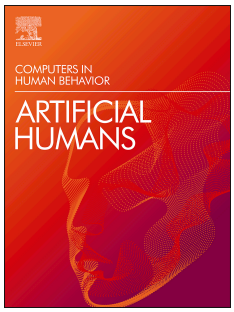
Making Moral Decisions With Artificial Agents As Advisors. An fNIRS Study

Eve Florianne Fabre, Damien Mouratille, Vincent Bonnemains, Grazia Pia Palmiotti, Mickael Causse



Making Moral Decisions With Artificial Agents As Advisors. An *f*NIRS Study

Eve Florianne Fabre, Damien Mouratille, Vincent Bonnemains, Grazia Pia Palmiotti, Mickael Causse

PII: S2949-8821(24)00056-2

DOI: https://doi.org/10.1016/j.chbah.2024.100096

Reference: CHBAH 100096

To appear in: *Computers in Human Behavior: Artificial Humans*

Received Date: 6 June 2024

Revised Date: 9 October 2024

Accepted Date: 11 October 2024

Please cite this article as: Fabre E.F., Mouratille D., Bonnemains V., Palmiotti G.P. & Causse M., Making Moral Decisions With Artificial Agents As Advisors. An *f*NIRS Study, *Computers in Human Behavior: Artificial Humans*, https://doi.org/10.1016/j.chbah.2024.100096.

Making Moral Decisions With Artificial Agents As Advisors. An *f*NIRS Study

Eve Florianne Fabre [1 *], Damien Mouratille [2], Vincent Bonnemains [3],

Grazia Pia Palmiotti [1] & Mickael Causse [1]

[1] ISAE Supaero, 10 avenue Edouard Belin, 31400 Toulouse, France

[2] Ecole National de l'Aviation Civile, 7 Avenue Edouard Belin, 31400 Toulouse, France

[3] DTIS, ONERA, 2 Avenue Edouard Belin, 31000 Toulouse, France

Keywords: Moral Dilemma; Decision-Making; Artificial Intelligence; Advisor; *f*NIRS

Author Note

The present research was funded by a postdoctoral grant awarded by the AXA research fund to Eve F. Fabre. We would like to thank Pr. Frédéric Dehais and Dr. Sébastien Scanella for their help with the *f*NIRS device.

* Correspondence should be addressed to Eve F. Fabre, e-mail: eve.fabre@isae-supaero.fr

Abstract

Artificial Intelligence (AI) is on the verge of impacting every domain of our lives. It is increasingly being used as an advisor to assist in making decisions. The present study aimed at investigating the influence of moral arguments provided by AI-advisors (i.e., decision aid tool) on human moral decision-making and the associated neural correlates. Participants were presented with sacrificial moral dilemmas and had to make moral decisions either by themselves (i.e., baseline run) or with AI-advisors that provided utilitarian or deontological arguments (i.e., AI-advised run), while their brain activity was measured using an *f*NIRS device. Overall, AI-advisors significantly influenced participants. Longer response times and a decrease in right dorsolateral prefrontal cortex activity were observed in response to deontological arguments than to utilitarian arguments. Being provided with deontological arguments by machines appears to have led to a decreased appraisal of the affective response to the dilemmas. This resulted in a reduced level of utilitarianism, supposedly in an attempt to avoid behaving in a less cold-blooded way than machines and preserve their (self-)image. Taken together, these results suggest that motivational power can led to a voluntary up- and down-regulation of affective processes along moral decision-making.

1. INTRODUCTION

1.1. The Revolution of Artificial Intelligence

The revolution of Artificial Intelligence (AI) is impacting almost every domain of our life. At home, social chatbots (Henkel et al., 2020; Pentina et al., 2023) are being increasingly used as personal assistants (e.g., Alexa, Siri, Chat GPT, Google Assistant), friendship companions (e.g., Replika, Anima, Kajiwoto, Microsoft XiaoIce) or even relational agents for (mental) healthcare (e.g., Woebot; Wysa; Youper; Stade et al., 2023). At work, intelligent decision support systems are now being used as advisors in healthcare, finance, and the military to name a few (Černevičienė & Kabašinskas, 2022; Macey-Dare, 2023; Shortliffe, & Sepúlveda, 2018; Wasilow & Thorpe, 2019). These systems allow the fusion and analysis of large amounts of data to reveal complex relationships, provide recommendations to workers and help them make faster and more informed decisions compared to when the data is analyzed manually (Eom & Kim, 2006). The increasing use of AI-agents as advisors, cooperators and sometimes even delegates, raises the question of whether and to what extent do AI decision aids (i.e., AI advisors) modulate human beings' judgment and decision-making (e.g., Collart et al., 2015; Crandall et al., 2018; Grgić-Hlača, et al., 2019; Hanson et al., 2024; Köbis et al., 2021; Ladak et al., 2024). The present study aimed at investigating this question, and more specifically whether moral advice provided by AI-advisors can influence human beings' moral decisions.

1.2. Using Moral Dilemmas to Study Moral Decision-Making

Moral dilemmas, particularly sacrificial ones, have been extensively used to explore moral judgment and decision-making, alongside the affective and cognitive processes underlining these judgments and decisions (e.g., Christensen & Gomila, 2012; Greene et al., 2001, 2015). When confronted with a sacrificial moral dilemma (Patil et al., 2021), one usually has to decide between a *utilitarian option*, which involves sacrificing the few to save the many (Mill, 1998), and a *deontological option*, which entails avoiding causing harm to individuals unrelated to the

situation and refraining from intervening (Kant, 2013). Individuals' moral decisions tend to vary depending on the dilemma, as reflected by the disparity in moral preferences commonly observed between dilemmas such as the trolley dilemma and the footbridge dilemma (Christensen & Gomila, 2012). In both dilemmas, one must envision that a trolley is hurtling at full speed along a track, where five workers stand, and is about to strike and kill these workers. In the trolley dilemma, the sole means of saving the workers is to pull a lever, which would divert the trolley onto a sidetrack where only one worker is present (Foot, 1978; Thomson, 1985). Conversely, in the footbridge dilemma, saving the workers involves pushing an overweight individual off the footbridge and into the path of the oncoming trolley (Thomson, 1976). Despite the fact that the two options to save the workers adhere to a utilitarian principle of sacrificing one life to save five, the greater majority of people opts for pulling the lever in the trolley dilemma but refuse to push the overweight person off the footbridge.

Exploring the reasons behind the preferences for utilitarian versus deontological options has been a central focus of numerous studies in the field of moral neuroscience (Casebeer, 2003; Greene et al, 2001, 2004, 2015; Moll et al., 2005). Under some circumstances, the utilitarian option can trigger a strong negative affective reaction, reflected by the activity of the ventromedial prefrontal cortex (vmPFC; Moll & de Oliveira-Souza, 2007) and the temporo-parietal junction (TPJ). This strong negative reaction occurs for instance when it involves killing someone as an intended means for the greater good like in the footbridge dilemma (Foot, 1978), and usually results in a preference for the deontological option. In other circumstances, such as when the death of an individual unrelated to the situation is a foreseen but unintended consequence (e.g., Sarlo et al., 2012), the utilitarian option triggers an increased activity of brain regions associated with working memory and problem solving, such as the dorsolateral prefrontal cortex (DLPFC; Moll et al., 2005), reflecting the greater cognitive effort made to control the negative affective response. In this case, the utilitarian option is more likely to be

preferred (Greene, 2015; Tassy et al., 2012; Zheng et al., 2018). Overall, this literature aligns with the dual process theory of moral judgment, which postulates that moral decisions result from two competing processing systems: a fast, automatic, and emotional System 1 and a slow, controlled and cognitively costly System 2 (Greene, 2015).

1.3. The Influence of Human and Artificial Advisors

Human beings are organized in social groups, wherein they constantly exchange knowledge (Gariépy et al., 2014) and seek advice (e.g., Polman & Ruttan, 2022). As a result, they exhibit a heightened sensitivity to the influence of their peers (Cialdini & Goldstein, 2004; Yu et al., 2021). They tend to align their judgments and behaviors with those of their peers (e.g., Braams et al., 2019; Chung et al., 2020; Fabre et al., 2022) and being swayed by their advice (e.g., Bonaccio & Dalal, 2006; Leong & Zaki, 2018; McCoy & Natsuaki, 2015). The extent of an advisor's influence is highly dependent on the advisee's trust, the advisors' expertise and reliability (Bonaccio & Dalal, 2006; for a meta-analysis see Bailey et al., 2023). Overall, neuroscience studies have shown an increased activity in the lateral orbitofrontal cortex and the vmPFC in response to advice that contradicts one's own opinions, and greater activity in the ventral striatum and the anterior cingulate cortex when the difference in opinion with the advisor is low (e.g., Biele et al., 2011; Campbell-Meiklejohn et al., 2010; Izuma, 2017; Meshi et al., 2012). The activity of these brain regions is known to increase in response to respectively negative and positive outcomes, suggesting that sharing similar opinions is socially rewarding for advisees (Izuma, 2017). Interestingly, the activity of left lateral orbitofrontal cortex and the DLPFC was found to be positively correlated to the influence of the advice (Meshi et al., 2012; Zhang & Gläscher, 2020).

In the last decades, an important number of studies have demonstrated that non-human agents can also have a significant influence on humans (for an overview, see Hertz & Wiese, 2019). The influence of non-human agents' advice was found to vary as a function of the non-

human advisors' perceived expertise (as for human advisors), but also depending on their similarity to humans in terms of appearance, empathy or non-verbal cues (e.g., Benitez et al, 2017; Chidambaran et al., 2012; Goodyear et al., 2016, 2017). Part of this literature focused specifically on the influence of non-human agents on moral judgment and/or decision-making (Jackson & Williams, 2019; Köbis et al., 2021; Leib et al., 2024; Straßmann et al., 2020). Some studies have shown that artificial advisors can have a corruptible impact on humans, increasing for instance their tendency to break ethical rules (e.g., cheat or inflict harm to another individual) for their own profit (e.g., monetary gain; Köbis et al., 2021; Sandoval et al., 2016). While there is a substantial body of literature examining how humans judge the decisions of machines in sacrificial moral dilemma situations (e.g., Awad et al., 2018, 2019; Bonnefon et al., 2016; Malle et al., 2015), the influence of non-human advisors on human decision-making in these specific situations has received very limited attention (Hanson et al., 2024; Straßmann et al., 2020).

1.4. The present study

The aim of the present study was to investigate the influence of moral arguments provided by AI-advisors, acting as decision aid tool, on human moral decisions and the associated neural correlates. Participants had to decide on sacrificial moral dilemmas (Patil et al., 2021) that were either categorized as *utilitarian dilemmas* (e.g., the trolley dilemma; Foot, 1978; Thompson, 1985) or *deontological dilemmas* (e.g., the footbridge dilemma; Thomson, 1976), depending on the general preference for the utilitarian option or the deontological option (i.e., rating study). In the baseline run, participants made their decisions by themselves, while in the AI-advised run, AI-advisors had already pre-selected one of the two options and provided participants with an argument to justify these pre-decisions. In the AI-advised run, participants also had the possibility to delegate the execution of the decisions to the AI-advisors by not answering for more than 15s. In this run, participants' prefrontal cortex activity was assessed

using an *f*NIRS system (Balconi & Fronda, 2020; Dashtestani et al., 2018, 2019; Lee & Yun, 2017; Strait & Scheutz, 2014).

