also measure performance by using reaction time, and the proportion of trials in which participants chose to attack. By manipulating the level of system autonomy and measuring implicit and explicit indicators of agency and moral decision-making, we sought to shed light on the complicated dynamics of interaction between human and autonomous systems in morally charged contexts.

The aim of our study was to investigate the effects of interaction with AI with different moral behavior on the participants' decision-making and their sense of agency in morally demanding situations. Based on previous findings[14,16], we hypothesized that agency is lower during interaction with autonomous systems than when people make decisions alone, as indicated by a longer time estimation between action and outcome in the first case, and that judgement of responsibility is also lower after this interaction. We also hypothesized that the moral decision-making would also be affected by the interaction with autonomous systems, with a higher number of attacks during the aggressive AI-assisted situations compared to situations without AI or with the support of conservative AI[44].

As the research questions addressed in our study have received little attention in the literature, we also decided to conduct exploratory analyses of psychological factors. In doing so, we wanted to establish a relationship between moral decision-making, our measures of implicit and explicit agency and other relevant psychological variables, such as personality traits, assessed through self-report questionnaires.

This study seeks to contribute to filling the gap in the existing literature by examining the effects of human-autonomous system interaction on moral decision-making, SoA, and responsibility. By shedding light on the cognitive and behavioral implications of this interaction, we aim to provide useful insights to promote responsible and morally conscious use of autonomous technologies in various domains. Our findings have significant implications for the development and deployment of autonomous technologies, especially in areas in which conflicting fundamental values are at stake.

Method

Participants

Thirty participants took part in the study (Mage = 25.5; SD = 8.2; 6 women, 24 men). One participant was excluded due to incomplete data, so the final sample consisted of 29 participants. The sample size was estimated using G*Power[45], with a small-to-medium size effect f of 0.2, a threshold for significance α set at 0.05, and a power 1-β at 0.80. Based on these values, the estimated sample size was 28 participants. However, to compensate for potential data losses and exclusions, a total of 30 participants was targeted. To participate in the study, participants had to have notions of International Humanitarian Law (IHL), and thus about what is legally allowed and forbidden in the conduct of armed conflict. In order to fulfil this criterion, we included third- and fourth-year students of the Royal Military Academy (who were trained in IHL) and officers of the Royal Military Academy in the study. None of the participants, including the officers, had any previous experience of drone piloting as this was not a prerequisite for participation. The participants were recruited with the help of a student officer in the course of his Bachelor Thesis. The study was conducted in accordance with the principles of the Declaration of Helsinki. Participants were informed about the general purpose and duration of the experiment and about their rights as participants in psychological research before giving their consent. Participation was voluntary and participants were informed that they could withdraw their participation at any time without justification and without consequences. Written and signed informed consent was obtained from each participant after the reading of a letter of information, providing general information about the study and prior to the experiment. The study was approved by the local ethical committee of the Faculty of Psychology and Educational Sciences of Ghent University (2022/047).

Stimuli and procedures

The datasets generated and analysed during the current study are available in the Open Science Framework[46]. The experiment was programmed and presented using MATLAB 2020b and the Psychophysics Toolbox Extension. The experiment was conducted on a laptop (screen resolution: 1920*1080 pixels) and responses were recorded using an AZERTY keyboard, with the left and right arrow keys used as response keys in the experiment. The experiment was conducted in an experimental room in the Department of Life Sciences at the Royal Military Academy. Participants were seated on a chair in front of a table with the laptop screen approximately 40 cm away from them. They were informed that during the experiment they would assume the role of a drone operator fighting in a war scenario. They would defend their country and protect their fellow citizens fighting to protect their fellow citizens from a foreign invader, defend the territorial integrity of their country and help repel the invaders at their own borders. As part of a drone unit, their task was to consult reports on the areas their drone flights over, analyse a static image of the zone and decide whether a target— consisting of enemy, infrastructure and/or civilians — should be bombed (or not), being aware of the complexity of the decision based on International Humanitarian Law. Each trial of the task consisted, indeed, of an aerial view of an urbanized area, with several possible target information. Participants were asked to choose between attacking the enemy or deciding not to attack by pressing respectively the left or right arrow key on the keyboard. In addition, they were informed through written reports displayed in the right part of the screen (see more detail below and in Fig. 1) that both choices (i.e., to attack or not to attack) could result in a range of potential collateral damage (see below for more details). Participants underwent training sessions prior to the start of the experiment to ensure they fully understood the task and could perform it correctly.

In this study, we used a 3 × 3 within-subject design, with "Uncertainty" and "Autonomy" as independent variables. During the experiment, participants were tested on the three different types of trials representing three levels of "Uncertainty": Moral trials, Attack (A) trials, and No Attack (NA) trials. In all trials, a group of enemies (represented by a red dot), a group of civilians (represented by an orange dot) and/or the position of an infrastructure (represented by a blue dot) were superimposed on the aerial view in different position combinations. On the right side of the screen, next to the view, a report informed the participants about the risks and values parameters.

To define morally challenging situations, 128 situations with different risks, benefits, and costs combinations were presented in a first pre-test phase to a group of 5 expert military officers, who were asked to identify the most morally challenging possible scenarios. The situations contained combinations reporting the following parameters: (1) the strategic importance of the target (i.e., the military advantage in case of attack or the loss in case of non-attack), with either an advantage at the tactical (winning a battle), operational (winning a campaign) or strategic (winning the war) level; (2) the potential destruction of civilian objects and infrastructure, expressed as risk (0, 25, 50, 75, 100%) and value (low, medium, high value); (3) the potential loss of civilian life or civilian injury, expressed as risk (0, 25, 50, 75, 100%) and number (1, 10, 20, 30, 50, 100 + persons); (4) the potential loss of allied forces (personnel and material) expressed as risk (0, 25, 50, 75, 100%); and (5) the criticality of the consequences in case of collateral damage from the attack or if the target is not hit (tactical, operational, strategic). The combinations of the various parameters were meant to create situations of uncertainty, imposing moral demands to the decision (having to choose between probable benefits and probable costs, with various material and human values at stake). After the first selection, 30 situations were defined as not challenging at all, and 55 as the most morally demanding. Among the 30 not morally challenging situations, 15 were situations in which it was obvious that an attack was expected (high losses for the enemy, without collateral damages) and 15 and 15 in which the decision not to attack was obvious (no enemy in sight). We called the first Attack trials, and the second type No Attack trials (see Table 1 for an example of each situation).

In a second pre-test phase, the 55 most morally demanding situations were presented to a second group of 5 different expert military officers who were asked to indicate in which situation they would initiate an attack and in which they would not. As a result, for the experiment we selected 15 situations amongst the 55 most morally challenging scenarios in which at least one of our second set of five experts chose he/she would attack in that situation. We called these situations Moral Decision-Making (Moral) trials. It is important to highlight that expert military officers found it quite difficult to make decisions - mostly choosing to avoid the attack, as they decide to attack only 20% of the time - which confirms that the scenarios we have designed are indeed morally challenging. Furthermore, the participants were not informed about the decision of the expert military officers. The experiment consisted of three blocks, with each block containing a certain level of support from an intelligent system (Level 0 = No Assistance; Level 1 = Aggressive AI; Level 2 = Conservative AI), which represents

