



HAL
open science

“Fire! Do Not Fire!”: A new paradigm testing how autonomous systems affect agency and moral decision-making

Adriana Salatino, Arthur Prevel, Emilie Caspar, Salvatore Lo Bue

► To cite this version:

Adriana Salatino, Arthur Prevel, Emilie Caspar, Salvatore Lo Bue. “Fire! Do Not Fire!”: A new paradigm testing how autonomous systems affect agency and moral decision-making. 2023. hal-04454490

HAL Id: hal-04454490

<https://hal.univ-lille.fr/hal-04454490>

Preprint submitted on 13 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **“Fire! Do Not Fire!”: A new paradigm testing how autonomous systems affect agency and**
2 **moral decision-making**

3

4 Adriana Salatino^{a*}, Arthur Prével^{b*}, Émilie Caspar^c, & Salvatore Lo Bue^a

5

6 *^aDepartment of Life Sciences, Royal Military Academy, Hobbema 8, 1000, Brussels, Belgium*

7 *^bUniv. Lille, CNRS, UMR 9193 – SCALab – Sciences Cognitives et Sciences Affectives, F-59000*

8 *Lille, France*

9 *^cMoral & Social Brain Lab, Department of Experimental Psychology, Ghent University*

10

11 *Corresponding authors: adriana.salatino@mil.be and arthur.prevel@univ-lille.fr

12

13

14

15 **Abstract**

16 Autonomous systems have pervaded many aspects of human activities. However, research suggest
17 that the interaction with those machines may influence human decision-making processes. These
18 effects raise ethical concerns in moral situations. We created an ad hoc setup to investigate the
19 effects of system autonomy on moral decision-making and human agency in a trolley-like
20 dilemma. In a battlefield simulation, participants had to decide whether to initiate an attack
21 depending on conflicting moral values. Our results suggest that our paradigm is suitable for future
22 research aimed at understanding the effects of system autonomy on moral decision-making and
23 human agency.

24

25 **Keywords:** Human-autonomous systems interaction; Human performance; Level of system
26 autonomy; Moral decision-making; Sense of agency

27

28 **1. Introduction**

29 In the last decades, the use of autonomous systems has become increasingly widespread in
30 various fields of human activity. From driving (Ayoub et al., 2019; Chan, 2017) to aviation
31 (Anderson et al., 2018; Chialastri, 2012; Valdés et al., 2018), medicine (Kawamoto et al., 2005;
32 Sutton et al., 2020), and military defense and security (Mayer, 2015), there is almost no area where
33 these technologies are not massively deployed. As these technologies become more widespread,
34 the inevitable trend toward even more machine autonomy will lead to profound changes in the role
35 played by humans during the execution of tasks. From workers in industry, agriculture, and
36 transportation, to consumers in their daily lives, advances observed in technologies show that
37 humans are rapidly moving from the position of direct tasks executors (with the help of
38 mechanistic machines) to supervising tasks performed directly by intelligent machines with high
39 level of autonomy.

40 The main reason for the massive deployment of autonomous systems resides in the many
41 benefits these systems offer to users. Several laboratory experiments have shown that the
42 introduction of some level of autonomy in tasks can have substantial effects in terms of users'
43 decision-making and performance. For example, studies have shown that autonomous systems can
44 help people in detecting task-relevant cues and ignoring irrelevant cues in the environment
45 (Chavaillaz et al., 2018; Goh et al., 2005), improve the decisions made by human subjects in
46 complex situations and reduce the number of errors they make (MacMillan et al., 1997; Rovira et
47 al., 2007; Sarter & Schroeder, 2001), or even reduce the time to make correct decisions (Chavaillaz
48 et al., 2018). In addition, the introduction of autonomy has been shown to reduce users' mental
49 workload and thus increase their ability to monitor multiple tasks simultaneously (e.g., Chen &
50 Barnes, 2012; Wright et al., 2018). Laboratory experiments have also shown that the effects

51 produced by the interaction with autonomous systems on human decision -making and performance
52 depend on several factors (Parasuraman & Riley, 1997; Mosier & Manzey, 2019). In particular,
53 autonomous systems operating at the decision stage of a task or with a high level of autonomy are
54 generally associated with the largest benefits in terms of human performance (e.g., Endsley &
55 Kaber, 1999; Manzey et al., 2012; Rovira et al., 2007). Based on these results, one could easily
56 conclude that more autonomy is always better for the users, whether they are pilots flying on a
57 plane, physicians analyzing the test results from a patient, or military drone operators deployed in
58 a war zone.

59 However, the involvement of autonomous systems technologies into human activities has
60 not been systematically associated with positive effects. Several studies have shown that their use
61 can also have significant negative effects. Parasuraman et al. (1993) provided a classic
62 demonstration thereof. They reported low level of detection of automation failures with highly
63 reliable systems, an effect they called automation complacency (also named automation
64 overreliance). Other examples of negative effects are loss of situational awareness (Endsley, 2017),
65 skill decay (Haslbeck & Hoemann, 2016; Volz & Dorneich, 2020), performance decrement in
66 return-to-manual control (Endsley & Kiris, 1995), or increased workload with too many
67 autonomous systems to monitor (Wang et al., 2009). Interestingly, the detrimental effects caused
68 by the cooperation with autonomous systems seem to be directly related with the stage and level
69 of autonomy of those systems. For example, Rovira et al. (2007) reported lower rates of correct
70 decisions with highly reliable systems with high levels of autonomy (i.e., decision systems)
71 compared to systems with lower levels of autonomy (i.e., information systems). Thus, while more
72 autonomy seems to be clearly beneficial when the system's recommendations are correct, the

73 negative effects seem also to be more pronounced for higher levels of autonomy in machines that
74 are, most of the time, imperfect.

75 In this context, another important aspect of human-autonomous systems interaction that
76 has received increasing attention in recent years is the impact of autonomy on human agency
77 (Berberian, 2012, 2019, Coyle et al., 2012; Zanatto et al., 2021). The sense of agency (SoA),
78 defined as the *feeling of causing changes in the external world by controlling one's own voluntary*
79 *actions* (Jeannerod, 2003; Haggard, 2017; Burin et al., 2017; Pyasik, Salatino et al., 2019), is
80 recognized as an important aspect of human consciousness. Because SoA enables us to perceive
81 ourselves as causal agents, it is the basis for intentional behavior (Haggard & Tsakiris, 2009), and
82 is closely related to moral responsibility (Moretto et al., 2011; Caspar, Christensen, Cleeremans,
83 & Haggard, 2016).

84 The recent interest in this topic has been triggered by the possibility offered by the
85 “Intentional Binding” effect to implicitly measure the SoA. The Intentional Binding is a
86 phenomenon by which the perceived time between an action and its outcomes is modulated by the
87 intentionality of that action. Time appears compressed in situations where the person is active,
88 while time appears stretched in situations where the person is passive (Haggard et al., 2002).
89 Measuring the SoA by using this effect usually consists of asking subjects to estimate the time
90 interval between an action they perform and the consequences of that action. Numerous studies
91 have now shown that the time estimation between action and outcome is a valid implicit measure
92 of SoA (e.g., Christensen et al., 2019; Imaizumi & Tanno, 2019; Malik & Obhi, 2019; Haggard,
93 Clark, and Kalogeras, 2002; Moore and Obhi, 2012) and is preferable to a subjective measurement
94 of responsibility, which is usually obtained by a direct report of how people attribute the effects of

95 their own actions (Saito et al., 2015), which is subject to social desirability and other biases (e.g.,
96 Blackwood et al., 2003; Wegner & Withley, 1999).

97 In one of the first studies investigating the effects of autonomy on SoA by using the
98 Intentional Binding paradigm (Berberian et al., 2012), participants took part in a flight simulation
99 and were assisted in their task by different levels of automation. Berberian and colleagues' results
100 showed a decrease in SoA with increasing levels of automation, suggesting that agency decreases
101 with higher levels of automation. Further evidence using the same paradigm can be found in the
102 study by Coyle and coworkers (2012), who investigated how assistance, in a machine-assisted
103 point-and-click task, affects the user's SoA. Their results suggest that, up to a certain point, the
104 computer could assist users while also allowing them to maintain a sense of control over their
105 actions and outcomes, hence of their agency. More recently, Zanatto et al. (2021) showed a similar
106 negative impact of automation on SoA, and that the mental workload may also play a role in
107 reducing agency. Taken together, these studies suggest that automation technology may affect the
108 mechanism underlying human agency.

109 Hence, the evidence of negative effects on human decisions and performance, and the
110 evidence of a decrease in the implicit and explicit SoA (Berberian et al., 2012; Coyle et al., 2012;
111 Vantrepotte et al., 2022), that might result from the interaction with the autonomous systems, have
112 serious performance and safety implications. Engineers working on the development of new forms
113 of autonomous technologies should be aware of these effects and take them into account.
114 Furthermore, these results have important implications when those systems are used in sensitive
115 or moral domains such as in medicine or in the military, in which decisions of life and death have
116 to be made.