Behavioral and brain activity data were first analyzed in terms of change of mind rate (compared to the baseline), as a function of the argument valence (i.e., validating versus contradicting the decision made in the baseline run). The data were also further analyzed as a function of the type of the dilemma (i.e., utilitarian/deontological) and the type of argument (i.e., utilitarian/deontological). Overall, we predicted to observe a greater proportion of utilitarian decisions in response to utilitarian dilemmas, and a greater proportion of deontological decisions in response to deontological dilemmas. Regarding brain activity, we predicted to observe a greater activity of the DLPFC in response to deontological dilemmas than to utilitarian dilemmas (Greene et al., 2001, 2004), reflecting the greater cognitive effort made to control the initial negative affective response to deontological dilemmas (Jeurissen et al., 2014; Lee & Yun, 2017). Regarding the influence of AI-advisors, we predicted to observe 1) a greater change of mind rate in response to contradicting arguments than to validating arguments and 2) a significant increase in utilitarian decisions in response to utilitarian arguments and/or a significant decrease in utilitarian decisions in response to deontological arguments in the AI-advised run (compared to the baseline run), with a potential interaction effect with the type of dilemma. We predicted that the influence of the AI-advisors would ensue from participants' motivation to engage in a greater cognitive effort to process contradicting arguments, which might be reflected by the greater activity of the DLPFC in response to contradicting arguments than to validating arguments (Jeurissen et al., 2014; Lee & Yun, 2017). We did not have specific predictions regarding participants' decisions to use AI advisors as delegates (i.e., to outsource the execution of their moral decisions) due to conflicting findings in previous studies. Some studies have found that participants retained control of action execution (Bigman & Gray, 2018; Castelo et al., 2019; Dietvorst et al., 2015, 2018; Gogoll &

Uhl, 2018), while others found that they can cede their control authority to non-human agents (Gombolay et al., 2015; Leyer & Schneider, 2019; Robinette et al., 2016).

2. METHODS

2.1. Participants

A review of the literature for sample size estimation – performed with the Pingouin package (v.0.5.3) in Python (v3.11) – revealed the necessity to record at least 8 participants to observe both the effect of the dilemma type on *f*NIRS (Lee & Yun, 2017) and the effect of advice type on performances (Goodyear et al., 2016) with a power ($1 - \beta$) = 95% and a significance level α = 5%. In addition, the sample size estimation for the effect of the interaction between dilemma type and argument type could not be computed since this has never been done with *f*NIRS to our knowledge. The closest result can be obtained by computing the interaction between dilemma type and response type on RMI data (Dashtestani et al., 2018) which resulted in a sample size of ~32 participants required to observe this effect with a power of 95% and a significance of 5%. 40 French participants (14 females, *Mage* = 28 years old) participated in the study. All were right-handed and had normal or corrected-to-normal vision. None of the participants reported a history of prior neurological disorder. They received no compensation for their participation.

2.2. Ethical concerns

All participants were informed of their rights and gave written informed consent for participation in the study. This study was carried out in accordance with both the Declaration of Helsinki and the recommendations of the Research Ethics Committee of the University of XXX in XXX (XXX no. 2017-045).

2.3. Materials

2.3.1. Selection of the dilemmas and the arguments

Two rating studies were conducted to select and balance the experimental materials. A first rating study was conducted to select an equal number of utilitarian dilemmas and deontological dilemmas, the two groups of dilemmas being balanced in terms of choice ratio (e.g., a utilitarian dilemma with a utilitarian choice rate of 80% was balanced with a deontological dilemma with a deontological choice rate of 80%). A second rating study was conducted to select one utilitarian argument and one deontological argument per dilemma (i.e., respectively arguing that *"it is necessary to sacrifice the few to save the many"* versus that *"unrelated people should not be sacrificed to save others"*) – the two classes of arguments being balanced in terms of persuasiveness. The rating studies are described in detail in the supplementary materials. The 18 dilemmas and the 36 arguments that were used in the present study are presented in Table 1S.

2.3.2. Experimental apparatus

The experiment was designed to record a hemodynamic neuronal response using *f*NIRS. The *f*NIRS signals were acquired by the NIRScout device manufactured by the NIRx Company (Germany). The NIRScout system has a 7.8125 Hz resolution and contains eight sources and seven detectors placed on the subject's scalp as shown in Figure 1. Eight optodes were placed close to the sources (i.e., less than 1 cm) to create eight short-channels. Each "source–detector" pair was placed close enough to each other (~ 3 cm) to form a *f*NIRS channel, generating 16 channels. Three regions of interest (ROI) were defined: 1) the rostral-lateral prefrontal area (RLPFC) composed by the S4-D3, S4-D4, S3-D3, S5-D4, S5-D5 and S6-D5 source-detector pairs, 2) the right dorsolateral prefrontal cortex (right-DLPFC) composed by S1-D1, S2-D1, S2-D2, S3-D2 and S3-D1 source-detector pairs and 3) the left dorsolateral prefrontal cortex (left-DLPFC) composed by S6-D6, S7-D6, S7-D7, S6-D7 and S8-D6 source-detector pairs (see Figure 1).

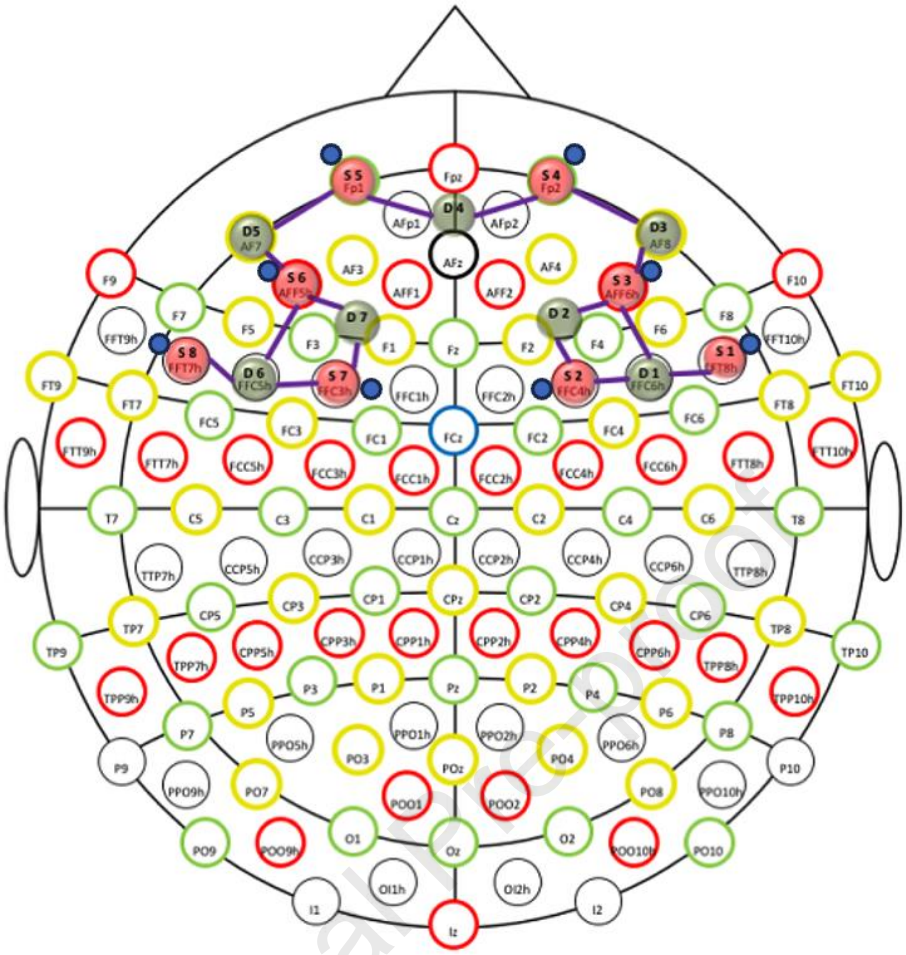


Figure 1. Illustration of the *f*NIRS montage in international 10-10 coordinate space. Montage with 8 × 9 frontal source-detector pairs. Sources are colored in red, detectors are colored in green, optodes for short-channels are colored in blue, and channels are indicated by purple lines.

2.4. Task

The experiment was divided into two runs. In the baseline run, participants were presented with 18 sacrificial moral dilemmas (i.e., nine utilitarian and nine deontological), each time with two possible options: one utilitarian and one deontological. Participants read the dilemma and then pressed the spacebar to make it disappear. After a 1-second blank screen, the two options were displayed side by side. Each type of option appeared half of the time on the left side of the screen and half of the time on the right side. Participants had to choose one of the two options by pressing a key on an AZERTY keyboard: the "*a*" key for the option on the left or the "*p*" key for the option on the right (see Figure 2).