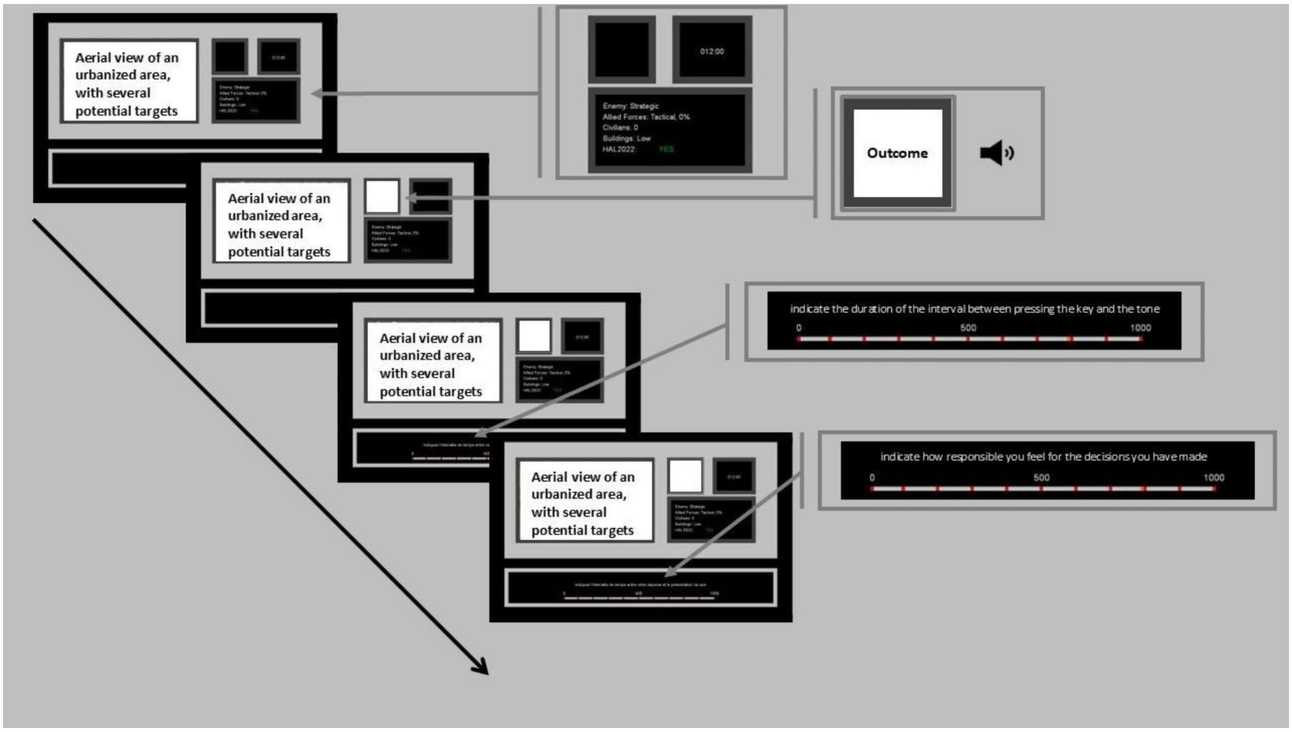
Aerial view of an urbanized area, with several potential targets
012:00
Enemy: Strategic
Allied Forces: Tactical, 0%
Civilians: 0
Buildings: Low
HAL2022 YES
Outcome
Aerial view of an urbanized area, with several potential targets
indicate the duration of the interval between pressing the key and the tone
0
500
1000
Aerial view of an urbanized area, with several potential targets
indicate how responsible you feel for the decisions you have made
0
500
1000
Aerial view of an urbanized area, with several potential targets

Fig. 1. Experiment set-up (schematic representation of the trials). Each trial of the task consisted of an aerial visual image of an urbanized area with different combinations of blue, orange, and red dots representing civilian, infrastructural and military targets, respectively (displayed for 15,000 ms or until the participant made a decision). To help participants make the decision, a report was displayed on the right-hand side with information on the risks, costs and benefits of an attack, as well as a timer display. Participants were asked to make their choice by pressing the selected response button, and soon after an image was displayed showing the outcome of the decision. To collect a Temporal Binding (TB) measurement, a horizontal line under the aerial image and an auditory tone were displayed after the responses. To obtain a judgement of responsibility a horizontal line (from – 100 to 100) was placed across the TB bar.

the variable “Autonomy”. In Level 0, the participants were not supported by any intelligent system. In Level 1, the participants were supported by a simulated AI with an “aggressive” approach, called “HAL 2022”. HAL 2022 always suggested to attack the enemy, except during NA trials. Finally, in Level 2, participants were supported by a simulated AI with a “conservative” approach, called “TAL 2022”. TAL 2022 always suggested to not attack the enemy except during A trials. Participants were not informed of the AI’s attitude but were told that they would be assisted by two different AIs, both programmed to help the operator make the best decision with respect to the International Human Law. In total, each block consisted of 15 Moral trials, 15 A trials, and 15 NA trials.

The experimental setup of this task is shown in Fig. 1: (1) First a loading bar was presented for a delay chosen randomly between 1,000 ms and 2,000 ms to signal a new trial to the participants. (2) The aerial view with different combinations of blue, red and orange dots was displayed for 15,000 ms or until the participant pressed one of the two response buttons. On the right side, the report with the information on risks and benefits and a count display was shown. (3) Participants were asked to make their choice by pressing the selected response button, and soon after an image showing the outcome of the decision was displayed in the top right corner. There were four possible outcomes. In the case where the decision was to launch an attack, two different images could be displayed: an image showing a destroyed tank, indicating that the enemy was hit, with no collateral damage or consequences for allies/civilians, or an image depicting a destroyed civilian building, indicating that the enemy was hit, but with collateral consequences for civilian buildings and/or civilian victims. In the case where the decision was to not launch an attack, two different images could be shown: an image showing that there were no direct losses for allies associated with the participant’s decision, or an image depicting allies injured by the enemy, indicating that the enemy gained an advantage from the decision not to attack. The nature of the outcome depended on the type of trial and on the risks for the different parameters. In case of attack, the presentation of a destroyed tank (i.e., enemy hit without collateral damages) was selected randomly with a probability based on the information reported on the risks of destruction of civilian objects and infrastructure (e.g., 50%) and loss of civilian life or civilian injury (e.g., 75%, then 0.875). In case of a non-attack, here the presentation of moving troops (i.e., moving enemies without allied losses) was selected randomly with a probability based on the information reported on the risk of loss of allied forces (e.g., 25%). (4) Pressing the key was followed by a horizontal line below the aerial view and a tone (frequency: 400 Hz) for 200, 500, or 800 ms. (5) Participants were asked to indicate the duration of the interval between pressing the key and the tone on a horizontal scale from 0 ms to 1000 ms (to collect a TB measurement, panel D). (6) Finally, a horizontal line from – 100 to 100

1. Moral decision-making
Importance of the target: *OPERATIONAL*
Risk for civilian objects: *75%*
Value of the civilian objects if they are hit: *Medium*
Risk for civilian persons: *50%*
Number of civilian if they are hit: *50*
Risk for military own or allied troops: *75%*
Consequence if military own or allied troops are hit: *Operational*
2. Attack sure
Importance of the target: *STRATEGIC*
Risk for civilian objects: *0*
Value of the civilian objects if they are hit: *Low*
Risk for civilian persons: *0*
Number of civilian if they are hit: *0*
Risk for military own or allied troops: *0*
Consequence if military own or allied troops are hit: *Strategic*
3. No attack
Importance of the target: *TACTICAL*
Risk for civilian objects: *100%*
Value of the civilian objects if they are hit: *High*
Risk for civilian persons: *100%*
Number of civilian if they are hit: *100+*
Risk for military own or allied troops: *100%*
Consequence if military own or allied troops are hit: *Strategic*

Table 1. Example of moral decision-making, attack and no attack situations. The table contains one example each of Moral (1), Attack (2) and No Attack (3) situation. For each of these situations are indicated: the importance of the target, the risk of destruction of civilian objects and infrastructure, the risk of loss of life or injury to civilians, the potential number of civilians if they are hit, and the criticality of the consequences in case of collateral damage from the attack or in the case of a non- attack.

appeared across the TB horizontal bar, and participants were asked to indicate on a scale from – 100 (not at all responsible) to 100 (totally responsible) how responsible they felt for the decisions they made (to obtain a judgement of responsibility - Responsibility). The trials were separated by an interval of 2,000 ms. The scenarios were presented in a random order to minimize potential order effects.