117 To date, however, very little is known about how the interaction with an autonomous
118 machine affects SoA and the decisions made by someone engaged in a moral scenario, and how
119 this is influenced by the level of autonomy of the system. Indeed, it is possible that interacting with
120 autonomous systems to make moral decisions negatively affects the moral and ethical decision-
121 making process and the resulting actions, particularly in tasks and domains of moral value such as
122 in the military context (Christensen et al., 2012; Cushman et al., 2013, 2017).

123 In recent years, research in the field of moral decision making and autonomy has focused
124 mainly on the rules and/or algorithms that can be assigned to an autonomous system to perform
125 ethical responses in moral situations (Arkin et al., 2011; Jiang et al., 2021). Surprisingly, until
126 recently, little attention has been paid to understanding how a human agent's ethical behavior in
127 moral decision-making situations can be influenced by its interaction with an autonomous system
128 (Köbis et al., 2021). The available data suggest a mixed picture of the effects of autonomous
129 systems in social and moral decision-making situations. Indeed, while some recent evidence
130 suggests that interaction with automation could lead to the promotion of prosocial behaviors (such
131 as fairness and cooperation, see, e.g., de Melo et al., 2018, 2019), other studies have shown that it
132 could also lead to unethical behaviors (Cohn et al., 2022; Leib et al., 2021). Concerning SoA, while
133 some studies report that a moral context increases it (e.g., Moretto et al., 2011) and that a higher
134 SoA is associated with higher prosocial decision-making (Caspar et al., 2022), it is not clear
135 whether this is still true when decisions are made in collaboration with an autonomous intelligent
136 machine.

137 Considering the lack of research on how autonomous systems impact the SoA and decision-
138 making in a moral context, and how this can be modulated by the level of autonomy of the system,
139 in the present study, we aimed to build an ad hoc setup to investigate how the level of system

140 autonomy affects SoA and the moral decision-making. To this end, we developed a task in which
141 participants (military cadets) played the role of drone operators on a simulated battlefield and had
142 to decide whether or not to initiate an attack, based on the presence of enemies and the risk that
143 allies might also be harmed. Participants were exposed to three types of trials representing three
144 types of uncertainty (*Moral Decision-Making* Trials, *No Risk* Trials, and *No Enemy* Trials) with
145 three different levels of system autonomy, including no system assistance, information assistance
146 (i.e., the system gives processed-information on the presence of enemies and the risk for allies),
147 and decision assistance (i.e., the system provides a recommendation on the best decision to make).
148 In our study, SoA is measured both at the implicit level, using the Intentional Binding paradigm,
149 and at the explicit level through a subjective assessment of responsibility (using an ad-hoc scale).
150 We also measured performance by using reaction time, the proportion of trials in which
151 participants chose to attack, and the proportion of choices leading to the fewest ally losses (called
152 *utilitarian choices* in our task).

153 The primary purpose of this research was to develop and test a new paradigm to investigate
154 how the interaction with autonomous systems can affect the SoA and the decisions made by people
155 when facing moral choices, and how the level of autonomy of the system influences this effect.
156 Based on previous findings (Berberian et al., 2012, Coyle et al., 2012; Zanatto et al., 2021), we
157 hypothesized that agency decreases with increasing levels of system autonomy, as indicated by a
158 longer time estimation between action and outcome, and lower subjective judgments of
159 responsibility at higher levels of system autonomy. We also hypothesized that if the SoA would
160 be affected by the autonomous system, as well as the sense of responsibility associated with SoA,
161 the moral decision-making would also be affected, with the number of attacks increasing as system
162 autonomy increases (Caspar et al., 2018; Goh et al., 2005, Chavallaz et al., 2018). In addition, we

163 expected shorter reaction times and more utilitarian choices with higher levels of system
164 autonomy. Crucially, by validating the new paradigm we propose with the present study, we hope
165 to pave the way for new quantitative studies to understand how the interaction with autonomous
166 systems affects agency and decision- making in a moral context. In turn, understanding these
167 effects better could help in the development of safer and more efficient autonomous systems in the
168 future.

169 **2. Method**

170 *2.1. Participants*

171 A total of 31 participants took part in the study ($M_{\text{age}} = 22$, $SD = 2.28$, $\text{Range} = 19 - 36$, 7
172 women, 24 men). Participants were cadets at the Royal Military Academy of Belgium and were
173 recruited with the help of a Master student officer in the course of his thesis. Participants were in
174 their third and fourth year of study, meaning they had notions of International Humanitarian Law,
175 and thus about what is legally allowed and forbidden in the conduct of armed conflicts. The sample
176 size was estimated using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), with a small-to-
177 medium size effect f of 0.2, a threshold for significance α set at 0.05, and a power $1 - \beta$ at .80. Based
178 on these values, the estimated sample size was 28 participants. To compensate for potential data
179 losses and exclusions, a total of 32 participants was targeted. However, considering the limited
180 pool and recruitment period, 31 participants were finally included in this study. The study was
181 performed in accordance with the principles expressed in the declaration of Helsinki and with the
182 protocol of the local ethics review board at the Faculty of Psychology and Educational Sciences at
183 Ghent University. Participants were informed of the general purpose and the duration of the
184 experiment, and about their rights as participants in psychological research before giving their
185 consent. Participation was voluntary and participants could withdraw their participation at any
186 time without justification and without consequences. Written informed consent was obtained from
187 each participant before the experiment.

188

189 *2.2. Stimuli and procedure*

190 All the material can be found on the Open Science Framework (Salatino et al., 2023). The
191 experiment was programmed and presented using MATLAB 2020b and the Psychophysics

192 Toolbox extension. The experiment was run on a laptop computer (display resolution: 2560*1600
193 pixels) and responses were collected via an AZERTY keyboard with left and right arrow keys
194 serving as the response keys used in the experiment. The experiment was conducted in an
195 experimental room in the department of Life Sciences at the Royal Military Academy. Participants
196 were seated on a chair, in front of a table with the screen of the laptop being at approximately 40
197 cm from the participants. Each trial of the task consisted of a 50 by 50 grid with dark grey cells
198 (RGB = [064, 064, 064]) presented on a black screen (RGB = [000, 000, 000]; see Fig. 1, below).
199 Participants were instructed that the grid represented a radar display used to inform on the position
200 of allies and enemies. Out of the 2500 cells, 100 were colored in light grey (RGB = [192, 192,
201 192]) and each light grey cell represented the position of a group of 10 allies. The number of allied
202 forces (i.e., light grey cells) was kept constant across trials, but their position varied randomly from
203 trials to trials. In addition, in 66% of trials one of the 2500 cells was colored in white (RGB = [255,
204 255, 255]) to represent the presence and the position of an enemy. Maximum one cell was colored
205 in white during trials (i.e., maximum of one enemy). When an enemy (i.e., a white cell) was
206 presented on a trial, participants were asked to choose between attacking the enemy or not by
207 pushing the left (attack) and right (no attack) arrows of the keyboard (i.e., Action 1- A1, if they
208 decide to attack, and Action 2- A2, if they decide to not attack). Participants were instructed that
209 not pushing the “attack” button whenever an enemy was present would result in the death of five
210 allies because of the continued hostile activities of that enemy. Instead, pushing the “attack” button
211 whenever an enemy was present would result in the death of the enemy, but with sometimes a risk
212 of collateral damages considering the position of the allies on the 50 by 50 grid radar display (see
213 details in the next paragraph below). Collateral damage was likely, but not sure, when grey cells

214 were separated from the enemy by less than five empty cells. When no enemy was shown on the
215 grid, participants were asked to not push on the “attack” button.

216

217 *Fig. 1, here*

218

219 During the experiment, participants were tested on three different types of trials,
220 representing three levels of Uncertainty: *No Risk* trials, *No Enemy* trials, and *Moral Decision-*
221 *Making (Moral DM)* trials. In the *No Risk trials*, an enemy was shown on the radar display and
222 allies were separated from five cells of more from the enemy, meaning there were no risk of
223 collateral damage. In the *No Enemy trials*, no enemy was shown on the radar display. Finally, in
224 *Moral DM trials* an enemy was shown on the radar display and up to three groups of allies were
225 on the first to the fifth range of cells next to the enemy. Here, participants received the instruction
226 that during those trials there was a risk of collateral damages if they decided to attack. More
227 exactly, participants were instructed that allies found on the 1st range of cells next to the position
228 of the enemy had a .50 probability of being killed if they decided to attack, while the probability
229 of collateral damages decreased with the distance according to the following *range – probability*
230 combinations: 2nd range – .40, 3rd range – .30, 4th range – .20, and 5th range – .10. The number of
231 allies that could be killed during an attack varied from 10, 20, or 30 allies. Groups of allies were
232 systematically on the same range of cells and *Moral DM* trials consisted in 15 different *number of*
233 *allies by probabilities of collateral damages* combinations. Thus, we expected those trials to be
234 morally challenging because participants were asked to choose between (1) pushing the attack
235 button to neutralize the enemy, but with the risk of killing allies during the attack, or (2) not
236 pushing the attack button and let the enemy kill 5 allies anyway.