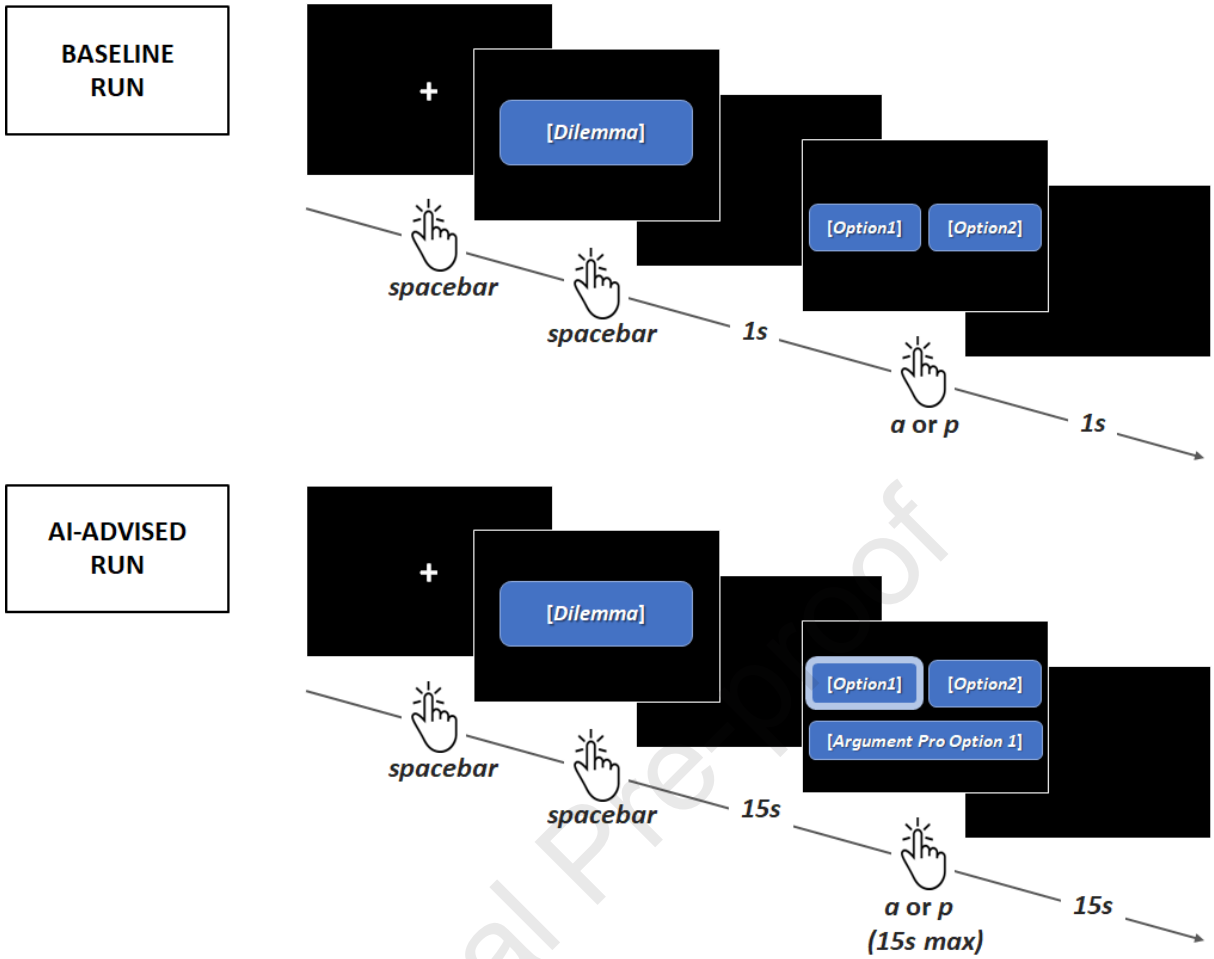


Figure 2. Illustrations of a trial in 1) the baseline run (top) and 2) the AI-advised run (bottom).

In the AI-advised run, participants were presented with the same sacrificial dilemmas as in the baseline run, but with different AI-advisors serving as moral advisors (i.e., a decision-aid tool based on artificial intelligence). Each dilemma was presented twice: once with an AI-advisor that had preselected the utilitarian option and provided a utilitarian argument and once with an AI-advisor that had preselected the deontological option and provided a deontological argument. As in the baseline run, participants read the dilemma and then pressed the spacebar to make it disappear. After a 15s blank screen, participants were presented with the two options (the option preselected by the AI-advisor was circled in light blue) and the argument justifying the pre-decision of the artificial advisor. In this run too, participants had to choose one of the two options by pressing a key on an AZERTY keyboard: the "*a*" key for the option on the left

or the "*p*" key for the option on the right. They also had the possibility to remain inactive for 15s and let the AI-advisor execute its pre-decision (i.e., delegation). Participants were given no information regarding the characteristics of the AI-advisors (e.g., reliability, design).

2.5. Procedure

After signing the informed consent form, participants were asked to enter the experimental room and read the instructions for the experiment. They were informed that in the first phase of the experiment they would be presented with different sacrificial dilemmas and their task was to choose between two options. There was no time limit for them to give their answer. It was emphasized that there were no right or wrong answers, and they should make their decisions based on their personal preferences. Participants were invited to ask any questions before the experiment began. They then sat in a comfortable seat in front of the 1920 x 1080 computer screen and performed the baseline run. At the end of the baseline run, the NIRScout system was placed on their heads. They were equipped with the NIRScap, to which 23 optodes were attached. The NIRX system was then calibrated. Pre-tests prompted us to record participants' brain activity only during the AI-advised run to prevent them from experiencing headaches triggered by the *f*NIRS system. Concurrently to the installation of the *f*NIRS system, participants were given the written instructions for the second phase of the experiment (i.e., AI-advised run). They were informed that in the second phase of the experiment, their task would be the same as in the first phase, but this time they would receive advice from different AI-advisors. It was explained that the arguments presented had been pre-generated by various AIs, each trained on distinct sets of human moral decision-making arguments. Participants were asked to imagine that they were making decisions with the assistance of these AI-advisors, which could automatically analyze each situation and provide guidance. Once ready, participants performed the AI-advised run. At the end of the latter, the experimenter removed the NIRScout system from the participants' head. Participants were then

asked to fill the French version of the Interpersonal Reactivity Index (IRI; composed of a empathic concern, a perspective taking, a personal distress and a fantasy subscale; Davis, 1980; Gilet et al., 2013), the Gudjonsson's compliance scale that we translated in French (Gudjonsson, 1989) and the French version of the Negative Attitudes towards Robots Scale (NARS; composed of an interaction, a social and an emotion subscale; Dinet & Vivian, 2015; Nomura et al., 2006). When filling the NARS and the Gudjonsson's compliance scale, participants had to indicate their level of agreement with the statements on respectively a 9-point Likert scale (from – 4 = *totally disagree* to 4 = *totally agree*) and a 7-point Likert scale (from 1 = *totally disagree* to 7 = *totally agree*). When filling the IRI, participants had to indicate the extent to which the statements described them on a 7-point Likert scale (from 1 = *does not describe me at all* to 7 = *describes me perfectly*). After completing the questionnaires, participants were debriefed. They were asked about their general impressions of the experiment and their thoughts on the possibility of delegating the execution of actions to AI-advisors. They were invited to ask any questions they had. Finally, they were thanked for their participation and left the experimental room.

2.6. Data measures

2.6.1. Behavioral measures

Decisions. Participants' decisions were assessed in terms of utilitarian decisions. Delegations (see below) were counted as utilitarian decisions when the AI-advisors had preselected the utilitarian option.

Changes of mind. To further analyze participants' decisions, the decisions made in the AI-advised run were compared to those made in the baseline run (for each dilemma), a measure referred to as changes of mind in the present study. Changes of mind were assessed in terms of 1) changes of mind from the deontological option to the utilitarian option, 2) changes of mind

from the utilitarian option to the deontological option and 3) indiscriminate changes of mind. Delegations were taken into account based on the option preselected by the AI-advisors.

Delegations versus active decisions. A delegation was defined as an absence of participants' response for more than 15s that triggered the execution of the AI-advisors' pre-decision. They were assessed in terms of active decisions.

Response times. In the baseline run, response times were defined as the time period between the onset of the options and the beginning of the participants' responses. In the AI-advised run, they were defined as the time period between the onset of the options and the AI-advisors' arguments and the beginning of the participants' responses. The response times above or below two standard deviations around the mean and those associated with delegations were eliminated (Ratcliff, 1993). Response times were assessed in terms of mean response times.

2.6.2. *f*NIRS measures

*f*NIRS data was preprocessed by using MNE-Python (Gramfort et al., 2013), its extension MNE-NIRS (Luke et al., 2021) and following the guidelines from Yücel and colleagues (2021). First, raw data were converted into optical density data. Then, a channel pruning was applied using the scalp-coupling index (SCI) for each channel. SCI is an indicator of the quality of the connection between the optodes and the scalp, which looks for the presence of a prominent synchronous signal in the frequency range of cardiac signals across the photo-detected signals (Pollonini et al., 2014). Channels with a scalp-coupling index below 0.75 were marked as bad channels (less than 10%) and were interpolated with the nearest channel providing good data quality. Visual checking of data was also performed. In order to remove baseline shift and spike artifacts, temporal derivative distribution repair was applied (Fishburn et al., 2019). To remove systemic signals contaminating the brain activity, short-separation regression was used: short-channel data was subtracted from the standard long-channel data (Zhou et al., 2020). Beer-Lambert Law was applied to transform optical density into oxygenated

(HbO) and deoxygenated (HbR) hemoglobin concentration changes with a partial pathlength factor of six. Data were filtered using a third-order zero-phase Butterworth bandpass filter with cutoff frequencies of 0.01 to 0.5 Hz to remove instrumental and physiological noise.

2.7. Data analysis

The data of 9 participants were set aside either because they were incomplete (i.e., participants had not entirely filled the questionnaires or the *f*NIRS signal was not recorded due to a technical problem) or of insufficient quality (i.e., the *f*NIRS was too noisy to be analyzed). The data were analyzed as a function of the type of dilemma and the type of argument. The data were also analyzed based on the valence of the arguments (i.e., whether the arguments provided by the AI-advisors contradicted or validated the choice made by participants in the baseline run). In this second type of analysis, the type of dilemma factor was not included, due to the variable proportion of validating and contradicting arguments among each class of dilemmas.

2.7.1. Behavioral data

Due to the binary nature of decisions, the non-normality of the data and the repeated-measures design of the study, we chose to use Generalized Estimating Equation (GEE) models to analyze participants' decisions, change of minds and delegations. Response times were log-transformed due to their non-normal distribution and analyzed using ANOVA or dependent t-tests.

Decisions. A 2 x 3 [Dilemma Type (utilitarian, deontological) x Argument Type (no argument, utilitarian argument, deontological argument)] binary logistic regression was conducted on participants' decisions (with 0 = deontological choice and 1 = utilitarian choice). A manual stepwise analysis was performed to remove non-significant interactions from the model and pairwise comparisons were carried out to further examine significant effects ($\alpha < .05$).

Changes of mind. Four analyses were conducted to investigate participants' changes of minds between the baseline run and the AI-advised run. A 2 x 2 [Dilemma Type (utilitarian, deontological) x Argument Type (utilitarian argument, deontological argument)] binary logistic regression was conducted on participants' decisions to change their mind (with 0 = no change of mind and 1 = change of mind), without discriminating the nature of the change of mind (i.e., from a deontological choice to a utilitarian one and the opposite). Two supplementary 2 x 2 [Dilemma Type (utilitarian, deontological) x Argument Type (utilitarian argument, deontological argument)] binary logistic regressions were conducted on participants' decisions to change their mind 1) from the utilitarian option to the deontological option and 2) from the deontological option to utilitarian option in response to the AI-advisors' arguments (with 0 = no change of mind and 1 = change of mind). Finally, a 2 [Argument Valence (validating, contradicting)] binary logistic regression was conducted on participants' decisions to change their mind (with 0 = no change of mind and 1 = change of mind).

Delegations versus active decisions. A 2 x 2 [Dilemma Type (utilitarian, deontological) x Argument Type (utilitarian, deontological)] and a 2 [Argument Valence (validating, contradicting)] binary logistic regressions were performed on active decisions (with 0 = delegation and 1 = active decision). Manual stepwise analyses were performed to remove non-significant interactions from all the models and pairwise comparisons were carried out to further examine significant effects ($\alpha < .05$).

Response times. Because participants made their decision after reading the dilemmas in the baseline run versus after reading the advice of AI-advisors in the AI-advised run, the response times of each run were analyzed separately. A 2 [Dilemma Type (utilitarian, deontological)] dependent t-test was performed on the log-transformed response times when no arguments were provided to the participants (i.e., baseline run). Two analyses were performed on the response times measured in the AI-advised run. A 2 x 2 [Dilemma Type (utilitarian, deontological) x

Argument Type (utilitarian argument, deontological argument)] ANOVA and a 2 [Argument Valence (contradicting, validating)] paired t-test were performed on the log-transformed response times measured in the AI-advised run. Tukey post-hoc tests were carried out to further examine significant effects ($\alpha < .05$) of the ANOVA analysis.