At the end of the experimental session, participants were asked to fill out four questionnaires assessing several personality traits. Those questionnaires included (1) the Levenson Self-Report Psychopathy Scale (LSRP)[47], (2) the Moral Foundation Questionnaire (MFQ)[48], (3) the Locus of CONTROL scale[47] and (4) the Checklist for Trust between People and Automation[49]. We included these questionnaires to investigate the relationship between moral decision-making, agency and personality traits. Previous research has shown that the loss of agency can affect the attribution of responsibility, potentially leading to moral disengagement from our actions[50]. Furthermore, the findings of Caspar et al.[19,43] suggest that a diminished sense of agency correlates with increased antisocial behavior. Therefore, we used the LSRP scale and the MFQ to investigate potential correlations between agency, responsibility and psychopathy (measured with LSRP) and foundational domains in moral decision-making (measured with MFQ). In addition, we used the Locus of Control Scale because, among other individual differences, the Locus of Control has been shown to be associated with ethical behavior[51,52]. It has been shown that people with an internal locus of control are more likely to see the connection between their own behavior and its consequences. They are therefore more willing to take responsibility for their ethical behavior and act according to what they believe to be right[53].

As people's trust in AI has not yet been sufficiently researched despite increasing attention, and as it is crucial to investigate it further in order to avoid wrong decisions when accepting or rejecting AI input[54], we measured people's trust in AI here using the Trust in Automation Scale. We hypothesized that higher levels of trust would lead to greater compliance with AI advice, with negative correlations with response time, sense of agency and perceived responsibility.

Measurements and analysis

We used five dependent variables in this study: Decision, Response Time, Agency, and Responsibility. Decision (A1) was expressed by the proportion of trials on which participants decided to attack (i.e., A1 choice, in percentage). Response Time (RT) was the mean response time (in seconds) on each trial. Agency was measured by Temporal Binding (TB, in milliseconds). The TB score was calculated by subtracting each interval (Int) estimate given by the subject from the actual response tone interval (either 200, 500 or 800 ms), i.e., Int_{Actual} - $Int_{Estimate}$. As (more) agency is associated with the reduction of the interval estimate compared to an actual interval (IntActual > IntEstimate), a higher TB score means higher agency and a lower TB score means lower agency. After this computation, these scores were averaged for each Uncertainty by Autonomy condition. A measure of Responsibility was obtained soon after the TB measurement, using the self-reported responsibility scale described above.

In addition, we examined the changes in TB and Responsibility based on the decision to attack or not (dependent variable A1). Lastly, to examine the relationship between the outcome and responsibility ratings, we conducted an additional analysis in which we examined how positive and negative outcomes affected participants' subjective sense of responsibility in different conditions. As the moral trials are the most interesting for our objectives and the "Attack" and "No-Attack" conditions are considered control trials, we conducted this analysis exclusively on the "Moral Decision Making" trials. Analyses were performed using JASP version 0.17.2. We performed separate repeated-measures ANOVAs for A1, RT, TB and Responsibility with Uncertainty (Moral, A, and NA) trials and Autonomy (Level 0, Level 1, Level 2) as within-subject factors. In addition, Responsibility was compared by means of a repeated measures ANOVA with Autonomy as within-subjects factors. For each dependent variable, only data of participants within +/- 2.5 SDs were considered [Only one participant for two dependent variables (RTs and subjective judgement) in the entire data set was identified as an outlier. The outliers were identified and removed before performing the main statistical analyses. However, we also analysed the data with these outliers and the results did not change compared to the original analyses, with the exception of the responsibility judgement ($p = .072$)]. Greenhouse-Geisser correction was applied where sphericity was violated.

Lastly, we conducted exploratory analyses to examine whether there is a correlation between moral decision-making, our measures of implicit and explicit agency, and other relevant psychological variables. Correlations between the Levenson Self-Report Psychopathy Scale (LSRP), Moral Foundation Questionnaire (MFQ), Locus of CONTROL scale, the Checklist for Trust between People and Automation and A1, RT, TB and Responsibility were estimated using Pearson's Correlations. Lastly, we conducted exploratory analyses to test if there is an effect on the nature of the outcome received (positive or negative outcome) on the choice made by participants at subsequent trial.

The primary focus of our analysis concerned the presence of a main effect of Uncertainty on A1, for which we expected a higher percentage of attacks during A trials, and a lower percentage during NA trials, and on RT, with expected shorter response time in the not morally challenging trials compared to the Moral trials, as evidence of the moral conflict produced by the scenarios. We also predicted that the Moral trials would result in an approximate 50% mean attack rate in the No-AI conditions, suggesting that decision- making in these trials is likely to be driven by random factors. As the scenarios selected were the most challenging situations in which participants had to weigh up competing ethical considerations given the risk to human life and the potential consequences, we expected that it would be extremely difficult to take a decision here. Regarding TB, following the results of Moretto et al.[55] we expected a main effect of Uncertainty with shorter TB, indicating an increase of sense of agency, during Moral trials in comparison with the two trials. We also expected a main effect of Autonomy on RT, congruent with the main effect of Uncertainty, on TB, with less interval binding, indicating a decrease of sense of agency, and on Responsibility, with lower Responsibility with increased level of autonomy in line with the conclusions of Berberian et al.[14].

The threshold selected for significance was $p < .05$. Raw data, scripts, and processed data can be found on the Open Science Framework.

Results

Analyses on A1

The analysis on A1 (Fig. 2, panel A) revealed a main effect of Uncertainty (F (1.14, 28.65) = 171.13, $p < .001$, $\eta p^2 = 0.87$) and post hoc tests showed that all comparisons were significant (all $p < .001$) with more A1 choices during A (mean = 93.20, SE = 1.43) in comparison with NA trials (mean = 2.63, SE = 0.54) and Moral trials (mean = 49.11, SE = 3.60). However, although it is close to the significance, the analysis revealed no significant effect of Autonomy (F (1.60, 40.07) = 3.10, $p = .070$, $\eta p^2 = 0.11$) on A1. Furthermore, a significant interaction (Fig. 2, Panel B) between Uncertainty and Autonomy was found (F (1.88, 47.01) = 7.16, $p = .002$, $\eta p^2 = 0.22$), with a simple main effect of Autonomy in Moral trials ($p = .03$) but not in other Uncertainty conditions (all $ps > 0.09$).

Analyses on RT

The analysis on RT (Fig. 3) revealed a significant effect of Uncertainty (F (1.86, 50.26) = 32.24, $p < .001$, $\eta p^2 = 0.54$) with post hoc tests showing significant longer RT during Moral trials (mean = 5.38, SE = 0.25) than during A (mean = 4.17, SE = 0.21) and NA trials (mean = 3.92, SE = 0.19), with all $p < .001$, but not between A and NA ($p = .32$). Neither significant differences between No Risk and No Enemy trials ($p = .96$), nor a significant effect of Autonomy ($p = .97$) were found.

Analyses on TB

The analysis on TB (Fig. 4) revealed a significant effect of Uncertainty (F (1.79, 48.27) = 6.78, $p = .003$, $\eta p^2 = 0.20$) with post hoc tests showing a significant ($p = .002$) difference between Moral trials (mean = 290.47, SE = 18.04) and NA trials (mean = 252.48, SE = 17.96), but not between Moral trials and *A* ($p = .29$, mean = 270.22, SE = 18.39), with Bonferroni's correction. No significant effects of Autonomy were found ($p = .63$).