237 The experiment consisted of three blocks, with each block including a specific level of
238 Autonomy from intelligent system (*Level 0, Level 1, Level 2*). In *Level 0*, the block with the lowest
239 level of assistance from intelligent system, participants received basic visual assistance to
240 determine the position of allies relatively to the position of an enemy shown on the grid. During
241 *No Risk* and *Moral DM* trials, by clicking with the help of the computer mouse on a white cell
242 while the 50 by 50 grid was shown on the screen, the cells on the 6th range next to the position of
243 the enemy turned blue (RGB = [000, 000, 255]) until the participants made their decision. This
244 visual information was designed to help the participants to detect how far allied forces were from
245 the enemy's position and to compute the risk of collateral damages if they decided to press the
246 attack button. However, allies found within that area were at risk for collateral damages, with the
247 risk depending on the range.

248 Then, in *Level 1* of Autonomy, in addition to the basic visual assistance found in *Level 0*,
249 a 12 grade-scale was shown on the right part of the screen (see Fig. 1, right up) participants were
250 instructed that this scale indicated the risk in terms of losses of allied forces if the attack button
251 was pressed. The risk was computed based on the number of allies within the blue area (i.e., 10,
252 20, or 30) and the probability of collateral damages based on their position (i.e., .50, .40, .30, .20,
253 or .10). The scale ranged from yellow (very low risk) to red (very high risk).

254 Finally, in *Level 2* of Autonomy, the visual assistance of *Level 0* was still included, but in
255 addition participants were assisted by a decision-support system that on each trial made a yes/no
256 recommendation on the best decision to make (see Fig. 1, right bottom). This recommendation was
257 based in each trial on the choice associated with the lowest expected losses in terms of allied forces.
258 When the expected losses were lower for pressing the attack button the 'yes' cue was highlighted,
259 while the 'no' cue was highlighted when the expected losses were lower by pressing the no attack

260 button (except during the trial where the number of allies was 10 and the probability was .50, in
261 which expected losses were equal for both choice). Participants were not informed on the
262 computation on which the recommendations were determined. Overall, each block consisted of 15
263 *No Risk* trials, 15 *No Enemy* trials, and 15 *Moral DM* trials.

264 The experimental setup of this task is shown in Fig. 1 (left): (1) First a loading bar was
265 presented for delay chosen randomly between 1000ms and 2000ms to signal a new trial to the
266 participants. (2) This was followed by a blackout screen for 500ms and then the presentation of
267 the 50 by 50 grid, displayed for 15000ms or until the participant pressed one of the two response
268 keys. (3) Participants were asked to confirm their choice by pressing the selected response key
269 again, or they had the possibility to change their choice by pressing the other key. (4) Responses
270 were followed by the presentation of blackout screen for a random duration of either 200, 500, or
271 800ms, and a tone (frequency: 400Hz) for 200ms. Finally, (5) participants were asked to report the
272 duration of the interval between their confirmation choice and the tone on a horizontal scale
273 ranging from 0 ms to 1000 ms. Trials were separated by an interval of 1000ms.

274

275 2.3. *Measurements and analysis*

276 We used five dependent variables in this study: Decision, Utilitarian Choice, Response
277 Time, Agency, and Subjective Responsibility. Decision (A1) was expressed by the proportion of
278 trials on which participants decided to attack. Utilitarian Choice (UC) was expressed by the
279 proportion of choices implying the lowest expected losses (in percentage). Response Time (RT)
280 was the mean response time (in seconds) on each trial. SoA was measured by Intentional Binding
281 (IB, in milliseconds). IB was computed by subtracting each interval estimate from the mean actual
282 response-tone interval (500ms) and averaged these scores for each Uncertainty X Autonomy

283 condition. Each block of Autonomy ended with a subjective judgment of responsibility (SubjA),
284 in which participants were asked to indicate how much they felt responsible of the decisions they
285 made on a scale from -100 (not responsible at all) to 100 (entirely responsible)¹.

286 Statistical analyses were performed using JASP version 0.17.2. We performed separate
287 repeated-measures ANOVAs for A1, UC, RT, and IB with Uncertainty (*No Risk trials, No Enemy*
288 *trials, Moral Decision-Making trials*) and Autonomy (*Level 0, Level 1, Level 2*) as within-subject
289 factors. In addition, SubjA was compared by means of a repeated measures ANOVA with
290 Autonomy as within-subject factors. For each dependent variable, only data of participants within
291 +/- 2.5 SDs were considered. Greenhouse-Geisser correction was applied where sphericity was
292 violated. We assessed Moral decision-making by several indicators.

293 The primary focus of our analysis concerned the presence of a main effect of Uncertainty
294 on 1) A1, for which we expected a higher percentage of attacks during No Risk trials and a lower
295 percentage during no Enemy trials, 2) UC, for which we expected an increased number in the No
296 Risk and No Enemy (control) trials compared to the Moral Decision-Making trials, and 3) RT,
297 with expected shorter response time in the No Risk and No Enemy trials compared to the Moral
298 Decision-Making trials. These effects were expected to provide evidence of the moral conflict
299 produced by the scenarios. Regarding IB, following the results of Moretto et al. (2011), we
300 expected a main effect of Uncertainty with shorter time interval, indicating an increase of SoA,
301 during *Moral Decision-Making* trials in comparison with the two control trials. We also expected
302 a main effect of Autonomy on 1) UC, with an increased rate of UC with the level of autonomy of
303 the task, 2) RT, congruent with the main effect of Uncertainty, 3) IB, with less IB, indicating a
304 decrease of SoA, with increased level of autonomy in line with the conclusions of Berberian et al.

¹ This range is commonly used in human contingency assessment (for recent examples, see Prével et al., 2021 or Vaghi et al., 2019).

305 (2012), and 4) SubjA, with lower SubjA with increased level of autonomy. These effects were
306 expected to provide evidence of the effectiveness of the paradigm we developed. The threshold
307 selected for significance was $p < .05$ with a two-tailed approach. Raw data, scripts, and processed
308 data can be found on the Open Science Framework (Salatino et al., 2023).
309
310

311 **3. Results**

312 *3.1. Analyses on A1 decisions*

313 The analysis on A1 decisions (i.e. the proportion of attacks) (Fig. 2) revealed a main effect
314 of Uncertainty ($F(1.06, 27.72) = 926.70, p < .001, \eta^2 = .97$) and post hoc tests showed that all
315 comparisons were significant (all p s $< .001$) with more a1 choices during *No Risk* trials (mean =
316 99.48, SEM = .20) in comparison with *Moral DM* trials (mean = 54.24, SEM = 1.77) and *No*
317 *Enemy* trials (mean = .81, SEM = .52). However, the analysis revealed no significant effect of
318 Autonomy ($F(1.78, 46.39) = 2.80, p = .07, \eta^2 = .09$) on A1, as well as no significant interaction
319 between Uncertainty and Autonomy ($F(2.76, 71.98) = 1.17, p = .32, \eta^2 = .04$).

320

321 *3.2. Analyses on Utilitarian Choice (UC)*

322 The analysis on UC (i.e., the choices implying the lowest expected losses, Fig. 3, Panel A)
323 revealed a significant effect of Uncertainty ($F(1.18, 31.96) = 95.76, p < .001, \eta^2 = .78$). Post hoc
324 tests showed a significant difference between *Moral DM* trials (mean = 84.02, SEM = 1.09) and
325 *No Risk* trials (mean = 99.30, SEM = 0.27), and between *Moral DM* and *No Enemy* trials (mean =
326 99.21, SEM = .50) (all p s $< .001$), with a reduced number of UC during *Moral DM* trials, but not
327 between *No Risk* and *No Enemy* trials ($p = 1.000$). The analysis (Fig. 3, Panel B) also revealed a
328 significant effect of Autonomy ($F(1.76, 47.55) = 8.67, p < .001, \eta^2 = .24$). Post hoc tests showed
329 no significant difference between *Level 0* (mean = 94.10, SEM = 1.07) and *Level 1* (mean = 92.59,
330 SEM = 1.17) ($p = .17$) and between *Level 0* and *Level 2* (mean = 95.84, SEM = .88) ($p = .089$). A
331 significant difference was found in the proportion of UC between *Level 1* and *Level 2* ($p < .001$)
332 with more UC on *Level 2*. Finally, the analysis revealed no significant interaction between
333 Uncertainty and Autonomy ($F(2.36, 63.80) = 2.48, p = .08, \eta^2 = .08$).

334

335 *3.3. Analyses on Response Time (RT)*

336 The analysis on RT (Fig. 4, Panel A) revealed a significant effect of Uncertainty ($F(1.73,$
337 $46.95) = 68.76, p < .001, \eta^2 = .71$) with post hoc tests showing that all comparisons were
338 significant (all $p_s \leq 0.36$), with longer RT during *Moral DM* trials (mean = 4.62, SEM = .20) than
339 during *No Risk* (mean = 2.27, SEM = .12) and *No Enemy* trials (mean = 2.82, SEM = .14). In
340 addition, the analysis revealed no significant effect of Autonomy ($F(1.68, 45.36) = 0.667, p = .49,$
341 $\eta^2 = .02$), but a significant interaction (Fig. 4, Panel B) between Uncertainty and Autonomy (F
342 $(3.43, 92.74) = 3.81, p = .009, \eta^2 = .12$), with a simple main effect of Autonomy in *No Enemy*
343 trials ($p = 0.03$), but not in other Uncertainty conditions (all $p > .37$).