2.7.2. Subjective data

To investigate whether participants' empathy level predicted participants' utilitarianism rate, we conducted a multiple regression analysis with participants' utilitarianism rate as dependent variable and the IRI subscales ratings (i.e., perspective taking, empathic concern and personal distress subscales) as predictors. A second and a third multiple regression analysis was conducted to investigate whether participants' attitudes toward AI, empathy level and tendency to comply predicted participants' propensity to 1) change their mind and to 2) delegate the execution of the decision, when the arguments of AI-advisors contradicted versus validated participants decisions in the baseline run, with participants' change of mind rate as dependent variable and the ratings of IRI subscales, the NARS subscales (i.e., interaction, social and emotion) and the Gudjonsson's compliance scale as predictors. Manual stepwise analyses were performed to remove the effects with a $p > .15$ from the models (Field, 2013).

2.7.3. *f*NIRS data

The generalized linear model (GLM) approach was used to quantify the amplitude of evoked hemodynamic responses per ROI and Condition. The GLM was fit to the long-channel data and included all principal components of short-detector channels to account for extracerebral and physiological signal components (Gagnon et al., 2014). The design matrix for the GLM was generated by convolving a boxcar function at each event-onset time with the SPM canonical hemodynamic response function (Abraham et al., 2014). The GLM was performed with a lag-1 autoregressive noise model, to account for the correlated nature of the *f*NIRS signal components. Individual beta estimates were then averaged for each ROI, weighted by the

standard error. As the oxyHb response is considered as a more sensitive indicator of changes in regional blood flow (Yücel, 2021), we chose to focus on the variations in oxyHb level. The use of ROIs as a factor may lead to unintended statistical bias due to the optical properties that may differ systematically between ROIs (Herold et al., 2018).

Dilemma. For the following analyses, the duration of boxcar function was 15 seconds and drift orders accounting for signal components up to 0.033 Hz were included as regression factors. To compare the hemodynamic responses observed for utilitarian dilemmas and deontological dilemmas, three 2 [Argument Valence (contradicting, validating)] paired t-test were performed on β values extracted from GLM separately for each ROI. To test whether the hemodynamic responses observed for utilitarian dilemmas and deontological dilemmas varied between the second and the third presentation, three 2 x 2 [Dilemma Type (utilitarian, deontological) x Presentation Order (second time, third time)] ANOVA were performed on β values extracted from GLM separately for each ROI. Tukey post-hoc tests were carried out to further examine significant effects ($\alpha < .05$) of the ANOVA analysis.

Dilemma & Argument. Due to the important variability of the measurements performed in the literature (e.g., Balconi & Fronda, 2020; Dashtestani et al., 2018, 2019; Greene et al., 2004; Lee & Yun, 2017; Strait & Scheutz, 2014), the hemodynamic responses observed as a function of both dilemma and argument types were assessed according to three temporal periods: 1) time-locked to the presentation of the argument, 2) time-locked 15 seconds before the response and 3) in the [-8s ; 8s] interval around the response. The duration of the boxcar function depended on the temporal periods: 30 seconds for the first, 15 seconds for the second and 16 seconds. Drift orders included as regression factors were also designed based on the temporal periods, with drift orders accounting for signal components up to 0.016 Hz for the first, 0.033 Hz for the second and 0.031 Hz for the third temporal period. For each of these three measures, three 2 x 2 [Dilemma Type (utilitarian, deontological) x Argument Type (utilitarian argument,

deontological argument)] ANOVAs were performed on β values extracted from GLM, on each of the three ROIs (i.e., nine analyses in total). Tukey post-hoc tests were carried out to further examine significant effects ($\alpha < .05$) of the ANOVA analyses.

Argument valence. The hemodynamic responses observed as a function of the argument valence were also assessed according to three temporal periods: 1) time-locked to the presentation of the argument, 2) time-locked 15 seconds before the response and 3) in the [- 8s; 8s] interval around the response. The duration of the boxcar function depended on the temporal periods: 30 seconds for the first, 15 seconds for the second and 16 seconds. Drift orders included as regression factors were also designed based on the temporal periods: drift orders accounting for signal components up to 0.033 Hz for the first, 0.063 Hz for the second and 0.031 Hz for the third). For each of these three measures, three 2 [Argument Valence (contradicting, validating)] paired t-tests were performed on β values extracted from GLM, on each of the three ROIs (i.e., nine analyses in total).

3. RESULTS

3.1. Behavioral results

3.1.1. Decisions

The 2 x 3 [Dilemma Type (utilitarian, deontological) x Argument Type (no argument, utilitarian argument, deontological argument)] binary logistic regression conducted on decisions revealed a significant main effect of dilemma type [*B* (SE) = - 1.502 (.186), CI $_{(95\%)}$ = (- 1.867, - 1.137), Wald χ^2 (1) = 65.203, $p < .001$; see Figure 3.A.], with utilitarian dilemmas predicting for a high rate of utilitarian decisions (*M* = 69.41 %, *SD* = 26.85) than deontological dilemmas (*M* = 33.81 %, *SD* = 24.35). The analysis also revealed a main effect of the argument type [utilitarian argument: *B* (SE) = - .094 (.109), CI $_{(95\%)}$ = (- .307 , .120), Wald χ^2 (1) = .737, $p = .391$; deontologist argument*: B* (SE) = .286 (.129), CI $_{(95\%)}$ = (.034, .538), Wald χ^2 (1) = 4.930, $p < .05$; with no argument as dummy], with significantly lower utilitarian decision rate when AI-

advisors provided a deontological argument (M = 46.59 %, SD = 32.80) than a utilitarian argument (M = 55.02 %, SD = 31.00, $p < .01$) and when no argument was provided (M = 53.23 %, SD = 29.51, $p < .05$; see Figure 3.B.). No difference in utilitarian decision rate was found when no argument was provided and when a utilitarian argument was provided by the AI-advisor (p = .389).

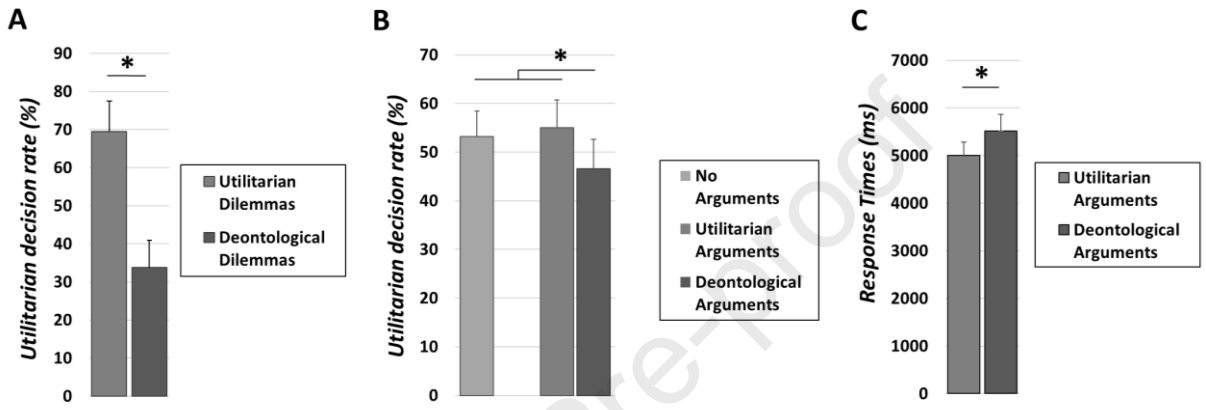


Figure 3. Illustration of A) the utilitarian decision rate as a function of the dilemmas type (with utilitarian dilemmas in mid-grey and deontological dilemmas in dark gray), B) the argument type (with no argument in light grey, utilitarian arguments in mid-grey and deontological arguments in dark grey); and C) response times as a function of the argument type in the AI-advised run.

3.1.2. Changes of mind between the baseline run and the AI-advised run

Changes of mind as a function of the argument type. The analysis revealed that no difference in change of mind rate was found as a function of the dilemma type [B (SE) = .221 (.310), CI $_{(95\%)}$ = (- .386, .828) , Wald χ^2 (1) = .508, p = .476] and the argument type [B (SE) = .060 (.131), CI $_{(95\%)}$ = (- .196, .316), Wald χ^2 (1) = .211, p = .646].

Changes of mind from the utilitarian to the deontological option. The analysis revealed a significant main effect of argument type [B (SE) = .808 (.256), CI $_{(95\%)}$ = (.307, 1.309), Wald χ^2 (1) = 9.987, $p < .01$; see Figure 4.A], with deontological arguments (M = 8.42 %, SD = 11.82) predicting for a greater amount of changes of mind from a utilitarian to a deontological decision

than utilitarian arguments (*M* = 3.94 %, *SD* = 5.59). The main effect of dilemma type [*B* (SE) = .282 (.293), CI (95%) = (- .292, .856), Wald χ^2 (1) = .926, *p* = .336] did not reach significance.

Changes of mind from the deontological to the utilitarian option. The analysis revealed a significant main effect of argument type [*B* (SE) = - 1.204 (.344), CI (95%) = (- 1.877, -.531), Wald χ^2 (1) = 12.291, *p* < .001; see Figure 4.A], with utilitarian arguments predicting for a greater amount of changes of mind from the deontological to the utilitarian option (*M* = 5.73 %, *SD* = 7.38) than deontological arguments (*M* = 1.79 %, *SD* = 3.33). The main effect of dilemma type [*B* (SE) = .100 (.436), CI (95%) = (- .755, .955), Wald χ^2 (1) = .053, *p* = .818] did not reach significance.

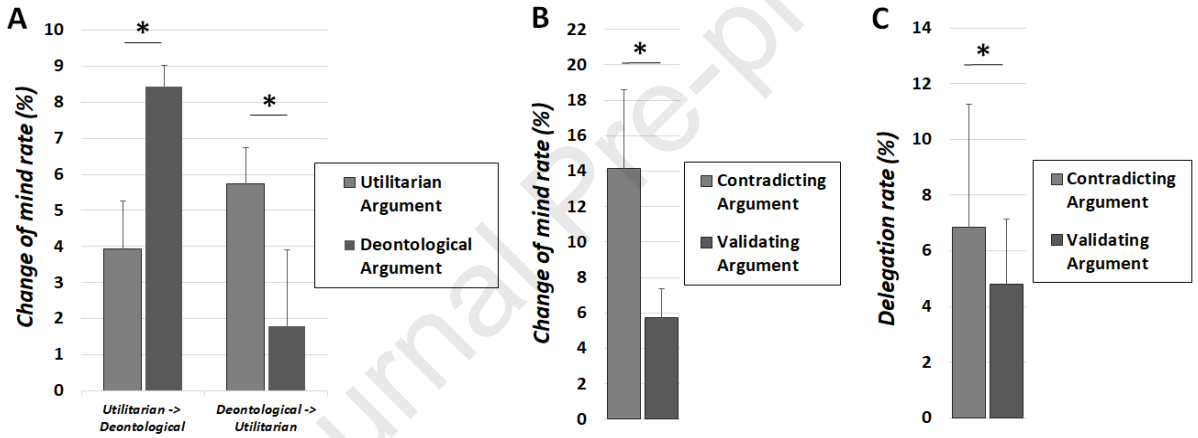


Figure 4. Illustration of A) the change of mind rate as a function of the argument type from deontological option to utilitarian option (left) and from utilitarian option to deontological option, B) the change of mind rates and C) the delegation rates as a function of the valence of the arguments.