Analyses on responsibility

The analysis on Responsibility (Fig. 5) revealed a significant main effect of Autonomy (F (1.81, 48.96) = 5.29, $p = .01$, $\eta p^2 = 0.16$). Post hoc tests (Bonferroni's correction) showed a significant difference between Level 0 (mean = 119.81, SE = 1.62) and Level 1 (mean = 111.60, SE = 2.32) ($p = .02$) and between Level 0 and Level 2

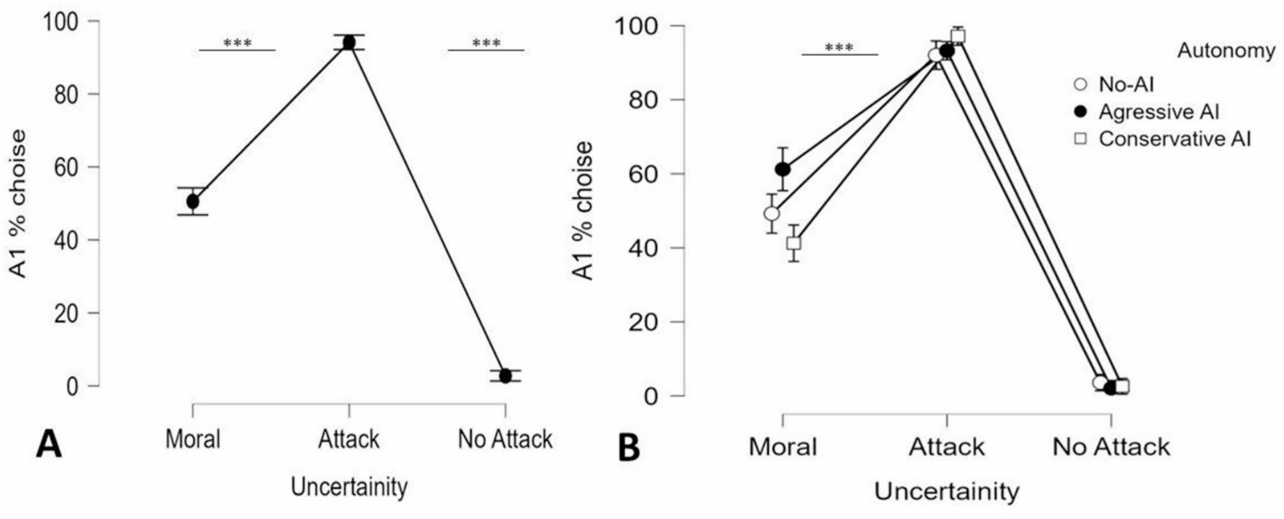

Fig. 2. Proportion of A1 action (i.e., percentage of attacks performed). During the Moral Decision-Making (Moral) trials there were significantly fewer A1 choices compared to Attack (A) ($p < .001$); in addition, the difference between A and No Attack (NA) was significant ($p < .001$, panel **A**). A significant interaction ($p < .001$) between the type of trials (i.e., Uncertainty) and Autonomy was found (panel **B**), with a simple main effect of Autonomy in Moral trials ($p < .001$). *** $p < .001$ = significant.

(mean = 113.23, SE = 2.15) ($p = .01$), with larger responsibility rating during Level 0 in both cases, but not between Level 1 and Level 2 ($p = 1$). No significant effects of Uncertainty were found ($p = .31$).

Analyses of changes in TB and responsibility based on A1 decision

The analysis indicates no significant difference based on the decision for the temporal binding variable ($p > .448$). However, a significant interaction was found between the decision and AI for the Responsibility (F (1.48, 25.14) = 5.58, $p = .02$, $\eta p^2 = 0.25$), with a significant difference "Aggressive AI" ($p = .008$). Specifically, participants reported a higher sense of responsibility when they chose not to attack compared to when they chose to attack.

Analyses of outcomes and personality scales

The ANOVA for the outcomes showed that the proportion of A1 decisions at trial t was not affected by the type of outcome presented at trial t-1, regardless of whether it was positive or negative (all *ps* > 0.391). In addition, the results show that the type of feedback does not affect the level of subjective sense of responsibility (all *ps* > 0.396). The correlation analyses between the personality scales and A1, RT, TB and Responsibility showed that none of the analyses or tested correlations were significant at $p < .05$ (all *ps* > 0.1).

Discussion

This study investigated the effects of input received from an intelligent autonomous system on moral decision-making and the sense of agency. Using an ad hoc task in which participants acted as drone operators and had to decide whether to initiate an attack or not on a simulated battlefield, this study allowed us to explore the effect of AI behavior on the human agent in morally challenging situations. By including different types of trials which comprised both moral and non-moral situations, and by involving the support of two different intelligent systems ("aggressive" vs. "conservative"), our results showed that in moral situations decisions are guided by the autonomous system.

Based on previous findings[14,16], we expected both implicit and explicit agency to decrease during interaction with autonomous systems. We also expected that moral decision-making would also be affected by the interaction with autonomous systems and that the nature of the moral value of the machines would influence participants' moral decisions, with more attacks during the interaction with the "aggressive" AI than with the "conservative" AI.

Our results were consistent with our expectations, as a significant interaction was found between trial types and autonomy, suggesting that human decision-making in morally challenging situations is influenced by the behavior of autonomous systems. Our results show that participants' moral decisions were significantly influenced by the system's input, confirming that human decisions in morally complex situations can be influenced by recommendations from a decision support system. These results are consistent with previous findings showing that AI-advisors can significantly influence human moral decision-making[35,38,56]. They are also in line with previous research in the military context [e.g.[13,57]] and confirm the influence of autonomy on human performance and decision-making. In addition, we also found an effect of trial type, as the frequency of attacks increased significantly when there was no risk and decreased when the attack was not safe.

An important consideration to take into account is the tendency for humans to rely on external cues, including social and environmental information, when faced with uncertainty[58,59], especially in complex situations