344

345 *3.4. Analyses on Intentional Binding (IB)*

346 The analysis on IB (Fig. 5) revealed a significant effect of Uncertainty ($F(1.56, 46.87) =$
347 $8.12, p = .002, \eta^2 = .21$) with post hoc tests showing a significant difference between *Moral DM*
348 trials (mean = 118.85, SEM = 13.58) and *No Risk* trials (mean = 147.04, SEM = 13.99) and
349 between *Moral DM* and *No Enemy* trials (mean = 145.53, SEM = 13.69) (all $p < 0.005$), with
350 shorter intervals reported in *No Risk* and *No Enemy trials* in comparison with *Moral Decision-*
351 *Making* trials. No significant differences were found between *No Risk* and *No Enemy* trials ($p =$
352 1.000). Concerning Autonomy, the analysis revealed no significant effect ($F(1.76, 52.82) = 0.67,$
353 $p = .495, \eta^2 = .022$) and no significant interaction between Uncertainty and Autonomy ($F(3.52,$
354 $105.86) = 1.29, p = .27, \eta^2 = .04$).

355

356 *3.5 Analyses of subjective judgment of responsibility (SubjAs)*

357 The analysis on SubjAs (i.e., how much participants felt responsible of the decisions made,
358 Fig. 6) revealed a significant effect of Autonomy ($F(1.61, 45.15) = 15.72, p < .001, \eta^2 = .36$).
359 Post hoc tests showed a significant difference between *Level 0* (mean = 84.13, SEM = 3.04) and
360 *Level 1* (mean = 71.72, SEM = 4.60) ($p = .006$) and between *Level 0* and *Level 2* (mean = 58.62,
361 SEM = 69) ($p < .001$), with larger subjective responsibility rating during *Level 0* in both cases, but
362 not between *Level 1* and *Level 2* ($p = .065$).

363

364

365 **4. Discussion**

366 In the present pilot study, we aimed to develop a new paradigm to investigate how the sense
367 of agency and (moral) decision-making are influenced by the type of input received from an
368 intelligent autonomous system. To this end, we programmed a task in which participants, in the
369 role of drone operators, had to decide whether (or not) to initiate an attack on a simulated
370 battlefield, during different trial types and with the support of an intelligent system at three levels
371 of autonomy. Ultimately, the overall goal of this research agenda was to develop a paradigm to
372 investigate the mechanisms involved in human-AI interactions in the context of morally
373 challenging situations and to better understand the determining factors in the effect of (inputs
374 received from) autonomous systems on the decisions of human agents faced with moral choices.

375 Our results show that our new paradigm was sensitive enough to discriminate between
376 moral and non-moral situations. Indeed, participants in our study had a .54 likelihood to initiate an
377 attack in our moral situations, while they refrained from attacking when there were no enemies,
378 and systematically attacked in situations where an enemy was present, but an attack posed no risk
379 for allied troops. Furthermore, the moral situations were characterized by less utilitarian choices
380 than our control condition, reflecting the uncertainty of the situation. Notably, the assistance of the
381 autonomous system increased the number of utilitarian choices, demonstrating the influence of the
382 machine on decision-making. Finally, the moral situations were characterized by a longer reaction
383 time, indicating the participants' hesitation when they had to decide to fire or to not fire when
384 allies' life were at stake. Thus, this pilot study thus provides a preliminary paradigm for
385 investigating research questions related to the effects of human-machine interactions in moral
386 decision-making situations.

387 Our results were partially consistent with our expectations, as the number of attacks
388 increased when there was no risk and decreased when there was no enemy, confirming that our
389 control trials were working properly as well. However, no effects related to the level of autonomy
390 were found regarding the proportion of attacks. One possibility is that human choices, when made
391 in the context of moral dilemmas, are not influenced by the level of autonomy of a system, contrary
392 to choices made in a context without these dilemmas. However, this would be surprising, given
393 that previous studies conducted in military scenarios (e.g., Chen & Joyner, 2009; Rovira et al.,
394 2007) have found an influence of the level of autonomy on human performance and decision.
395 Another interpretation for the lack of an effect of level of autonomy on the rate of A1 choices relies
396 on the way the task was designed (and in particular the possibility for participants to calculate
397 maybe too easily the expected losses for both alternatives in each choice), resulting in ceiling/floor
398 effects in the control conditions and an A1 rate of around 50% during *Moral Decision-Making*
399 trials.

400 With respect to the Utilitarian Choice, i.e., the proportion of choices that lead to the least
401 loss of allies, we expected an increase in UC in trials without moral conflict and an effect of
402 autonomy leading to an increase in these choices at the highest level of Autonomy (as evidence of
403 an effect of autonomous system on moral decision). Our results confirmed our expectations, with
404 utilitarian choices increasing as a function of Uncertainty and Autonomy. Considering that the
405 recommendations made with the highest level of system support were based on the lower expected
406 losses computed, our results showed that our participants' moral choices were significantly
407 changed by the input received from the system. This suggests that human choices can be influenced
408 by the recommendations received from a decision support system, independently of the morally-
409 unmorally challenging nature of the situation.

410 In relation to the Response Time, we expected that participants would take a longer time
411 to make a decision on trials with a moral conflict and Autonomy would shorten participants'
412 response times. Our results confirmed our expectation, with RTs being longer in moral situations,
413 suggesting that participants took longer to make a decision when a moral conflict was present, and
414 that the highest level of system support shortened their response time in the No-Enemy trials. This
415 last result is consistent with previous findings from laboratory experiments showing that
416 autonomous systems can help users in detection tasks (e.g., Goh et al., 2005), although it is
417 surprising that this effect was found only for the absence of target but not for the presence of target
418 during trials without risk, which is somehow inconsistent with previous findings (e.g., Chavaillaz
419 et al., 2018). One possible reason for this result could be that the target used in our task (i.e., a
420 white square among grey squares on a black background) was too noticeable and thus too easy to
421 recognize for the *Level 1* and *Level 2* functions to be useful to participants.

422 Regarding SoA, given that a decrease in agency in human-machine interactions has been
423 previously reported (Berberian et al., 2012; Vantrepotte et al., 2022), we firstly expected that
424 participants would show a decrease in the Intentional Binding and subjective sense of
425 responsibility at higher levels of support, indicating a decrease in the SoA. Consistent with our
426 hypothesis, our results showed a decrease in the SoA at the explicit level with higher levels of
427 support. This result is also consistent with the recently described human tendency to attribute moral
428 responsibility to non-human agents which may lead people to be willing to blame them (Furlough,
429 et al., 2021; Kneer and Stuart, 2021; Liu and Du, 2022). Nevertheless, our results did not show a
430 decrease in SoA at the implicit level, which is inconsistent with our expectations and previous
431 studies (Berberian et al., 2012). However, this discrepancy is not completely surprising
432 considering that a dissociation between the two levels of measures in the SoA has already been

433 reported in previous studies (e.g., Synofzik et al., 2008; Moore and Obhi, 2012; Saito et al., 2015).
434 It has been suggested that at the explicit level a higher-order conceptual judgement of being an
435 agent is formed and that this aspect of SoA is closely related to higher-level sources of information
436 such as social and contextual cues (Synofzik et al., 2008), suggesting that Intentional Binding and
437 explicit judgments of agency do not share the same processes (Dewey & Knoblich, 2014). Since
438 the two measurement systems are separable (Saito et al., 2015), it is possible that dissociation
439 between implicit and explicit measurements occurred in our study. In particular, in our new
440 paradigm the three autonomy conditions may not have been so different as to yield a significant
441 difference at the implicit level of agency, but only at the explicit level.

442 Still regarding SoA, consistent with previous findings (Moretto et al., 2011), we also
443 expected an increase in SoA during Moral Decision-Making trials. Consequently, we expected that
444 moral decision making would also be affected. Because SoA appears to be closely related to moral
445 responsibility (Moretto et al., 2011; Caspar, et al., 2016) and this is reduced by the level of
446 autonomy of the machine, we expected that a decrease in the sense of responsibility would lead to
447 a change in the number of attacks at the highest level of system autonomy. However, contrary to
448 our expectations, the results showed a significant decrease in the SoA in the *Moral Decision-*
449 *Making* trials at the implicit level (i.e., in the IB). One possible explanation for these results is
450 related to the human tendency to take more responsibility for positive than for negative events,
451 which seems to be a mechanism for increasing self-esteem (Bradley, 1978; Greenberg et al., 1992;
452 Yoshie and Haggard, 2013). However, it has been pointed out there is a tendency to overestimate
453 one's agency, and that this bias is stronger when the outcome of an action is positive rather than
454 neutral or negative (Wegner & Wheatley, 1999; Haggard, 2017). Because the risk of a potential
455 hit to allies was constantly present in the Moral Decision-Making trials in our task, it is possible

456 that participants in these trials had a reduced SoA and responsibility, and disengaged from the
457 situation due to the risk of negative dramatic consequences of their actions. -Alternatively, it is
458 also possible that these results are related to the young cadets' lower SoA, which has already been
459 described by Caspar and colleagues (Caspar et al., 2018). However, these results should be taken
460 with caution, considering a flaw in the preparation of the Matlab script used to run the experiment,
461 and the possibility that the time intervals used for the time estimations (200, 500, and 800ms) were
462 not completely assigned equally across Uncertainty conditions (i.e., that each time interval are
463 shown five times per Uncertainty condition). Indeed, the program was designed to generate for
464 each three blocks (corresponding to the three levels of Autonomy) 15 presentations of each interval
465 duration (45 intervals in total) presented in a random order across trials (45 trials/block). Albeit
466 randomly presented across Uncertainty conditions, it is possible that the number of presentations
467 of each interval duration was not perfectly the same across Uncertainty condition. Future
468 investigations will be therefore necessary to determine whether the SoA decreases when human
469 subjects interact with autonomous systems when making moral decisions compared to situations
470 that do not pose a moral challenge, or whether this result is due to the methodological flaw
471 observed in our experiment.