Changes of mind as a function of the argument valence. In the AI-advised run, 17 participants showed a greater change of mind rate in response to the contradicting arguments than the validating arguments, 10 participants showed the same amount of change of minds in response to the validating and the contradicting arguments, and 5 participants never changed their minds at all. The analysis revealed a main effect of argument valence [*B* (SE) = - .967 (251), CI (95%) = (- 1.469, - .483), Wald χ^2 (1) = 15.080, *p* < .001; see Figure 4.B], with contradicting arguments predicting for a greater change of mind rate (*M* = 14.16 %, *SD* = 19.93) than validating arguments (*M* = 5.73 %, *SD* = 9.16).

3.1.3. Delegations versus active decisions in the AI-advised run.

In the AI-advised run, 14 participants never delegated the execution of the decision to the AI-advisors, while 17 participants made at least one delegation. The results revealed a rather low mean delegation rate (M = 5.83 %, SD = 14.13).

Active decisions as a function of the argument type. The analysis performed on active decisions revealed no significant main effects of dilemma type effect [B (SE) = - .223 (.230), CI $_{(95\%)}$ = (- 675, .229), Wald χ^2 (1) = .936, p = .333] or argument type [B (SE) = .032 (.183), CI $_{(95\%)}$ = (- .327, .390), Wald χ^2 (1) = .030, p = .862].

Active decisions as a function of the valence of the argument. The analysis performed on active decisions revealed a significant main effect of argument valence [B (SE) = .417 (.186), CI $_{(95\%)}$ = (.053, .781), Wald χ^2 (1) = 5.029, p < .05], with significantly fewer active decisions when the AI-advisors provided contradicting arguments (M = 93.15 %, SD = 15.29) than validating arguments (M = 95.19 %, SD = 13.36; see Figure 4.C).

3.1.4. Response times

Response times in the baseline run. No difference in response times was found in response to utilitarian and deontological dilemmas [t (31) = - 1.260, p = .217, $CI_{95\%}$ = (- .135; .032)] when no arguments were provided by the AI-advisors (i.e., baseline run).

Response times as a function of the argument type. The analysis revealed a main effect of arguments type [F (1, 30) = 8.708, p < .01, ηp^2 = .225], with longer response times when AI-advisors provided deontological arguments (M = 5515 ms, SD = 1477) than utilitarian arguments (M = 5008 ms, SD = 1291; see Figure 3.C). The main effect of dilemma type [F (1, 30) = .356, p = .555, ηp^2 = .012] and Dilemma Type x Argument Type interaction [F (1, 30) = .635, p = .432, ηp^2 = .021] did not reach significance.

Response times as a function of the argument valence. No difference in response times was found as a function of the argument valence [t (31) = - 0.223, p = .825, $CI_{95\%}$ = (- .021; .017)].

3.1.5. Subjective & behavioral data

Participants' level of empathy was not predictive of their utilitarian decision rate in the first phase of the experiment [F (1, 29) = .291, p = .594, R^2 = .010, $R^2_{adjusted}$ = - .024; see Table 1 for a summary of the model]. However, in the second phase of the experiment participants' tendency to comply and tendency to experience positive emotions toward robots/A.I. significantly were predictive of participants' propensity both to change their mind [F (3, 27) = 7.790, p < .001, R^2 = .464, $R^2_{adjusted}$ = .404] and to delegate the execution of the action [F (3, 27) = 4.812, p < .01, R^2 = .348, $R^2_{adjusted}$ = .276] when being contradicted (versus validating) by the AI-advisors (see Table 1 for a summary of the model).

Insert Table 1 here

3.2. *f*NIRS results

The 2 x 2 [Dilemma Type (utilitarian, deontological) x Argument Type (utilitarian argument, deontological argument)] ANOVA performed on the hemodynamic responses time-locked to the presentation of the arguments and measured at the right ROI was the only analysis that reached significance. For the sake of clarity, non-significant analyses are reported in Table 2.

Insert Table 2 here

Hemodynamic response time-locked to the argument. The ANOVA analysis conducted on the right ROI revealed a significant main effect of argument type [F(1, 30) = 8.109, p < .01, np^2 = .213, see Figure 5], with a higher consumption of oxygenated hemoglobin observed in response to utilitarian arguments (M = .019, SD = .010) than to deontological arguments (M = - .013, SD = .013). The main effect of dilemma type [F(1, 30) = .006, p = .941, np^2 = .000] and the

Dilemma Type x Argument Type interaction [$F(1, 30) = 1.749$, $p = .196$, $np^2 = .055$] did not reach significance.

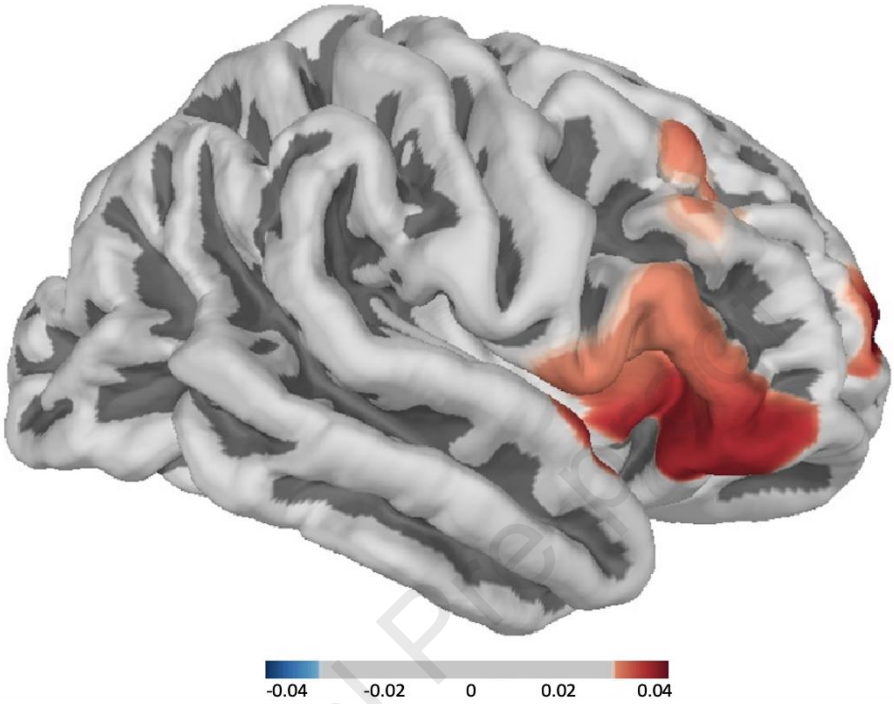


Figure 5. Topographic plot of the contrast between the responses to utilitarian arguments > deontological arguments.

4. DISCUSSION

The present experiment aimed at investigating the influence of AI-advisors on human moral decision-making and the associated brain correlates. Participants were presented with sacrificial dilemmas and were tasked with making moral decisions either independently themselves (i.e., baseline run) or with AI-advisors (i.e., AI-advised run). Their brain activity was assessed during the AI-advised run using an *f*NIRS system.

4.1. Deciding on utilitarian and deontological dilemmas

In both the baseline and the AI-advised runs, participants were significantly more likely to choose the utilitarian option in response to utilitarian dilemmas (i.e., ~ 70%) and the deontological option in response to deontological dilemmas (i.e., ~ 67%). This result serves as

a manipulation check confirming that the two types of dilemmas were properly defined and balanced based on the first rating study. In the AI-advised run, no modulation of DLPFC activity was observed as a function of the type of dilemma (not in line with Greene et al., 2001, 2004). It is possible that this lack of difference in brain activity stems from the fact that all dilemmas had already been presented during the baseline run. Participants may have engaged in less thorough analysis of the dilemmas when they were presented for the second and third time during the AI-advised run, resulting in an absence of brain activity difference. In contradiction with some previous studies (Gleichgerrcht & Young, 2013; Nasello et al., 2021; Takamatsu, 2018), participants' self-assessed level of empathy was not predictive of their level of utilitarianism. This lack of association between utilitarianism and empathy in the current study is not unexpected. A recent meta-analysis demonstrated that the role of empathy in moral judgments can strongly vary, oscillating between a small/moderate to limited/insignificant link depending on the moral decisions (Nasello & Triffaux, 2023).

4.2. Change of mind in response to AI-advisors' arguments

Overall, participants were significantly more likely to change their mind 1) from the utilitarian to the deontological option in response to deontological arguments than to utilitarian ones and 2) from the deontological to the utilitarian option in response to utilitarian arguments than to deontological ones. Moreover, participants changed significantly more their mind when the arguments of AI-advisors contradicted than validated the decisions they made in the baseline run. Taken together, these results demonstrate that AI-advisors can have a significant influence on moral decision-making (in line with Bai et al, 2023; Hertz & Weise; 2018; Köbis et al., 2021; Leib et al., 2024). While 17 participants were more likely to change their minds in response to contradicting arguments than to validating ones, 10 participants showed an equal change of mind rate in response to both types of arguments, and 5 participants never changed

their minds at all. Therefore, despite the statistical significance of these findings, it is important to note that the AI advisors' arguments did not affect all participants uniformly.

Interestingly, participants' propensity to change their mind when contradicted by AI-advisors was predicted both by their self-estimated tendency to comply (i.e., Gudjonsson's compliance scale) and their disposition to experience negative emotions toward non-human agents (i.e., inverted emotion NARS subscale). In other words, the influence of AI-advisors appears to depend on human advisees' susceptibility to obey and to fear non-human agents, as previously found with human advisors (e.g., Fabre et al., 2022).

4.3. Impact of the utilitarian versus deontological arguments of AI-advisors

Being advised to make deontological choices by the AI-advisors seems to have prompted participants to thoroughly reassess their decisions, as indicated by the longer response times observed in response to deontological arguments than to consequentialist ones. Interestingly, greater activity of right-DLPFC was also measured in response to utilitarian arguments than to deontological ones. In various previous studies, increased activity in the right-DLPFC was found to predict for (more) utilitarian decisions (e.g., Greene et al., 2004; Jeurissen et al., 2014; Lee & Yun, 2017; Patil et al., 2021; Zheng et al., 2018). The right-DLPFC is thought to underpin the cognitive control processes necessary to appraise the emotional response to moral dilemmas (Greene et al., 2004; Tassy et al., 2012). Being provided with deontological arguments by the AI-advisors appears to have decreased the appraisal of the affective response to the sacrificial moral dilemmas, ultimately resulting in a significant decrease in participants' utilitarianism.