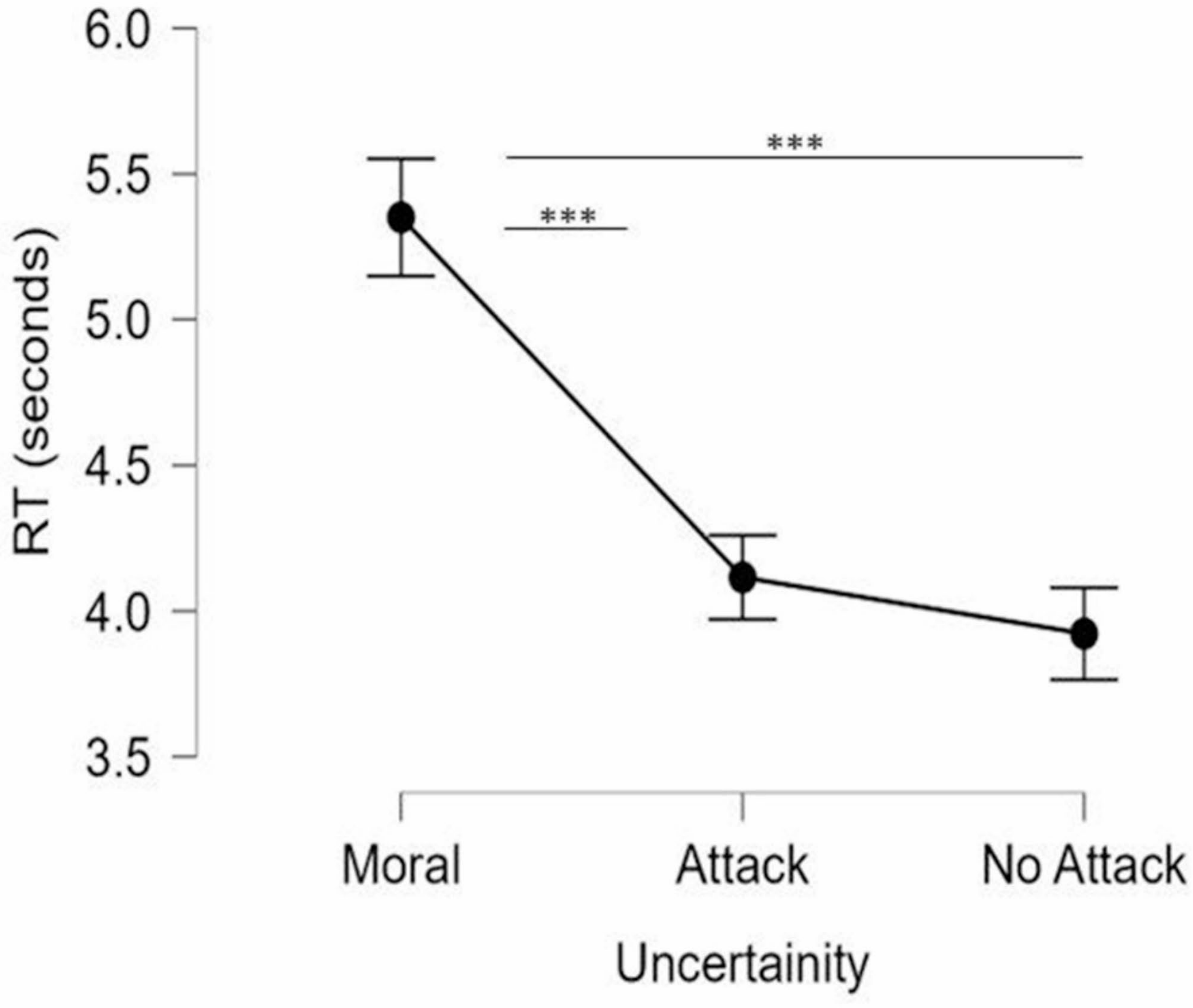

Fig. 3. Response time (in seconds). Participants took significantly more time to make a decision when the situation was morally challenging (Moral trials) than when it was obvious to attack or not to attack. (all $p < .001$). *** = significant.

where external input can provide additional clarity or certainty. In our study, morally challenging trials were intentionally designed to elicit high levels of uncertainty, so it is plausible that participants were more receptive to external information, including AI recommendations. However, a critical factor in Human-AI interaction is the phenomenon of 'anthropomorphism' towards intelligent systems[60–62], which has been shown to be an important determinant of trust in automation[63,64]. Anthropomorphism describes the human tendency to attribute human behaviours and characteristics, such as traits, emotions, motivations and intentions, to non-human entities, including computer applications, robots and machines. Given this tendency, we suggest that AI suggestions are not perceived as ordinary advice. Instead, this anthropomorphism combined with the phenomenon of trust in automation likely increases the weight given to AI recommendations, leading participants to frequently follow its advice, as our results show.

With regard to the Response Time (RT), we expected that participants would need more time to make a decision in trials with a moral conflict and that interaction with the Artificial Intelligence would shorten the participants' response times. Our results partially confirmed our expectation, as RTs were significantly longer in situations that our experts labelled as morally challenging, suggesting that participants need more time to make a decision in these situations and potentially suggesting an internal moral conflict. However, no significant effects were found in relation to the interaction with the AI. This last result is inconsistent with previous evidence from laboratory experiments showing that autonomous systems can help users in recognition tasks [see for example[44]]. Nevertheless, it is important to emphasize that the interaction between trial type and automation was close to significance ($p = .056$), so it is possible that our sample was not large enough to reach significance.

With regard to SoA, in line with previous results[55,65], we expected an increase in SoA during the Moral trials. Consistent with our expectations, the results showed a significant increase in SoA in these trials with

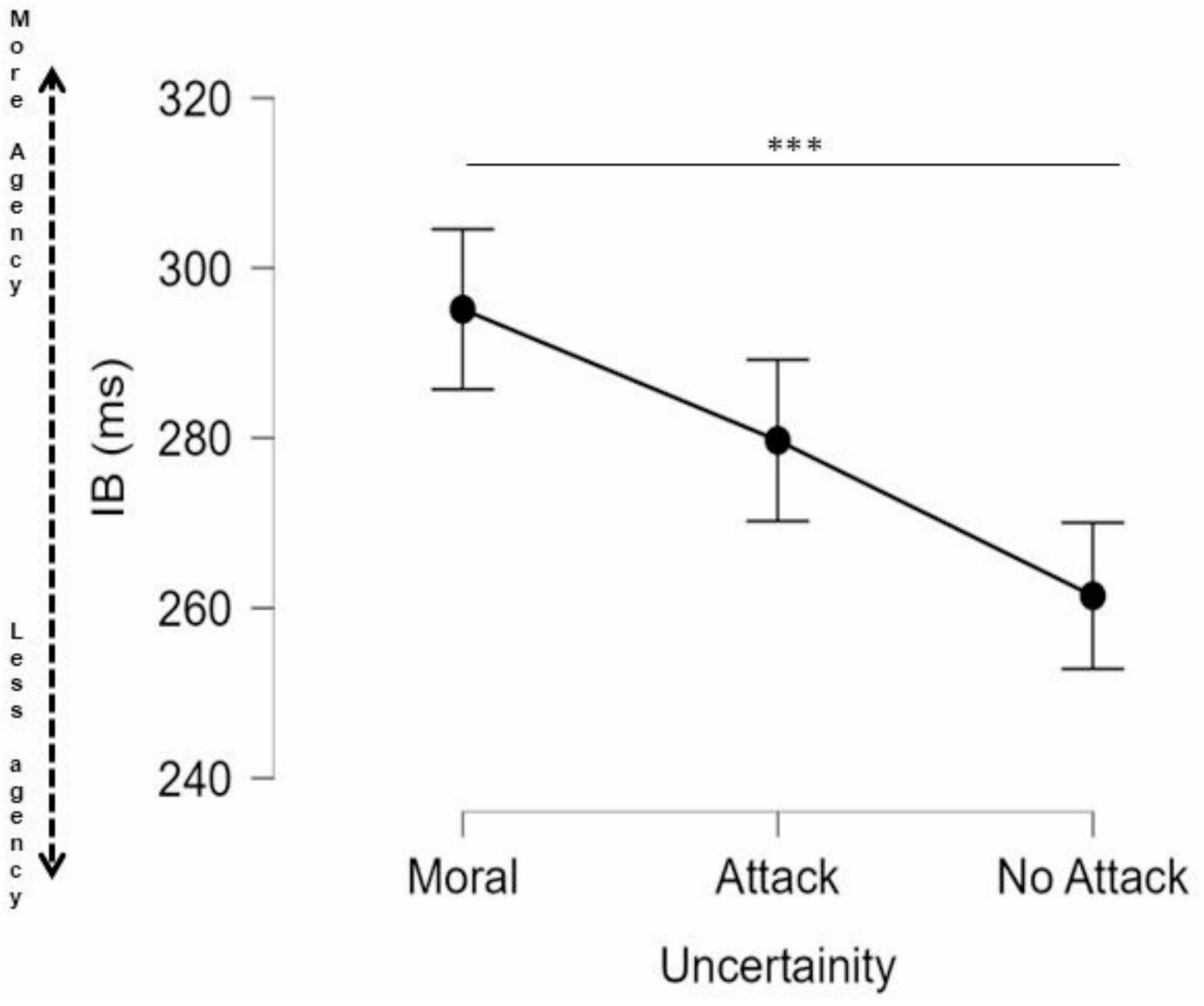

Fig. 4. Temporal binding (TB). A significant difference in the TB was found between Moral trials and NA trials ($p = .002$), but not between Moral and A trials, indicating an increase in the sense of agency during the Moral trials. *** = significant.

respect to the non-moral trials at the implicit level (i.e., in the TB). Thus, the present findings align well with the findings of Moretto and co-workers[55] and confirm the hypothesis that agency is experienced differently in morally significant contexts than in other contexts. Interestingly, we observed a difference between the morally challenging trials and the non-moral trials that did not lead to an attack. However, this difference was not present in the trials that resulted in an attack. It is possible that these latter trials, in which an attack even causes harm or death to enemies, are perceived as more morally challenging than scenarios in which an attack is avoided.