472 To summarize, our results show that human choices made in morally - challenging scenarios
473 can be differentially influenced by the recommendations received from different level of
474 autonomous systems, replicating and extending previous findings (e.g., Chen & Joyner, 2009;
475 Rovira et al., 2007 for studies in the military domain). This was measured by the small but
476 significant difference in the number of utilitarian choices between conditions and the decrement
477 in response time in the No Enemy trials. Interestingly, this effect was completed by a decrement
478 in the subjective measure of agency with higher levels of autonomy, which is consistent with

479 previous research (Barberian et al., 2012; Vantrepotte et al., 2022). At the same time, however,
480 our results show some inconsistencies with previous studies (Moretto et al., 2011; Barberian et al.,
481 2012), with less agency measured in the moral decision-Making trials and no significant difference
482 across Level of Autonomy at the implicit level of agency (IB). Thus, further experiments need to
483 be conducted to determine if the inconsistent results we found were related to the design of the
484 experiment as we suggested above. In particular, in addition to the change need on the Intentional
485 Binding measure, the scenario we used was closer to impersonal/neutral stimuli rather than a
486 moral/emotional context. For example, the radar screen shown to participants was quite schematic
487 and likely did not allow participants to properly imagine the context of the choices they were
488 making. In addition, they did not know the number of victims following their decision, and the
489 victims were not clearly shown as individuals. Thus, it could be that the images we used did not
490 have enough emotional content to reinforce/enhance the SoA. Since the sense of agency could be
491 also affected by the actions' outcome, which is missing in the current version of the task, to
492 overcome this issue, future research could improve our paradigm by using a less neutral task and
493 content with more moral and emotional valence.
494

495 **5. Conclusion**

496 Considering the increasing presence of intelligent autonomous systems in our daily lives, it is
497 crucial to conduct further research to better understand the implications in sensitive domains such
498 as the military context to provide input for the successful design of innovative automated systems.
499 Our findings suggest that the level of system autonomy influences participants' moral decision-
500 making and that input received from an intelligent autonomous system influences SoA. By
501 developing a valid paradigm for assessing the impact of human-machine interaction on moral
502 decision-making in the military, with the present study we pave the way for further lines of
503 research on the influence of autonomous systems on human moral behavior, considering the
504 current lack of research on this issue.

505

506

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

511
512 **Funding:** This research was funded by Belgian Defense – Royal Higher Institute of Defense, grant
513 number HFM20-03.

514

515 **Reference**

516 Ayoub, J., Zhou, F., Bao, S., & Yang, X. J. (2019, September). From manual driving to
517 automated driving: A review of 10 years of autou. In *Proceedings of the 11th international*
518 *conference on automotive user interfaces and interactive vehicular applications* (pp. 70-90).

519
520 Anderson, E., Fannin, T., & Nelson, B. (2018, September). Levels of aviation autonomy.
521 In *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)* (pp. 1-8). IEEE.

522
523 Arkin, R. C., Ulam, P., & Wagner, A. R. (2011). Moral decision making in autonomous
524 systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the*
525 *IEEE*, *100*(3), 571-589.

526
527 Berberian, B., Sarrazin, J. C., Le Blaye, P., & Haggard, P. (2012). Automation technology
528 and sense of control: a window on human agency. *PloS one*, *7*(3), e34075.
529 <https://doi.org/10.1371/journal.pone.0034075>.

530
531 Berberian, B. (2019). Man-Machine teaming: a problem of Agency. *IFAC-*
532 *PapersOnLine*, *51*(34), 118-123. <https://doi.org/10.1016/j.ifacol.2019.01.049>.

533
534 Blackwood, N. J., Bentall, R. P., Simmons, A., Murray, R. M., & Howard, R. J. (2003).
535 Self-responsibility and the self-serving bias: an fMRI investigation of causal
536 attributions. *NeuroImage*, *20*(2), 1076-1085. [https://doi.org/10.1016/S1053-8119\(03\)00331-8](https://doi.org/10.1016/S1053-8119(03)00331-8).

537 Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of
538 the fact or fiction question. *Journal of personality and social psychology*, 36(1), 56.

539
540 Burin, D., Pyasik, M., Salatino, A., & Pia, L. (2017). That's my hand! Therefore, that's my
541 willed action: How body ownership acts upon conscious awareness of willed
542 actions. *Cognition*, 166, 164-173. <https://doi.org/10.1016/j.cognition.2017.05.035>.

543
544 Caspar, E. A., Christensen, J. F., Cleeremans, A., & Haggard, P. (2016). Coercion changes
545 the sense of agency in the human brain. *Current biology*, 26(5), 585-592.
546 <https://doi.org/10.1016/j.cub.2015.12.067>.

547
548 Caspar, E. A., Cleeremans, A., & Haggard, P. (2018). Only giving orders? An experimental
549 study of the sense of agency when giving or receiving commands. *PloS one*, 13(9), e0204027.
550 <https://doi.org/10.1371/journal.pone.0204027>.

551
552 Caspar, E. A., Ioumpa, K., Arnaldo, I., Di Angelis, L., Gazzola, V., & Keysers, C. (2022).
553 Commanding or being a simple intermediary: how does it affect moral behavior and related brain
554 mechanisms?. *eneuro*, 9(5).

555
556 Chan, C. Y. (2017). Advancements, prospects, and impacts of automated driving
557 systems. *International journal of transportation science and technology*, 6(3), 208-216.
558 <https://doi.org/10.1016/j.ijtst.2017.07.008>.

559

560 Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2018). Automation in visual
561 inspection tasks: X-ray luggage screening supported by a system of direct, indirect or adaptable
562 cueing with low and high system reliability. *Ergonomics*, *61*(10), 1395-1408.
563 <https://doi.org/10.1080/00140139.2018.1481231>.

564
565 Chialastri, A. (2012). *Automation in aviation*. INTECH Open Access Publisher.

566
567 Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of
568 moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4),
569 1249-1264. <https://doi.org/10.1016/j.neubiorev.2012.02.008>.

570
571 Christensen, J. F., Di Costa, S., Beck, B., & Haggard, P. (2019). I just lost it! Fear and
572 anger reduce the sense of agency: a study using intentional binding. *Experimental brain*
573 *research*, *237*, 1205-1212. <https://doi.org/10.1007/s00221-018-5461-6>.

574
575 Chen, J. Y., & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of
576 imperfect automation and individual differences. *Human Factors*, *54*(2), 157-174.
577 <https://doi.org/10.1177/0018720811435843>.

578
579 Chen, J. Y., & Joyner, C. T. (2009). Concurrent performance of gunner's and robotics
580 operator's tasks in a multitasking environment. *Military Psychology*, *21*(1), 98-113.

581

582 Cohn, A., Gesche, T., & Maréchal, M. A. (2022). Honesty in the digital age. *Management*
583 *Science*, 68(2), 827-845.

584

585 Coyle, D., Moore, J., Kristensson, P. O., Fletcher, P., & Blackwell, A. (2012, May). I did
586 that! Measuring users' experience of agency in their own actions. In *Proceedings of the SIGCHI*
587 *conference on human factors in computing systems* (pp. 2025-2034).
588 <https://doi.org/10.1145/2207676.2208350>.

589

590 Cushman, F. (2013). Action, outcome, and value: A dual-system framework for
591 morality. *Personality and social psychology review*, 17(3), 273-292.
592 <https://doi.org/10.1177/1088868313495594>.

593

594 Cushman, F., Kumar, V., & Railton, P. (2017). Moral learning: Psychological and
595 philosophical perspectives. *Cognition*, 167, 1-10. <https://doi.org/10.1016/j.cognition.2017.06.008>.

596

597 de Melo, C. M., Marsella, S., & Gratch, J. (2018). Social decisions and fairness change
598 when people's interests are represented by autonomous agents. *Autonomous Agents and Multi-*
599 *Agent Systems*, 32, 163-187. <https://doi.org/10.1007/s10458-017-9376-6>.

600

601 de Melo, C. M., Marsella, S., & Gratch, J. (2019). Human cooperation when acting through
602 autonomous machines. *Proceedings of the National Academy of Sciences*, 116(9), 3482-3487.
603 <https://doi.org/10.1073/pnas.1817656116>.