Human beings are acutely aware that the moral decisions they make shape both their self-image and the way they are perceived by others (Bauman & Helzer, 2023; Brambilla et al., 2021; Carlson & Furr, 2009; Carlson et al., 2011; Macko, 2020; Plaks et al., 2021; Reynolds et al., 2019; Rom & Conway, 2018; Uhlmann et al., 2015). As a result, they tend to adjust both

their moral choices and their advice to align with the way they wish to present themselves (Jin & Peng, 2021; Macko, 2020; Polman & Ruttan, 2022; Rom & Conway, 2018; Trémolière & Rateau, 2023). Compared to those who behave in a deontological way, individuals who make utilitarian decisions are generally viewed as more competent, but less warm, trustworthy, empathetic, moral, prosocial, and attractive (Brown & Sacco, 2019; Capraro et al., 2018; Everett et al., 2018; Rom et al., 2017; Rom & Conway, 2018; Sacco et al., 2017; Uhlmann et al., 2013). Moral character being far more influential than competence when forming impressions (Brambilla et al., 2011, 2012; Goodwin et al., 2015; Leach et al., 2007), human beings tend to adapt their moral preferences to appear more deontological than they actually are (e.g., Bostyn & Roets, 2017; Lee et al., 2018; Reynolds et al., 2019; Sacco et al., 2017), as a strategical move to avoid the personal and social costs associated with utilitarianism (Brown & Sacco, 2019; Everett et al., 2018; Goldstein-Greenwood et al., 2020; Rom et al., 2017; Szekely & Miu, 2014; Uhlmann et al., 2013). Compared to non-human agents, human beings are less strongly expected by their peers to make utilitarian choices and are typically judged more negatively when they do, particularly in deontological dilemma situations (e.g., Chu & Liu, 2023; Malle et al., 2015; Zhang et al., 2022). Therefore, making a utilitarian decision after receiving a deontological argument from an AI-advisor exposed participants to both the personal and/or social costs of acting in a utilitarian manner (e.g., Reynolds et al., 2019; Sacco et al., 2017) and those of seeing themselves and/or being perceived by others as more cold-blooded than machines (Chu & Liu, 2023). We hypothesize that the reevaluation process that occurred when the AI-advisors provided deontological arguments reflects participants' motivation to preserve their (self-)image (Bostyn & Roets, 2017; Macko, 2020; Reynolds et al., 2019; Rom & Conway, 2018).

The present results contradict Greene's *moral dual-process theory*, which posits that deontological decisions mainly result from a fast, automatic and emotional decision process

underpinned by System 1 (Greene, 2015). However, they align with Moll and colleagues' view, asserting that cognition and emotion are continuously integrated during moral decision-making and that motivational power – here supposedly preserving one's (self-)image – can lead to a voluntarily down- or up-regulation of affective reactions (Moll et al., 2005, 2008; Ochsner et al., 2004).

4.4. Delegating the execution of the action to the AI-advisors

In the AI-advised run, participants were given the possibility to delegate the execution of the decision to the AI-advisors by not replying for 15 seconds. On average, participants chose to delegate about 6% of the time. The results revealed significant variability among participants in their willingness to delegate the decision execution to the AI-advisors, with 14 participants never opting to delegate, and 17 participants doing it at least once. During the post-experimental debriefing, all 14 participants who never chose to delegate reported feeling uncomfortable with the notion of allowing a machine to execute such actions, expressing a preference for maintaining control (in line with Bigman & Gray, 2018; Castelo et al., 2019; Dietvorst et al., 2015, 2018; Gogoll & Uhl, 2018). Conversely, the 17 participants who delegated the execution of actions at least once reported that certain decisions were so challenging to execute that they felt relieved to be able to delegate them to the AI-advisors (in line with Drugov et al., 2014; De Melo et al., 2016; Sloane & Moss, 2019). A significant increase in delegation (i.e., decrease in active decisions) was observed when AI-advisors provided arguments contradicting the decisions made by participants in the baseline run. Moreover, the propensity of participants to delegate the execution to AI-advisors after being contradicted versus validated was predicted both by their self-estimated propensity to comply and tendency to experience negative emotions toward non-human agents. Based both on literature and participants' feedback, this result suggests that participants' decisions to delegate the execution of the actions to the AI-advisors might reflect a willingness to surrender decision-making authority to the machine to avoid the

negative emotional response associated with the execution of the action (Drugov et al., 2014; De Melo et al., 2016; Sloane & Moss, 2019).

4.5. Limitations and further work

In the present study, participants were tasked with deciding on fictitious sacrificial dilemmas that may not fully represent the moral decisions individuals typically face in everyday life. Consequently, further research is warranted to assess the influence of AI moral advisors on the moral decisions humans regularly encounter in their personal and professional lives. This study is one of the first to investigate the brain activity associated with moral decision-making advised by AI-agents (Goodyear et al., 2016; 2017). As previous *f*NIRS studies investigating moral decision-making (Balconi & Fronda, 2020; Dashtestani et al., 2018, 2019; Lee & Yun, 2017; Strait & Scheutz, 2014), we focused on the activity of the PFC, a key structure involved in moral decision-making. However, the activity of various other brain structures underpinning advice taking (e.g., Goodyear et al., 2016, 2017; Meshi et al., 2012) and/or moral judgment/decision-making (e.g., Casebeer, 2003; Greene et al, 2001, 2004; Moll et al., 2005; Moll & de Oliveira-Souza, 2007), such as for instance the anterior cingulate cortex, the ventromedial PFC or the temporal-parietal junction, was not investigated due to the inherent constraints associated with the use of *f*NIRS (i.e., limited number of channels and measurements restricted to the outer cortex; Quaresima & Ferrari, 2019). Replicating the present study using *f*MRI would further improve our understanding of the involvement of both subcortical and (outer and inner) cortical structures during AI-advised moral decision-making.

Finally, further work is necessary to confirm that participants lowered their level of utilitarianism in response to the deontological arguments provided by the AI-advisors to preserve their (self-)image and to determine the extent to which this behavior is driven by self-oriented versus other-oriented motives (e.g., Reynolds et al., 2019; Sacco et al., 2017).

Conclusion

The results of the present study demonstrate that AI-agents can influence moral decision-making in human beings. They confirm the necessity to define clear rules to regulate the way artificial moral advisors are trained and used in everyday life (e.g., Constantinescu et al., 2022; IEEE, 2017; Kobis et al., 2019; Russell et al., 2015). Some authors have proposed to adjust the "behavior" of artificial moral advisors to the moral preferences of the user (e.g., Giubilini & Savulescu, 2018), which appears as a good solution at least for a personal use of this technology. Given the influence of artificial moral advisors, particular attention should be given to their use in a professional context – especially in the military (Svenmarck et al., 2018) – in which the moral preferences of the users might not align with those of the organization they work for (IEEE, 2017). Further research is therefore necessary to define the rules for proper and ethical use of artificial moral advisors (Irving & Askell, 2019).

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, *8*, 14. https://doi.org/10.3389/fninf.2014.00014

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332-2337. https://doi.org/10.1073/pnas.1911517117

Bai, H., Voelkel, J. G., Eichstaedt, J. C., & Willer, R. (preprint). Artificial Intelligence Can Persuade Humans on Political Issues. *Osf*

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2023). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, *42*(28), 24516-24541. https://doi.org/10.1007/s12144-022-03573-2

Balconi, M., & Fronda, G. (2020). Morality and management: an oxymoron? *f*NIRS and neuromanagement perspective explain us why things are not like this. *Cognitive, Affective, & Behavioral Neuroscience*, *20*, 1336-1348. https://doi.org/10.3758/s13415-020-00841-1

Bauman, C. W., & Helzer, E. G. (2023). Interpersonal consequences of moral judgments about others. In N. Ellemers, S. Pagliaro, and F. van Nunspeet (Eds.), *The Routledge International Handbook of the Psychology of Morality*. Abingdon, UK: Routledge. https://doi.org/10.4324/9781003125969-17

Benitez, J., Wyman, A. B., Carpinella, C. M., & Stroessner, S. J. (2017, August). The authority of appearance: How robot features influence trait inferences and evaluative responses. *In 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 397-404). IEEE.

Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The neural basis of following advice. *PLoS biology*, *9*(6), e1001089. https://doi.org/10.1371/journal.pbio.1001089

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34.https://doi.org/10.1016/j.cognition.2018.08.003

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, *101*(2), 127-151. https://doi.org/10.1016/j.obhdp.2006.07.001

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

Bostyn, D. H., & Roets, A. (2017). An asymmetric moral conformity effect: Subjects conform to deontological but not consequentialist majorities. *Social Psychological and Personality Science*, *8*(3), 323-330. https://doi.org/10.1177/1948550616671999

Braams, B. R., Davidow, J. Y., & Somerville, L. H. (2019). Developmental patterns of change in the influence of safe and risky peer choices on risky decision-making. *Developmental science*, *22*(1), e12717. https://doi.org/10.1111/desc.12717

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, *41*, 135–143. https://doi.org/10.1002/ejsp.744

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest!: Moral traits dominate group impression formation. *British Journal of Social Psychology*, *51*, 149–166. https://doi.org/10.1111/j.2044-8309.2010.02011.x

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. *Advances in Experimental Social Psychology*, *64*, 187–262. https://doi.org/10.1016/bs.aesp.2021.03.001

Brown, M., & Sacco, D. F. (2019). Is pulling the lever sexy? Deontology as a downstream cue to long-term mate quality. *Journal of Social and Personal Relationships*, *36*(3), 957-976. https://doi.org/10.1177/0265407517749331

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology*, *20*(13), 1165-1170. https://doi.org/10.1016/j.cub.2010.04.055

Capraro, V., Sippel, J., Zhao, B., Hornischer, L., Savary, M., Terzopoulou, Z., ... & Griffioen, S. F. (2018). People making deontological judgments in the Trapdoor dilemma are perceived to be more prosocial in economic games than they actually are. *PLoS One*, *13*(10), e0205066. https://doi.org/10.1371/journal.pone.0225850

Carlson, E. N., & Furr, R. M. (2009). Evidence of Differential Meta-Accuracy: People Understand the Different Impressions They Make. *Psychological Science*, *20*(8), 1033-1039. https://doi.org/10.1111/j.1467-9280.2009.02409.x

Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them? *Journal of Personality and Social Psychology*, *101*(4), 831–846. https://doi.org/10.1037/a0024297

Casebeer, W. D. (2003). Moral cognition and its neural constituents. *Nature Reviews Neuroscience*, *4*(10), 840-846. https://doi.org/10.1038/nrn1223

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809-825. https://doi.org/10.1177/0022243719851788

Černevičienė, J., & Kabašinskas, A. (2022). Review of multi-criteria decision-making methods in finance using explainable artificial intelligence. *Frontiers in artificial intelligence*, *5*, 35. https://doi.org/10.3389/frai.2022.827584

Chidambaram, V., Chiang, Y. H., & Mutlu, B. (2012, March). Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 293-300). https://doi.org/10.1145/2157689.2157798

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

Chu, Y., & Liu, P. (2023). Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently?. *Cognition*, *239*, 105575. https://doi.org/10.1016/j.cognition.2023.105575

Chung, D., Orloff, M. A., Lauharatanahirun, N., Chiu, P. H., & King-Casas, B. (2020). Valuation of peers' safe choices is associated with substance-naïveté in adolescents. *Proceedings of the National Academy of Sciences*, *117*(50), 31729-31737. https://doi.org/10.1073/pnas.1919111117

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, *55*, 591-621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Collart, J., Gateau, T., Fabre, E., & Tessier, C. (2015, January). Human-Robot Systems Facing Ethical Conflicts: A Preliminary Experimental Protocol. In *AAAI Workshop: AI and Ethics*.