It is important to highlight that our sample also comprised five trained officers, which could in some ways affect this result. Indeed, in a previous study Caspar and colleagues[66] investigated the sense of agency using TB in both cadets and officers and reported that trained officers showed a stronger sense of agency compared to cadets, suggesting that sense of agency may be influenced by rank in the military hierarchy.

Furthermore, as previous studies have reported a decrease in agency during human-machine interactions[14,16], we expected a decrease in agency at both implicit and explicit levels in the AI-assisted blocks[32]. Our results confirmed our expectation and showed a decrease of agency at the explicit level during the interaction with the AI. However, we did not observe a decrease in implicit agency, which contrasts with our expectations and previous studies[14]. However, this discrepancy is not entirely surprising given that, as reported in previous studies[22,26,29], there is a dissociation between the two levels of measures in SoA. Importantly, our results suggest that interacting with an autonomous system in a morally challenging context is likely to lead to different effects on the implicit SoA than interacting with such a system in a morally non-challenging context. The differing outcomes of our study may be explained by the different setting of Berberian's study. In Berberian's study, participants engaged in a flight simulation task supported by different levels of automation to solve a non-morally challenging situation in which they had to select an appropriate command in a flight plan in which a conflict occurred due to the presence of another aircraft [see[14] for more details]. Therefore, it is possible that

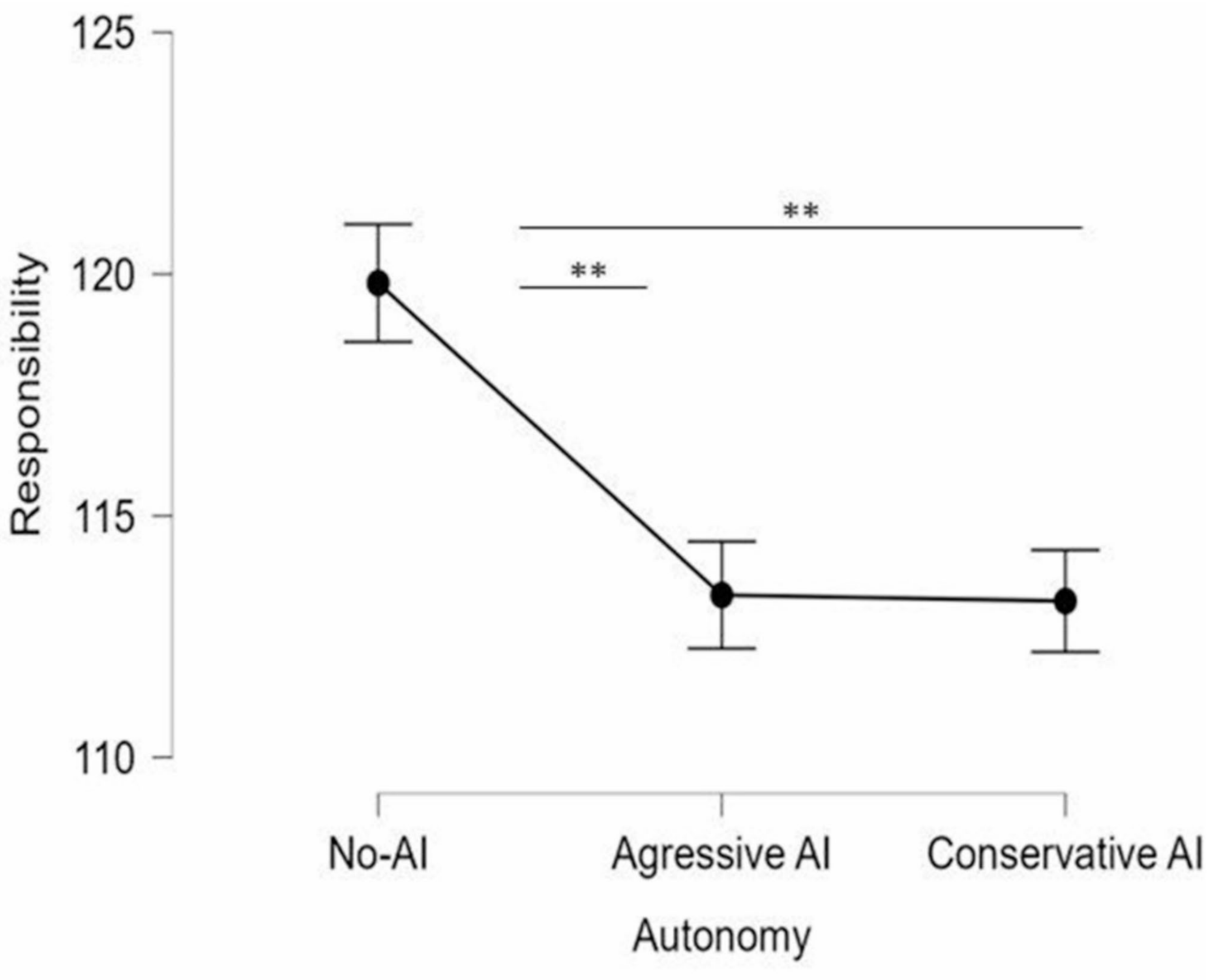

Fig. 5. Assessment of responsibility. A significant difference between Level 0 and Level 1 ($p = .02$) and between Level 0 and Level 2 ($p = .01$), with larger responsibility rating during Level 0. ** = significant.

interaction with autonomous systems influences agency in situations that do not pose a moral challenge, as in Berberian's study, compared to situations that require moral decisions, as in the present study. This finding is particularly consistent with recent findings on the human tendency to attribute moral responsibility to non-human agents, which may lead individuals to hold these agents accountable, assign blame, and clear their own name[67–70]. Interestingly, we also found a significant interaction between the decision and the AI for the subjective sense of responsibility, with the sense of responsibility being higher when participants chose not to follow the "aggressive AI" recommendation. This increased sense of responsibility when not following the AI recommendation opens the way for new, future investigations, as it contrasts with the current data in the literature on Human-AI interaction and responsibility[14,32].

Based on previous findings[43], we hypothesized that higher psychopathic traits would lead to more attacks in morally challenging situations, and to less sense of agency and responsibility. As the locus of control has been shown to be associated with ethical behavior[51,52], we also expected that external locus of control would positively correlate with increased number of attacks in the morally challenging situations, whereas internal locus of control would not lead to changes in moral DM in interactions with intelligent system support. In addition, since Dewey and Knoblich[30], found that locus of control does not correlate with Temporal Binding, we also expected no correlation between locus of control and agency. Furthermore, we investigated possible correlations between the number of attacks and the prioritisation of foundational domains in moral decision-making. Contrary to our expectation, our results showed no correlation between the psychopathic traits, locus of control and our variables of interest, suggesting that moral decision and the sense of agency and responsibility are not related to personality traits, nor to locus of control.

Moreover, we were interested in measuring the confidence in automation, as we hypothesized that higher levels of trust would lead to greater compliance with AI advice. Our results contradict our expectation and show no correlation between trust in automation and our variables of interest, suggesting that trust in automation does not correlate with moral decision-making, sense of agency and sense of responsibility.

An important limitation of this study relates to the statistical power to analyze the role of individual differences in moral decision-making. While our sample size was sufficient to detect within-subject effects, analyses exploring between-subjects individual difference measures, such as the personality tests used here, were underpowered to detect moderate or small effects. Therefore, as the lack of statistically significant results may be due to the limited sample size rather than the actual absence of such effects, these results should be interpreted with caution. Future research with larger samples is needed to draw more reliable conclusions about the role of individual differences in moral decision-making, especially in contexts where humans interact with AI in morally charged situations.

Furthermore, as previous findings[66] have shown the influence of hierarchy on the sense of agency, future studies should aim to investigate in depth the differences between ranks within the military ranks to better understand how hierarchical status modulates cognitive and ethical decision-making mechanisms. Such research could provide crucial insights into the specific processes underlying these effects and their implications for military operations. We acknowledge that our sample consists primarily of cadets, who cannot be equated with fully trained, operational military personnel. However, their inclusion provides a unique perspective as they are at an advanced stage of training and are familiar with military decision-making principles. Specifically, all participants had received formal training in International Humanitarian Law, a framework directly relevant to the moral dilemmas examined in this study.