604

605 Dewey, J. A., & Knoblich, G. (2014). Do implicit and explicit measures of the sense of
606 agency measure the same thing?. *PloS one*, 9(10), e110118.
607 <https://doi.org/10.1371/journal.pone.0110118>.

608
609 Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level
610 of control in automation. *Human factors*, 37(2), 381-394.
611 <https://doi.org/10.1518/001872095779064555>.

612
613 Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance,
614 situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3), 462-492.
615 <https://doi.org/10.1080/001401399185595>.

616
617 Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation
618 research. *Human factors*, 59(1), 5-27. <https://doi.org/10.1177/0018720816681350>.

619
620 Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible
621 statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior*
622 *research methods*, 39(2), 175-191.

623
624 Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing blame to robots: I. The
625 influence of robot autonomy. *Human factors*, 63(4), 592-602.
626 <https://doi.org/10.1177/0018720819880641>.

627

628 Greenberg, J., Pyszczynski, T., Burling, J., & Tibbs, K. (1992). Depression, self-focused
629 attention, and the self-serving attributional bias. *Personality and Individual Differences*, 13(9),
630 959-965. [https://doi.org/10.1016/0191-8869\(92\)90129-D](https://doi.org/10.1016/0191-8869(92)90129-D).

631
632 Goh, J., Wiegmann, D. A., & Madhavan, P. (2005). Effects of automation failure in a
633 luggage screening task: a comparison between direct and indirect cueing. In *Proceedings of the*
634 *Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 3, pp. 492-496). Sage CA:
635 Los Angeles, CA: SAGE Publications.

636
637 Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious
638 awareness. *Nature neuroscience*, 5(4), 382-385. <https://doi.org/10.1038/nn827>.

639
640 Haggard, P., & Tsakiris, M. (2009). The experience of agency: Feelings, judgments, and
641 responsibility. *Current Directions in Psychological Science*, 18(4), 242-246.
642 <https://doi.org/10.1111/j.1467-8721.2009.01644.x>.

643
644 Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews*
645 *Neuroscience*, 18(4), 196-207. <https://doi.org/10.1038/nrn.2017.14>.

646
647 Haslbeck, A., & Hoermann, H. J. (2016). Flying the needles: flight deck automation erodes
648 fine-motor flying skills among airline pilots. *Human factors*, 58(4), 533-545.
649 <https://doi.org/10.1177/0018720816640394>.

650

651 Imaizumi, S., & Tanno, Y. (2019). Intentional binding coincides with explicit sense of
652 agency. *Consciousness and Cognition*, 67, 1-15. <https://doi.org/10.1016/j.concog.2018.11.005>.

653
654 Jeannerod, M. (2003). The mechanism of self-recognition in humans. *Behavioural brain*
655 *research*, 142(1-2), 1-15. [https://doi.org/10.1016/S0166-4328\(02\)00384-4](https://doi.org/10.1016/S0166-4328(02)00384-4).

656
657 Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., ... & Choi, Y.
658 (2021). Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.
659 <https://doi.org/10.48550/arXiv.2110.07574>.

660
661 Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical
662 practice using clinical decision support systems: a systematic review of trials to identify features
663 critical to success. *Bmj*, 330(7494), 765.

664
665 Kneer, M., & Stuart, M. T. (2021). Playing the blame game with robots. In *Companion of*
666 *the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 407-411).
667 <https://doi.org/10.1145/3434074.3447202>.

668
669 Köbis, N., Bonnefon, J. F., & Rahwan, I. (2021). Bad machines corrupt good
670 morals. *Nature Human Behaviour*, 5(6), 679-685. <https://doi.org/10.1038/s41562-021-01128-2>.

671

672 Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2021). The corruptive
673 force of AI-generated advice. *arXiv preprint arXiv:2102.07536*.
674 <https://doi.org/10.48550/arXiv.2102.07536>.

675
676 Liu, P., & Du, Y. (2022). Blame attribution asymmetry in human–automation
677 cooperation. *Risk Analysis*, 42(8), 1769-1783. <https://doi.org/10.1111/risa.13674>.

678
679 MacMillan, J., Deutsch, S. E., & Young, M. J. (1997). A comparison of alternatives for
680 automated decision support in a multi-task environment. In *Proceedings of the Human Factors*
681 *and Ergonomics Society Annual Meeting* (Vol. 41, No. 1, pp. 190-194). Sage CA: Los Angeles,
682 CA: SAGE Publications. <https://doi.org/10.1177/107118139704100144>.

683
684 Mayer, M. (2015). The new killer drones: Understanding the strategic implications of next-
685 generation unmanned combat aerial vehicles. *International Affairs*, 91(4), 765-780.
686 <https://doi.org/10.1111/1468-2346.12342>.

687
688 Malik, R. A., & Obhi, S. S. (2019). Social exclusion reduces the sense of agency: Evidence
689 from intentional binding. *Consciousness and Cognition*, 71, 30-38.
690 <https://doi.org/10.1016/j.concog.2019.03.004>.

691
692 Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of
693 automated decision aids: The impact of degree of automation and system experience. *Journal of*

694 *Cognitive Engineering and Decision Making*, 6(1), 57-87.

695 <https://doi.org/10.1177/1555343411433844>.

696

697 Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: a
698 review. *Consciousness and cognition*, 21(1), 546-561.

699 <https://doi.org/10.1016/j.concog.2011.12.002>.

700

701 Moretto, G., Walsh, E., & Haggard, P. (2011). Experience of agency and sense of
702 responsibility. *Consciousness and cognition*, 20(4), 1847-1854.

703 <https://doi.org/10.1016/j.concog.2011.08.014>.

704

705 Mosier, K. L., & Manzey, D. (2019). Humans and automated decision aids: A match made
706 in heaven?. In *Human performance in automated and autonomous systems* (pp. 19-42). CRC Press

707

708 Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of
709 automation-induced 'complacency'. *The International Journal of Aviation Psychology*, 3(1), 1-23.

710

711 Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse,
712 abuse. *Human factors*, 39(2), 230-253. <https://doi.org/10.1518/001872097778543886>.

713

714 Pyasik, M., Salatino, A., Burin, D., Berti, A., Ricci, R., & Pia, L. (2019). Shared
715 neurocognitive mechanisms of attenuating self-touch and illusory self-touch. *Social cognitive and*

716 *affective neuroscience*, 14(2), 119-127. <https://doi.org/10.1093/scan/nsz002>.

717 Prével, A., Krebs, R. M., Kukkonen, N., & Braem, S. (2021). Selective reinforcement of
718 conflict processing in the Stroop task. *PloS one*, *16*(7), e0255430.
719 <https://doi.org/10.1371/journal.pone.0255430>.

720
721 Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on
722 decision making in a simulated command and control task. *Human factors*, *49*(1), 76-87.
723 <https://doi.org/10.1518/001872007779598082>.

724
725 Saito, N., Takahata, K., Murai, T., & Takahashi, H. (2015). Discrepancy between explicit
726 judgement of agency and implicit feeling of agency: Implications for sense of agency and its
727 disorders. *Consciousness and cognition*, *37*, 1-7. <https://doi.org/10.1016/j.concog.2015.07.011>.

728
729 Salatino, A., Prével, A., & Lo Bue, S. (2023, December 12). “Fire! Do Not Fire!”: A pilot-
730 new paradigm testing how autonomous systems affect agency and moral decision-making.
731 Retrieved from osf.io/98ycf

732
733 Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection
734 under time pressure and uncertainty: The case of in-flight icing. *Human factors*, *43*(4), 573-583.
735 <https://doi.org/10.1518/001872001775870403..>

736
737 Synofzik, M., Vosgerau, G., & Newen, A. (2008). I move, therefore I am: A new theoretical
738 framework to investigate agency and ownership. *Consciousness and cognition*, *17*(2), 411-424.
739 <https://doi.org/10.1016/j.concog.2008.03.008>.

740 Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker,
741 K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for
742 success. *NPJ digital medicine*, 3(1), 17. <https://doi.org/10.1038/s41746-020-0221-y>.

743
744 Vaghi, M. M., Cardinal, R. N., Apergis-Schoute, A. M., Fineberg, N. A., Sule, A., &
745 Robbins, T. W. (2019). Action-outcome knowledge dissociates from behavior in obsessive-
746 compulsive disorder following contingency degradation. *Biological Psychiatry: Cognitive*
747 *Neuroscience and Neuroimaging*, 4(2), 200-209. <https://doi.org/10.1016/j.bpsc.2018.09.014>.

748
749 Valdés, R. A., Comendador, V. F. G., Sanz, A. R., & Castán, J. P. (2018). Aviation 4.0:
750 more safety through automation and digitization. In *Aircraft technology*. IntechOpen.

751
752 Vantrepotte, Q., Berberian, B., Pagliari, M., & Chambon, V. (2022). Leveraging human
753 agency to improve confidence and acceptability in human-machine interactions. *Cognition*, 222,
754 105020. <https://doi.org/10.1016/j.cognition.2022.105020>.

755
756 Volz, K. M., & Dorneich, M. C. (2020). Evaluation of cognitive skill degradation in flight
757 planning. *Journal of Cognitive Engineering and Decision Making*, 14(4), 263-287.
758 <https://doi.org/10.1177/1555343420962897>.