Constantinescu, M., Vică, C., Uszkai, R., & Voinea, C. (2022). Blame it on the AI? on the Moral Responsibility of Artificial Moral Advisors. *Philosophy & Technology*, *35*(2), 35. https://doi.org/10.1007/s13347-022-00529-z

Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J. F., Cebrian, M., ... & Rahwan, I. (2018). Cooperating with machines. *Nature communications*, *9*(1), 233. https://doi.org/10.1038/s41467-017-02597-8

Dashtestani, H., Zaragoza, R., Kermanian, R., Knutson, K. M., Halem, M., Casey, A., ... & Gandjbakhche, A. (2018). The role of prefrontal cortex in a moral judgment task using functional near-infrared spectroscopy. *Brain and behavior*, *8*(11), e01116. https://doi.org/10.1002/brb3.1116

Dashtestani, H., Zaragoza, R., Pirsiavash, H., Knutson, K. M., Kermanian, R., Cui, J., ... & Gandjbakhche, A. (2019). Canonical correlation analysis of brain prefrontal activity measured by functional near infra-red spectroscopy (fNIRS) during a moral judgment task. *Behavioural brain research*, *359*, 73-80. https://doi.org/10.1016/j.bbr.2018.10.022

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85.

De Melo, C., Marsella, S., & Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *23*(2), 1-17. https://doi.org/10.1145/2890495

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, *64*(3), 1155-1170. https://doi.org/10.1287/mnsc.2016.2643

Dinet, J., & Vivian, R. (2015). Perception and attitudes towards anthropomorphic robots in France: Validation of an assessment scale. *Psychologie francaise*, *60*(2), 173-189. https://doi.org/10.1016/j.psfr.2015.05.002

Drugov, M., Hamman, J., & Serra, D. (2014). Intermediaries in corruption: an experiment. *Experimental Economics*, *17*, 78-99. https://doi.org/10.1007/s10683-013-9358-8

Eom, S., & Kim, E. (2006). A survey of decision support system applications (1995–2001). *Journal of the Operational Research Society*, *57*, 1264-1278. https://doi.org/10.1057/palgrave.jors.2602140

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of experimental social psychology*, *79*, 200-216. https://doi.org/10.1016/j.jesp.2018.07.004

Fabre, E. F., Matton, N., Beltran, F., Baragona, V., Cuny, C., Imbert, J. P., ... & Causse, M. (2022). Hierarchy in the cockpit: How captains influence the decision-making of young and inexperienced first officers. *Safety science*, *146*, 105536. https://doi.org/10.1016/j.ssci.2021.105536

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. sage.

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., & Medvedev, A. V. (2019). Temporal derivative distribution repair (TDDR): a motion correction method for fNIRS. Neuroimage, 184, 171-179. https://doi.org/10.1016/j.neuroimage.2018.09.025

Foot, P. (1978). The problem of abortion and the doctrine of double effect. In *Virtues and vices*. Oxford: Blackwell.

Gagnon, L., Yücel, M. A., Boas, D. A., & Cooper, R. J. (2014). Further improvement in reducing superficial contamination in NIRS using double short separation measurements. *Neuroimage*, *85*, 127-135. https://doi.org/10.1016/j.neuroimage.2013.01.073

Gariépy, J. F., Watson, K. K., Du, E., Xie, D. L., Erb, J., Amasino, D., & Platt, M. L. (2014). Social learning in humans and other animals. *Frontiers in neuroscience*, *8*, 58. https://doi.org/10.3389/fnins.2014.00058

Gilet, A. L., Mella, N., Studer, J., Grühn, D., & Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *45*(1), 42. https://doi.org/10.1037/a0030425

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy & technology*, *31*, 169-188. https://doi.org/10.1007/s13347-017-0285-z

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97-103. https://doi.org/10.1016/j.socec.2018.04.003

Goldstein-Greenwood, J., Conway, P., Summerville, A., & Johnson, B. N. (2020). (How) do you regret killing one to save five? Affective and cognitive regret differ after utilitarian and deontological decisions. *Personality and Social Psychology Bulletin*, *46*(9), 1303-1317. https://doi.org/10.1177/0146167219897662

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38–44. https://doi.org/10.1177/0963721414550709

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PloS one*, *8*(4), e60418. https://doi.org/10.1371/journal.pone.0203826

Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., & Shah, J. A. (2015). Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, *39*, 293-312. https://doi.org/10.1007/s10514-015-9457-9

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice taking from humans and machines: An fMRI and effective connectivity study. *Frontiers in Human Neuroscience*, *10*, 542. https://doi.org/10.3389/fnhum.2016.00542

Goodyear, K., Parasuraman, R., Chernyak, S., de Visser, E., Madhavan, P., Deshpande, G., & Krueger, F. (2017). An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social neuroscience*, *12*(5), 570-581. https://doi.org/10.1080/17470919.2016.1205131

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, 267. https://doi.org/10.3389/fnins.2013.00267

Grassian, V. (1981). Moral reasoning: Ethical theory and some contemporary moral problems. Prentice-Hall Wilmington California.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105-2108. https://doi.org/10.1126/science.10628

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400. https://doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D. (2015). The cognitive neuroscience of moral judgment and decision making. In J. Decety & T. Wheatley (Eds.), *The moral brain: A multidisciplinary perspective* (pp. 197–220). Boston Review.

Greene, J. D. (2016). Solving the trolley problem. A companion to experimental philosophy, 173-189. https://doi.org/10.1002/9781118661666.ch11

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-25. https://doi.org/10.1145/3544548.3580959

Gudjonsson, G. H. (1989). Compliance in an interrogative situation: A new scale. *Personality and individual differences*, *10*(5), 535-540. https://doi.org/10.1016/0191-8869(89)90035-4

Hanson, A., Starr, N. D., Emnett, C., Wen, R., Malle, B. F., & Williams, T. (2024, March). The Power of Advice: Differential Blame for Human and Robot Advisors and Deciders in a Moral Advising Context. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 240-249). https://doi.org/10.1145/3610977.3634942

Henkel, A. P., Čaić, M., Blaurock, M., & Okan, M. (2020). Robotic transformative service research: deploying social robots for consumer well-being during COVID-19 and

beyond. *Journal of Service Management*, *31*(6), 1131-1148. https://doi.org/10.1108/JOSM-05-2020-0145

Herold, F., Wiegel, P., Scholkmann, F., & Müller, N. G. (2018). Applications of functional near-infrared spectroscopy (fNIRS) neuroimaging in exercise–cognition science: a systematic, methodology-focused review. *Journal of clinical medicine*, 7(12), 466. https://doi.org/10.3390/jcm7120466

Hertz, N., & Wiese, E. (2018). Under pressure: Examining social conformity with computer and robot groups. *Human factors*, *60*(8), 1207-1218. https://doi.org/10.1177/0018720818788473

Hertz, N., & Wiese, E. (2019). Good advice is beyond all price, but what if it comes from a machine?. *Journal of Experimental Psychology: Applied*, 25(3), 386. https://doi.org/10.1037/xap0000205

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017). Ethically Aligned Design: a Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems Version 2 https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*, *4*(2), e14. https://doi.org/10.23915/distill.00014

Izuma, K. (2017). The neural bases of social influence on valuation and behavior. In J.-C. Dreher & L. Tremblay (Eds.), *Decision neuroscience: An integrative perspective* (pp. 199–209). Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-805308-9.00016-6

Jackson, R. B., & Williams, T. (2019, March). Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 401-410). IEEE. https://doi.org/10.1109/HRI.2019.8673123

Jeurissen, D., Sack, A. T., Roebroeck, A., Russ, B. E., & Pascual-Leone, A. (2014). TMS affects moral judgment, showing the role of DLPFC and TPJ in cognitive and emotional processing. *Frontiers in neuroscience*, *8*, 18. https://doi.org/10.3389/fnins.2014.00018

Jin, W. Y., & Peng, M. (2021). The effects of social perception on moral judgment. *Frontiers in psychology*, *11*, 557216. https://doi.org/10.3389/fpsyg.2020.557216

Kant, I. (2013). *Moral law: Groundwork of the metaphysics of morals.* Routledge.

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

Ladak, A., Loughnan, S., & Wilks, M. (2024). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, *33*(1), 27-34. https://doi.org/10.1177/0963721423120

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*, 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Lee, E. J., & Yun, J. H. (2017). Moral incompetency under time constraint. *Journal of Business Research*, *99*, 438-445. https://doi.org/10.1016/j.jbusres.2017.10.043

Lee, M., Sul, S., & Kim, H. (2018). Social observation increases deontological judgments in moral dilemmas. *Evolution and Human Behavior*, *39*(6), 611-621. https://doi.org/10.1016/j.evolhumbehav.2018.06.004

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? how ai-generated and human-written advice shape (dis) honesty. *The Economic Journal*, *134*(658), 766-784. https://doi.org/10.1093/ej/uead056

Leong, Y. C., & Zaki, J. (2018). Unrealistic optimism in advice taking: A computational account. *Journal of Experimental Psychology: General*, *147*(2), 170. https://doi.org/10.1037/xge0000382

Leyer, M. & Schneider, S. (2019). Me, you or AI? How do we feel about delegation. *In Proc. 27th European Conference on Information Systems (ECIS)* https://aisel.aisnet.org/ecis2019_rp/36

Luke, R., Larson, E., Shader, M. J., Innes-Brown, H., Van Yper, L., Lee, A. K., ... & McAlpine, D. (2021). Analysis methods for measuring passive auditory fNIRS responses generated by a block-design paradigm. *Neurophotonics*, *8*(2), 025008-025008. https://doi.org/10.1117/1.NPh.8.2.025008

Macey-Dare, R. (2023). ChatGPT and Generative AI Systems as Corporate Ethics Advisors. *Available at SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4413206

Macko, A. (2020). Contingencies of self-worth and the strength of deontological and utilitarian inclinations. The Journal of Social Psychology, 161(6), 664–682. https://doi.org/10.1080/00224545.2020.1860882

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117-124). https://doi.org/10.1145/2696454.2696458

Manfrinati, A., Sarlo, M., & Lotto, L. (2013). Un nuovo set di 60 dilemmi morali: dati normativi italiani per giudizi di accettabilità morale, tempi di decisione e valutazioni emozionali. *Giornale italiano di psicologia*, *40*(1), 211-230. https://doi.org/10.1421/73992

McCoy, S. S., & Natsuaki, M. N. (2018). For better or for worse: Social influences on risk-taking. *The Journal of social psychology*, *158*(2), 139-151. https://doi.org/10.1080/00224545.2017.1294139

Meshi, D., Biele, G., Korn, C.W., & Heekeren, H.R. (2012). How Expert Advice Influences Decision Making. *PLoS ONE* *7*(11): e49748. https://doi.org/10.1371/journal.pone.0049748

Mill, J. S. (1998). Utilitarianism, ed. Roger Crisp. *Oxford: Oxford University Press*, *5*, 3-4.