By focusing on this population, we aimed to explore how individuals with foundational military training — who are in the process of transitioning to decision-making roles in real-world military contexts — approach ethically and legally complex scenarios. While this sample is not fully representative of experienced military decision-makers, it provides important data on the cognitive and ethical processes of individuals being prepared to manage such dilemmas in operational settings. These findings lay the groundwork for future research to clarify how hierarchical and experiential factors shape decision-making and moral reasoning within military populations.

Lastly, in the present study, we aimed to investigate the effects of human-AI interaction on moral decision-making, sense of agency, and responsibility as separate outcomes. However, this study does not provide a fully integrated model of the relationships between these variables. Future research should explore how these variables interact within a more integrated framework, as such an approach could provide additional insights into the complex dynamics of human-AI interactions in morally challenging contexts. For example, in future experiments with a larger sample, it might be important to conduct a mediation analysis to investigate whether the sense of agency or sense of responsibility is a causal factor for the number of attacks or whether the sense of agency and the proportion of attacks are only influenced by the autonomous system's recommendations, with no direct influence between the two.

Conclusion

Our results suggest that interacting with AI may affect moral decision-making. In moral situations, decisions may be guided by the autonomous system's behavior. This study also suggests that being in a situation of moral decision-making increases the sense of agency, which is closely related to moral responsibility, but that interacting with an autonomous system does not affect it, at least at implicit level. However, when people are explicitly asked about their sense of responsibility, they indicate that they feel less responsible when an intelligent autonomous system has helped them make decisions, indicating that what they explicitly report does not correspond to how they subjectively experience the situation, which could indicate a self-serving bias[50].

The evidence of the negative impact on human decision-making and the evidence of a decrease in the sense of responsibility that could result from interaction with autonomous systems could have serious implications for the safety of their use in contexts implying moral decision-making. Thus, the behavior of AI systems needs to be carefully programmed and people should remain critical with regards to it, especially considering their significant impact on moral decision-making. As our study suggests, individuals tend to rely heavily on AI recommendations in morally challenging situations. This trust can amplify the impact of AI's programmed behavior on critical decisions. It is therefore important to take ethical considerations into account when developing AI systems and to ensure that their behavior promotes responsible and morally sound decision-making. With the increasing presence of intelligent autonomous systems in our daily lives it is crucial to increase research on the influence of autonomous systems on these aspects, as there is currently a lack of research on this topic and especially because the decision-making process is well documented in civilian samples, while studies on moral decision-making in the military or other life-and-death situations have only recently appeared in the literature[70] and show a mixed picture of the effects[71,72].

Data availability

The datasets generated and analysed during the current study are available in the Open Science Framework (Salatino, A., Prével, A., Caspar, E. A., & Lo Bue, S. (2025, February 3). Influence of AI Behavior on Human Moral decisions, Agency, and Responsibility. https://doi.org/10.17605/OSF.IO/ZYFQD.

Received: 18 September 2024; Accepted: 21 March 2025
Published online: 10 April 2025

References

1. Kawamoto, K., Houlihan, C. A., Balas, E. A. & Lobach, D. F. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj* **330** (7494), 765 (2005).
2. Sutton, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit. Med.* **3** (1), 17 (2020).

3. Černevičienė, J. & Kabašinskas, A. Review of multi-criteria decision-making methods in finance using explainable artificial intelligence. *Front. Artif. Intell.* **5**, 35 (2022).
4. Wasilow, S. & Thorpe, J. B. Artificial intelligence, robotics, ethics, and the military: A Canadian perspective. *AI Mag.* **40** (1), 37–48 (2019).
5. Rashid, A. B., Kausik, A. K., Hassan Sunny, A. & Bappy, M. H. A., Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *Int. J. Intell.Syst.*, 8676366 (2023).
6. Parasuraman, R. & Manzey, D. H. Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* **52** (3), 381–410 (2010).
7. Endsley, M. R. From here to autonomy: lessons learned from human–automation research. *Hum. Factors* **59** (1), 5–27 (2017).
8. Haslbeck, A. & Hoermann, H. J. Flying the needles: flight deck automation erodes fine-motor flying skills among airline pilots. *Hum. Factors* **58** (4), 533–545 (2016).
9. Volz, K. M. & Dorneich, M. C. Evaluation of cognitive skill degradation in flight planning. *J. Cogn. Eng. Decis. Mak.* **14** (4), 263–287 (2020).
10. Ahn, J., Kim, J. & Sung, Y. The effect of gender stereotypes on artificial intelligence recommendations. *J. Bus. Res.* **141**, 50–59 (2022).
11. Pelau, C., Dabija, D. C. & Ene, I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Comput. Hum. Behav.* **122**, 106855 (2021).
12. Shank, D. B., Graves, C., Gott, A., Gamez, P. & Rodriguez, S. Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Comput. Hum. Behav.* **98**, 256–266 (2019).
13. Rovira, E., McGarry, K. & Parasuraman, R. Effects of imperfect automation on decision making in a simulated command and control task. *Hum. Factors* **49** (1), 76–87 (2007).
14. Berberian, B., Sarrazin, J. C., Le Blaye, P. & Haggard, P. Automation technology and sense of control: a window on human agency. *PloS ONE* **7** (3), e34075 (2012).
15. Berberian, B. Man-machine teaming: a problem of agency. *IFAC PapersOnLine* **51** (34), 118–123 (2019).
16. Zanatto, D., Chattington, M. & Noyes, J. Human-machine sense of agency. *Int. J. Hum. Comput. Stud.* **156**, 102716 (2021).
17. Haggard, P. & Tsakiris, M. The experience of agency: feelings, judgments, and responsibility. *Curr. Dir. Psychol. Sci.* **18** (4), 242–246. https://doi.org/10.1111/j.1467-8721.2009.01644.x (2009).
18. Pyasik, M. et al. Shared neurocognitive mechanisms of attenuating self-touch and illusory self-touch. *Soc. Cognit. Affect. Neurosci.* **14** (2), 119–127 (2019).
19. Caspar, E. A. et al. Commanding or being a simple intermediary: how does it affect moral behavior and related brain mechanisms? *ENeuro* **9** (5) (2022).
20. Imaizumi, S. & Tanno, Y. Intentional binding coincides with explicit sense of agency. *Conscious. Cogn.* **67**, 1–15 (2019).
21. Malik, R. A. & Obhi, S. S. Social exclusion reduces the sense of agency: evidence from intentional binding. *Conscious. Cogn.* **71**, 30–38 (2019).
22. Moore, J. W. & Obhi, S. S. Intentional binding and the sense of agency: a review. *Conscious. Cogn.* **21** (1), 546–561 (2012).
23. Caspar, E. A., Cleeremans, A. & Haggard, P. The relationship between human agency and embodiment. *Conscious. Cogn.* **33**, 226–236 (2015).
24. Haggard, P., Clark, S. & Kalogeras, J. Voluntary action and conscious awareness. *Nat. Neurosci.* **5** (4), 382–385 (2002).
25. Christensen, J. F., Di Costa, S., Beck, B. & Haggard, P. I just lost it! Fear and anger reduce the sense of agency: a study using intentional binding. *Exp. Brain Res.* **237**, 1205–1212 (2019).
26. Saito, N., Takahata, K., Murai, T. & Takahashi, H. Discrepancy between explicit judgement of agency and implicit feeling of agency: implications for sense of agency and its disorders. *Conscious. Cogn.* **37**, 1–7 (2015).
27. Blackwood, N. J., Bentall, R. P., Simmons, A., Murray, R. M. & Howard, R. J. Self-responsibility and the self-serving bias: an fMRI investigation of causal attributions. *NeuroImage* **20** (2), 1076–1085 (2003).
28. Wegner, D. M. & Wheatley, T. Apparent mental causation: sources of the experience of will. *Am. Psychol.* **54** (7), 480 (1999).
29. Synofzik, M., Vosgerau, G. & Newen, A. I move, therefore I am: A new theoretical framework to investigate agency and ownership. *Conscious. Cogn.* **17** (2), 411–424 (2008).
30. Dewey, J. A. & Knoblich, G. Do implicit and explicit measures of the sense of agency measure the same thing? *PloS ONE* **9** (10), e110118 (2014).
31. Coyle, D., Moore, J., Kristensson, P. O., Fletcher, P. & Blackwell, A. I did that! Measuring users' experience of agency in their own actions. In *Proceedings of the SIGCHI conference on human factors in computing systems* 2025–2034. (2012).
32. Vantrepotte, Q., Berberian, B., Pagliari, M. & Chambon, V. Leveraging human agency to improve confidence and acceptability in human-machine interactions. *Cognition* **222**, 105020 (2022).
33. Arkin, R. C., Ulam, P. & Wagner, A. R. Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. *Proc. IEEE* **100** (3), 571–589 (2011).
34. Jiang, L. et al. Can machines learn morality? the delphi experiment. *arXiv* (2021).
35. Köbis, N., Bonnefon, J. F. & Rahwan, I. Bad machines corrupt good morals. *Nat. Hum. Behav.* **5** (6), 679–685 (2021).
36. de Melo, C. M., Marsella, S. & Gratch, J. Human cooperation when acting through autonomous machines. *Proc. Natl. Acad. Sci.* **116** (9), 3482–3487 (2019).
37. Cohn, A., Gesche, T. & Maréchal, M. A. Honesty in the digital age. *Manag. Sci.* **68** (2), 827–845 (2022).
38. Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M. & Irlenbusch, B. The corruptive force of AI-generated advice. *arXiv* (2021).
39. Xu, Y. et al. Artificial intelligence: A powerful paradigm for scientific research. *Innovation* **2** (4) (2021).
40. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25** (1), 44–56 (2019).
41. Parasuraman, R. & Riley, V. Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* **39** (2), 230–253 (1997).
42. Li, P., Han, C., Lei, Y., Holroyd, C. B. & Li, H. Responsibility modulates neural mechanisms of outcome processing: an ERP study. *Psychophysiology* **48** (8), 1129–1133 (2011).
43. Caspar, E. A., Cleeremans, A. & Haggard, P. Only giving orders? An experimental study of the sense of agency when giving or receiving commands. *PloS ONE* **13** (9), e0204027 (2018).
44. Chavaillaz, A., Schwaninger, A., Michel, S. & Sauer, J. Automation in visual inspection tasks: X-ray luggage screening supported by a system of direct, indirect or adaptable cueing with low and high system reliability. *Ergonomics* **61** (10), 1395–1408 (2018).
45. Faul, F., Erdfelder, E., Lang, A. G. & Buchner, A. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39** (2), 175–191 (2007).
46. Salatino, A., Prével, A., Caspar, E. A. & Lo Bue, S. Influence of AI behavior on human moral decisions, agency, and responsibility. https://doi.org/10.17605/OSF.IO/ZYFQD (2025).
47. Levenson, M. R., Kiehl, K. A. & Fitzpatrick, C. M. Assessing psychopathic attributes in a noninstitutionalized population. *J. Personal. Soc. Psychol.* **68** (1), 151 (1995).
48. Graham, J. et al. Mapping the moral domain. *J. Personal. Soc. Psychol.* **101** (2), 366 (2011).
49. Jian, J. Y., Bisantz, A. M. & Drury, C. G. Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* **4**, 53–71 (2000).
50. Bandura, A. Toward a psychology of human agency. *Perspect. Psychol. Sci.* **1**, 164–180 (2006).