759
760 Zanatto, D., Chattington, M., & Noyes, J. (2021). Human-machine sense of agency.
761 *International Journal of Human-Computer Studies*, 156, 102716.

762

763 Wang, H., Lewis, M., Velagapudi, P., Scerri, P., & Sycara, K. (2009, March). How search
764 and its subtasks scale in N robots. In *Proceedings of the 4th ACM/IEEE international conference*
765 *on Human robot interaction* (pp. 141-148). <https://doi.org/10.1145/1514095.1514122>.

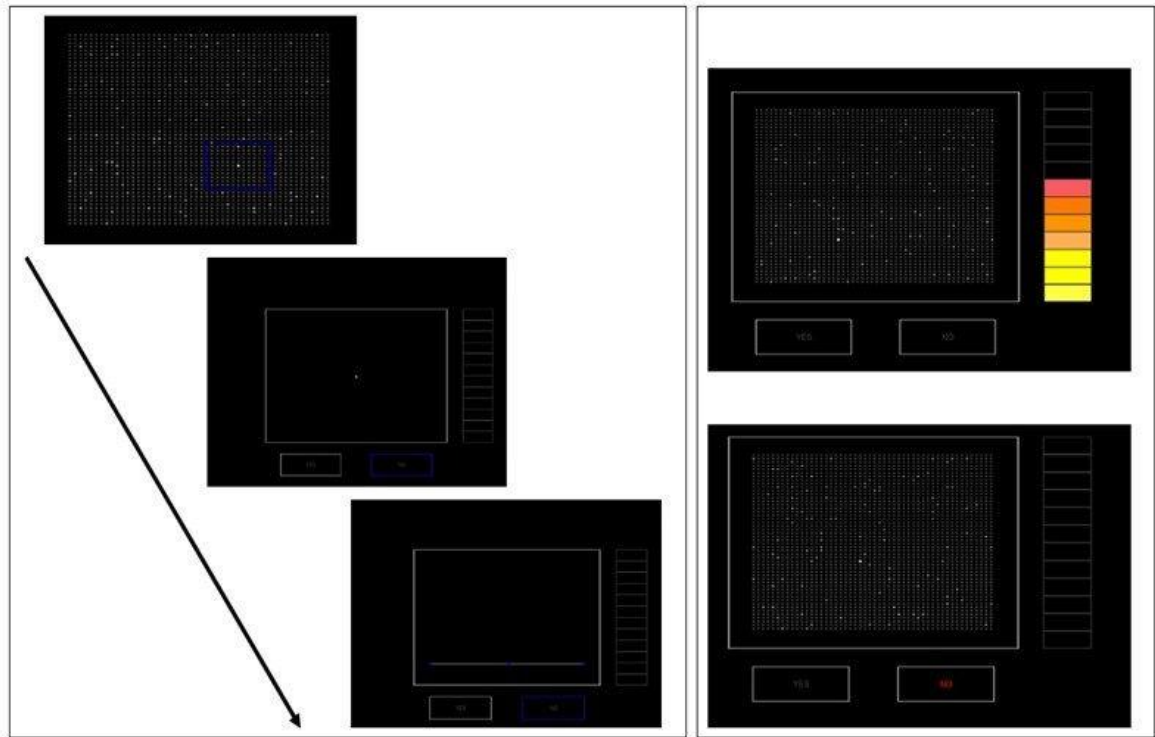
766
767 Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the
768 experience of will. *American psychologist*, *54*(7), 480. [https://doi.org/10.1037/0003-](https://doi.org/10.1037/0003-066X.54.7.480)
769 [066X.54.7.480](https://doi.org/10.1037/0003-066X.54.7.480).

770
771 Wright, J. L., Chen, J. Y., & Barnes, M. J. (2018). Human–automation interaction for
772 multiple robot control: the effect of varying automation assistance and individual differences on
773 operator performance. *Ergonomics*, *61*(8), 1033-1045.
774 <https://doi.org/10.1080/00140139.2018.1441449>.

775
776 Yoshie, M., & Haggard, P. (2013). Negative emotional outcomes attenuate sense of agency
777 over voluntary actions. *Current Biology*, *23*(20), 2028-2032.
778 <https://doi.org/10.1016/j.cub.2013.08.034>..

779

780 **Fig. 1.**



781

782 **Fig. 1. Experimental setup.** Each trial of the task consisted of a 50 x 50 grid of dark grey
783 cells on a black screen (**Left Panel**), representing a radar display informing of the position
784 of allies and enemies. Of the 2500 cells, 100 were colored light grey and represented the
785 position of a group of 10 allies, which varied in position from trial to trial.

786 In 66% of the trials, one of the 2500 cells was coloured white to represent an enemy and
787 participants were asked to choose whether or not to attack the enemy by pressing the left
788 (attack) or right (no attack) arrow key on the keyboard. The grid was displayed for 1500ms
789 or until the participant pressed one of the two response keys. After the response, a blackout
790 screen was displayed for a random duration (200, 500 or 800ms) and a tone (200ms).
791 Participants were asked to indicate the duration of the interval between their choice and the
792 tone on a horizontal scale from 0 ms to 1000 ms. *Level 1* of system autonomy is shown in
793 the right panel (**top**), where in addition to the basic visual assistance of *Level 0*, a scale has
794 been displayed indicating the risk in terms of losses of allied forces in the event of an attack.
795 In *Level 2* (**Lower Right Panel**), in addition to the visual assistance of *Level 0*, participants
796 were assisted by a decision support system that gave a yes/no recommendation for the best
797 decision to make.

798

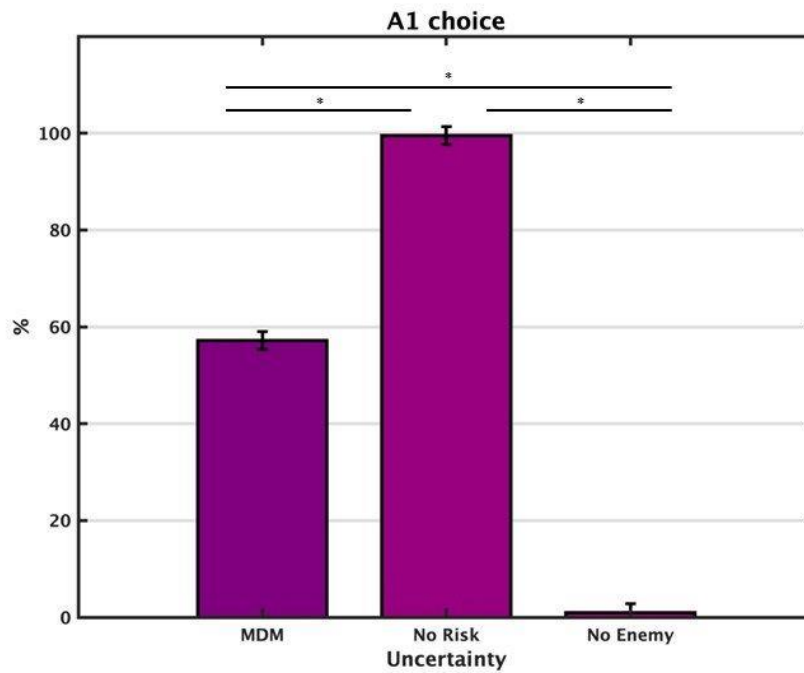
799

800

801

802

Fig. 2.



803

804

805

806

807

808

809

810

811

812

813

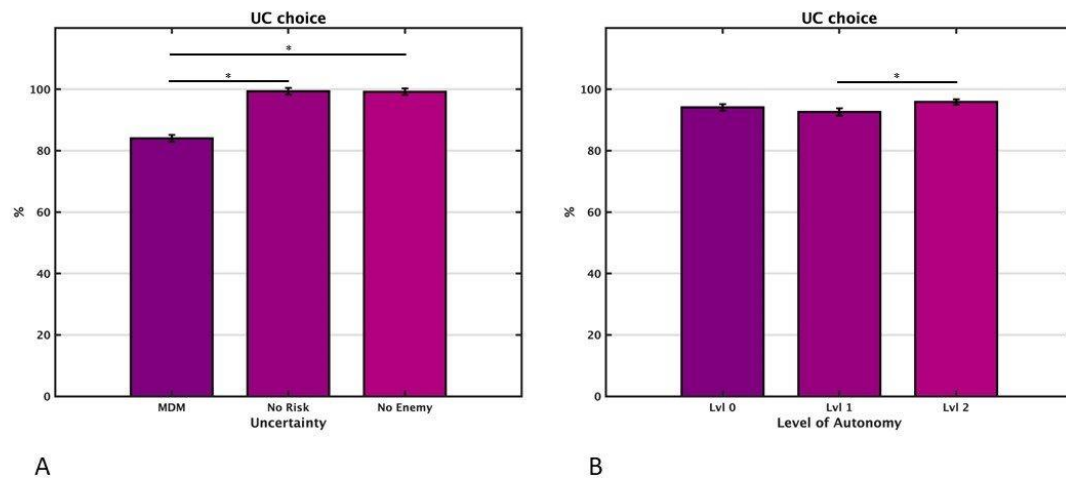
814

815

Fig. 2. Proportion of A1 action (i.e. attacks performed). A significant increase of A1 choices during *No Risk* trials in comparison with *Moral Decision-Making* and *No Enemy* trials were found (all $p < .001$). * = significant.

816

Fig. 3.



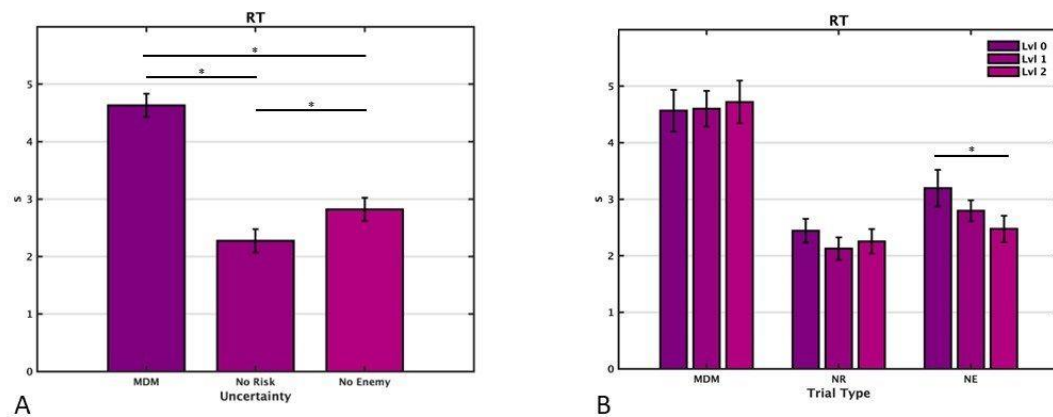
817

818 **Fig. 3. Proportion of UC (i.e. Utilitarian Choices).** A significant difference was found in
819 UC between *Moral Decision-Making* trials and *No Risk* trials, and between *Moral*
820 *Decision-Making* and *No Enemy* trials (all $p < .001$), with a reduced number of UC during
821 *Moral Decision-Making* trials, but not between *No Risk* and *No Enemy* trials (**Panel A**). A
822 significant difference was found in the proportion of UC between *Level 1* and *Level 2* ($p <$
823 $.001$) with more UC on *Level 2*. * = significant.

825

826

Fig. 4.



827

828 **Fig. 4. Response time (in seconds).** A significant increase in the RT was found during
829 *Moral Decision-Making trials* compared to the *No Risk* and *No Enemy trials* were found
830 (all $p < 0.36$, **Panel A**). In addition, a significant ($p = .009$) interaction was found between
831 Uncertainty and Autonomy (**Panel B**), with a simple main effect of Autonomy in *No Enemy*
832 *trials* ($p = 0.03$). * = significant.

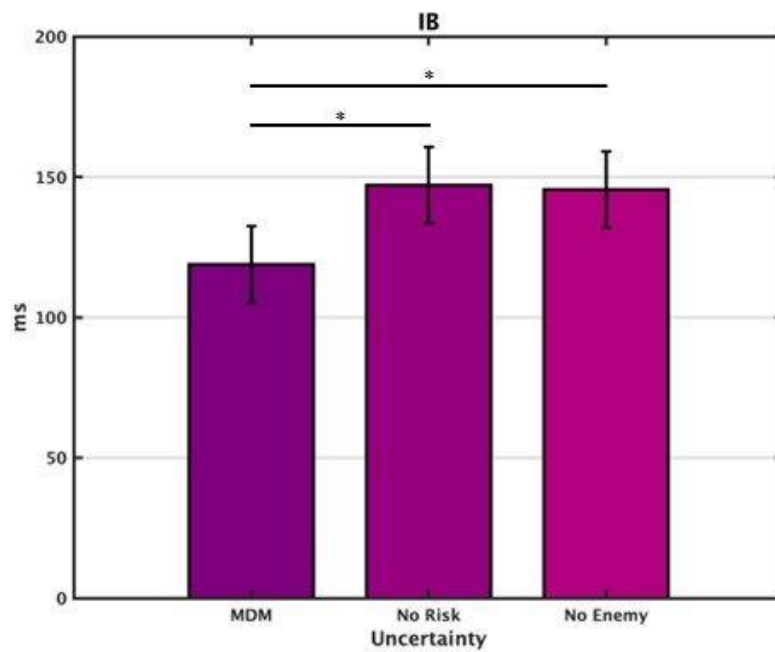
833

834

835

836

837 **Fig. 5.**



838

839 **Fig. 5. Intentional binding (IB).** A significant difference in the IB was found between
840 *Moral Decision-Making trials* and *No Risk trials*, and between *Moral Decision-Making*
841 *trials* and *No Enemy trials* (all $p < 0.005$), with shorter intervals reported in *No Risk* and
842 *No Enemy trials* compared to the *Moral Decision-Making trials*. * = significant.
843

844

845

846

847

848

849

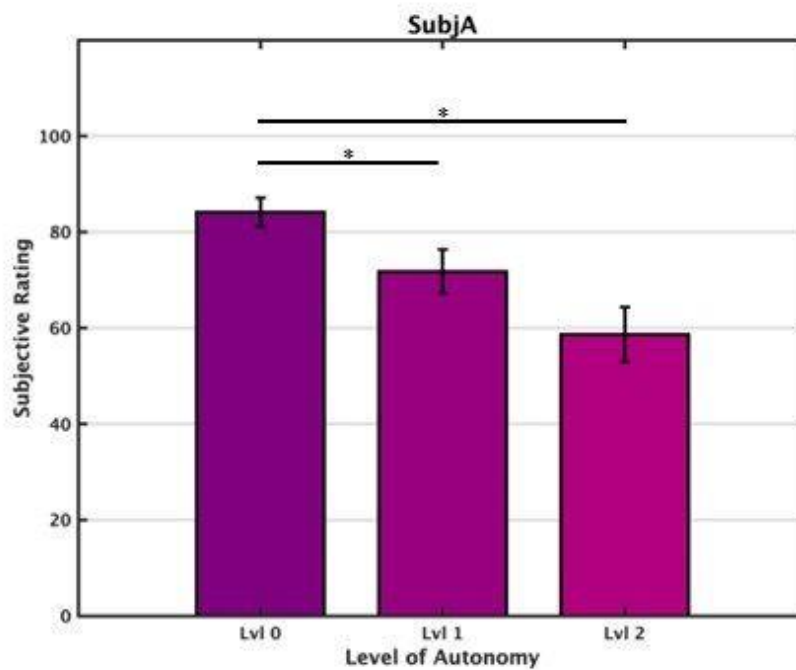
850

851

852

853

Fig. 6.

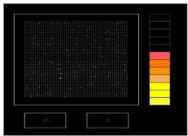
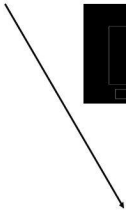
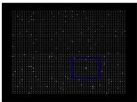


854

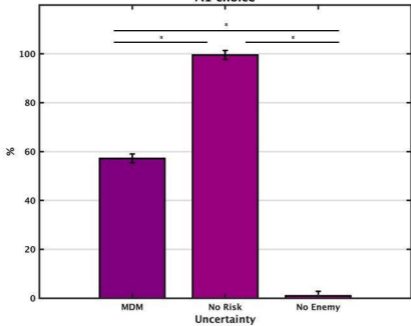
855 **Fig. 6. Subjective assessment of responsibility.** A significant decrease in subjective
856 judgement of responsibility was found in *Level 0* compared to *Level 1* ($p = .006$) and in
857 *Level 0* compared to *Level 2* ($p < .001$), with greater subjective judgement of responsibility
858 in *Level 0* compared to *Level 1* and *Level 2*. * = significant.

859

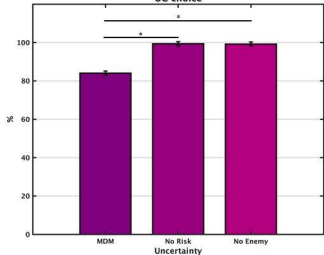
860



A1 choice

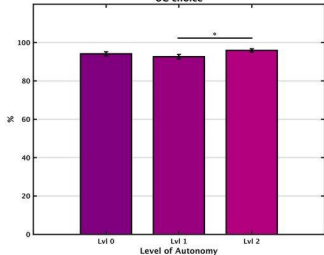


UC choice

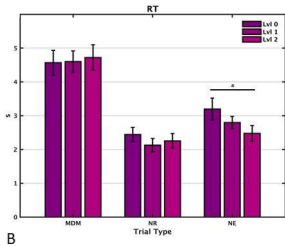
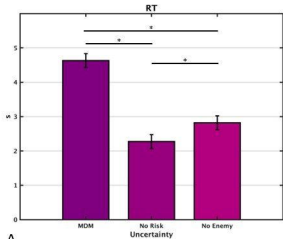


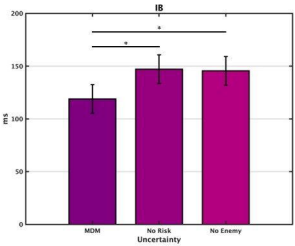
A

UC choice



B





SubjA