51. Klebe- Treviño, and Youngblood. 'Bad apples in bad barrels: A causal analysis of ethical decision making behavior'. *J. Appl. Psychol.* **75** (4), 378–385 (1990).
52. Forte, A. Locus of control and the moral reasoning of managers. *J. Bus. Ethics* **58**, 65–77 (2005).
53. Treviño, W. Reynolds, behavioral ethics in organizations: A review. *J. Manag.* **32**, (6), 951–990 (2006).
54. Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K. & Cagan, J. Human confidence in artificial intelligence and in themselves: the evolution and impact of confidence on adoption of AI advice. *Comput. Hum. Behav.* **127**, 107018 (2022).
55. Moretto, G., Walsh, E. & Haggard, P. Experience of agency and sense of responsibility. *Conscious. Cogn.* **20** (4), 1847–1854 (2011).
56. Fabre, E. F., Mouratille, D., Bonnemain, V., Palmiotti, G. P. & Causse, M. Making moral decisions with artificial agents as advisors. An fNIRS study. *BioRxiv*, 2024–2003 (2024).
57. Chen, J. Y. & Barnes, M. J. Supervisory control of multiple robots: effects of imperfect automation and individual differences. *Hum. Factors* **54** (2), 157–174 (2012).
58. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Ann. Rev. Psychol.* **62** (1), 451–482 (2011).
59. Krueger, J. I. (ed) *Social Judgment and Decision Making* (Psychology, 2012).
60. Bartneck, C., Kulić, D., Croft, E. & Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **1**, 71–81 (2009).
61. Waytz, A., Cacioppo, J. & Epley, N. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspect. Psychol. Sci.* **5** (3), 219–232 (2010).
62. Troshani, I., Rao Hill, S., Sherman, C. & Arthur, D. Do we trust in AI? Role of anthropomorphism and intelligence. *J. Comput. Inform. Syst.* **61** (5), 481–491 (2021).
63. Hoff, K. A. & Bashir, M. Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* **57** (3), 407–434 (2015).
64. Chiou, E. K. & Lee, J. D. Trusting automation: designing for responsivity and resilience. *Hum. Factors* **65** (1), 137–165 (2023).
65. Spaccasassi, C., Cenka, K., Petkovic, S. & Avenanti, A. Sense of agency predicts severity of moral judgments. *Front. Psychol.* **13**, 1070742 (2023).
66. Caspar, E. A., Lo Bue, S., Magalhães De Saldanha da Gama, P. A., Haggard, P. & Cleeremans, A. The effect of military training on the sense of agency and outcome processing. *Nat. Commun.* **11** (1), 4366 (2020).
67. Furlough, C., Stokes, T. & Gillan, D. J. Attributing blame to robots: I. The influence of robot autonomy. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **63** (4), 592–602 (2021).
68. Liu, P. & Du, Y. Blame attribution asymmetry in human–automation cooperation. *Risk Anal.* **42** (8), 1769–1783 (2022).
69. Kneer, M. & Stuart, M. T. Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* 407–411 (2021).
70. Oimann, A. K. & Salatino, A. Command responsibility in military AI contexts: balancing theory and practicality. *AI Ethics*, 1–11. (2024).
71. Verweij, D., Hofhuis, K. & Soeters, J. Moral judgement within the armed forces. *J. Mil. Ethics* **6** (1), 19–40 (2007).
72. Kimhi, S. & Kasher, A. Moral dilemmas in military situations: proportionality principle, religiosity, political attitudes, and authoritarian personality. *Mil. Psychol.* **27** (3), 169–184 (2015).

Author contributions

A.P., S.L.B., E.C.: Conceptualization, Methodology. A.P.: Software. A.S.: Investigation. A.S., A.P.: Data curation, Formal Analysis, Writing- Original draft preparation. E.C., S.L.B.: Supervision. S.L.B., A.P: Funding acquisition. A.S., A.P., E.C., S.L.B.: Writing- Reviewing and Editing.

Funding

This research was funded by Belgian Defense – Royal Higher Institute of Defense, grant number HFM23-03.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.