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of neuroscience*, *22*(7), 2730-2736. https://doi.org/10.1523/JNEUROSCI.22-07-02730.2002

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature reviews neuroscience, 6*(10), 799-809. https://doi.org/10.1038/nrn1768

Moll, J., & de Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in cognitive sciences*, *11*(8), 319-321. https://doi.org/10.1016/j.tics.2007.06.001

Moll, J., De Oliveira-Souza, R., & Zahn, R. (2008). The neural basis of moral cognition: sentiments, concepts, and values. *Annals of the New York Academy of Sciences*, *1124*(1), 161-180. https://doi.org/10.1196/annals.1440.005

Nasello, J. A., Dardenne, B., Blavier, A., & Triffaux, J. M. (2021). Does empathy predict decision-making in everyday trolley-like problems?. *Current Psychology*, 1-14. https://doi.org/10.1007/s12144-021-01566-1

Nasello, J. A., & Triffaux, J. M. (2023). The role of empathy in trolley problems and variants: A systematic review and meta-analysis. *British Journal of Social Psychology*. https://doi.org/10.1111/bjso.12654

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *7*(3), 437-454. https://doi.org/10.1075/is.7.3.14nom

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., & Gross, J. J. (2004). For better or for worse: neural systems supporting the cognitive down-and up-regulation of negative emotion. *Neuroimage*, *23*(2), 483-499. https://doi.org/10.1016/j.neuroimage.2004.06.030

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., ... & Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology*, *120*(2), 443. https://doi.org/10.1037/pspp0000281

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, 107600. https://doi.org/10.1016/j.chb.2022.107600

Plaks, J. E., Robinson, J. S., & Forbes, R. (2022). Anger and Sadness as Moral Signals. *Social Psychological and Personality Science*, *13*(2), 362-371. https://doi.org/10.1177/19485506211025909

Pollonini, L., Olds, C., Abaya, H., Bortfeld, H., Beauchamp, M. S., & Oghalai, J. S. (2014). Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hearing research*, 309, 84-93. https://doi.org/10.1016/j.heares.2013.11.007

Quaresima, V., & Ferrari, M. (2019). Functional near-infrared spectroscopy (fNIRS) for assessing cerebral cortex function during human behavior in natural/social situations: a concise review. *Organizational Research Methods*, *22*(1), 46-68. https://doi.org/10.1177/1094428116658959

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin, 114*(3), 510. https://doi.org/10.1037/0033-2909.114.3.510.

Reynolds, C. J., Knighten, K. R., & Conway, P. (2019). Mirror, mirror, on the wall, who is deontological? Completing moral dilemmas in front of mirrors increases deontological but not utilitarian response tendencies. *Cognition*, *192*, 103993. https://doi.org/10.1016/j.cognition.2019.06.005

Riva, P., Aureli, N., & Silvestrini, F. (2022). Social influences in the digital era: When do people conform more to a human being or an artificial intelligence?. *Acta Psychologica*, *229*, 103681. https://doi.org/10.1016/j.actpsy.2022.103681

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016, March). Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 101-108). IEEE. https://doi.org/10.1109/HRI.2016.7451740

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology*, *69*, 44-58. https://doi.org/10.1016/j.jesp.2016.09.007

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology*, *74*, 24-37. https://doi.org/10.1016/j.jesp.2017.08.003

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI magazine*, *36*(4), 105-114. https://doi.org/10.1609/aimag.v36i4.2577

Sacco, D. F., Brown, M., Lustgraaf, C. J., & Hugenberg, K. (2017). The adaptive utility of deontology: Deontological moral decision-making fosters perceptions of trust and likeability. *Evolutionary Psychological Science*, *3*(2), 125–132. https://doi.org/10.1007/s40806-016-0080-6

Sandoval, E. B., Brandstetter, J., & Bartneck, C. (2016, March). Can a robot bribe a human? The measurement of the negative side of reciprocity in human robot interaction. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 117-124). IEEE. https://doi.org/10.1109/HRI.2016.7451742

Sarlo, M., Lotto, L., Manfrinati, A., Rumiati, R., Gallicchio, G., & Palomba, D. (2012). Temporal dynamics of cognitive–emotional interplay in moral decision-making. *Journal of Cognitive Neuroscience*, *24*(4), 1018-1029. https://doi.org/10.1162/jocn_a_00146

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama*, *320*(21), 2199-2200. https://doi.org/10.1001/jama.2018.17163

Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nature Machine Intelligence*, *1*(8), 330-331. https://doi.org/10.1038/s42256-019-0084-6

Stade, E. C., Stirman, S. W., Ungar, L. H., Schwartz, H. A., Yaden, D. B., Sedoc, J., … Eichstaedt, j. C. (2023, April 24). Artificial intelligence will change the future of

psychotherapy: A proposal for responsible, psychologist-led development. https://doi.org/10.31234/osf.io/cuzvr

Strait, M., & Scheutz, M. (2014). Using functional near infrared spectroscopy to measure moral decision-making: effects of agency, emotional value, and monetary incentive. *Brain-Computer Interfaces*, *1*(2), 137-146. https://doi.org/10.1080/2326263X.2014.912886

Straßmann, C., Eimler, S. C., Arntz, A., Grewe, A., Kowalczyk, C., & Sommer, S. (2020). Receiving robot's advice: Does it matter when and for what?. In *Social Robotics: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings 12* (pp. 271-283). Springer International Publishing. https://doi.org/10.1007/978-3-030-62056-1_23

Svenmarck, P., Luotsinen, L., Nilsson, M., & Schubert, J. (2018, May). Possibilities and challenges for artificial intelligence in military applications. In *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting* (pp. 1-16).

Szekely, R. D., & Miu, A. C. (2014). Incidental emotions in moral dilemmas: The influence of emotion regulation. *Cognition and Emotion*, *29*(1), 64–75. https://doi.org/10.1080/02699931.2014.895300

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., ... & Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Social cognitive and affective neuroscience*, *7*(3), 282-288. https://doi.org/10.1093/scan/nsr008

Takamatsu R (2018). Turning off the empathy switch: Lower empathic concern for the victim leads to utilitarian choices of action. *PLoS ONE 13*(9): e0203826. https://doi.org/10.1371/journal.pone.0203826

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The monist*, *59*(2), 204-217.

Thomson, J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

Trémolière, B., & Rateau, P. (2023). You're heartless, I'm less: self-image and social norms in moral judgment. The Journal of General Psychology, 151(2), 112–137. https://doi.org/10.1080/00221309.2023.2218637

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326-334. https://doi.org/10.1016/j.cognition.2012.10.005

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72-81. https://doi.org/10.1177/1745691614556679

Wasilow, S., & Thorpe, J. B. (2019). Artificial intelligence, robotics, ethics, and the military: A Canadian perspective. *AI Magazine*, *40*(1), 37-48. https://doi.org/10.1609/aimag.v40i1.2848

Yu, H., Siegel, J. Z., Clithero, J. A., & Crockett, M. J. (2021). How peer influence shapes value computation in moral decision-making. *Cognition*, *211*, 104641. https://doi.org/10.1016/j.cognition.2021.104641

Yücel, M. A., Lühmann, A. V., Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., ... & Wolf, M. (2021). Best practices for fNIRS publications. *Neurophotonics*, 8(1), 012101-012101. https://doi.org/10.1117/1.NPh.8.1.012101

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, *101*, 104327. https://doi.org/10.1016/j.jesp.2022.104327

Zheng, H., Lu, X., & Huang, D. (2018). tDCS over DLPFC leads to less utilitarian response in moral-personal judgment. *Frontiers in Neuroscience*, *12*, 193. https://doi.org/10.3389/fnins.2018.00193

Zhou, X., Sobczak, G., McKay, C. M., & Litovsky, R. Y. (2020). Comparing fNIRS signal qualities between approaches with and without short channels. PloS one, 15(12), e0244186. https://doi.org/10.1371/journal.pone.0244186

Multiple Regression Analysis	Unstandardized Coefficients		Standardized Coefficients	*t*	*p*	CI (95%)	
	B	*SE*	*Beta*			*Inferior*	*Superior*
Utilitarinism in the baseline run							
Constant	.362	.313		1.157	.257	-.278	1.002
IRI_Perspective Taking subscale	.036	.066	.100	.539	.594	-.100	.172
Propensity to change mind							
Constant	- 72.620	23.639		- 3.072	.005	-121.122	- 24.117
Gudjonsson's compliance scale	11.892	2.830	.597	4.202	**.000***	6.036	17.685
NARS_Emotion subscale	3.545	1.429	.355	2.492	**.020***	.613	6.477
IRI_Perspective Taking subscale	8.630	4.306	.284	2.057	.055	-.205	17.464
Propensity to delegate							
Constant	-60.502	23.84		-2.538	.017	-109.419	-11.586
Gudjonsson's compliance scale	8.318	2.863	.457	2.905	**.007***	2.444	14.193
NARS_Emotion subscale	3.523	1.441	.385	2.445	**.021***	.566	6.479
IRI_Perspective Taking subscale	7.56	4.342	.272	1.741	.093	-1.35	16.47

Table 1. Summary of the multiple regression analyses performed on the questionnaire ratings and the decision rates.

ANOVA	Left ROI			Rostral ROI			Right ROI		
	F(1,30)	*p*	*ηp²*	*F(1,30)*	*p*	*ηp²*	*F(1,30)*	*p*	*ηp²*
Dilemma x Presentation Order (2nd vs 3rd)									
Dilemma	.120	.732	.004	1.194	.283	.038	.481	.493	.016
Order	.231	.634	.008	.013	.909	.000	.027	.869	.001
Dilemma x Order	.256	.616	.008	**4.668**	**.039***	**.135**	.013	.910	.000
Dilemma x Argument (Time Locked to Argument)									
Dilemma	2.387	.133	.007	1.562	.221	.049	.006	.941	.000
Argument	.348	.560	.011	1.078	.308	.035	**8.109**	**.007***	**.213**
Dilemma x Argument	.037	.849	.001	.061	.806	.002	1.749	.196	.055
Dilemma x Argument (Before Answer)									
Dilemma	.025	.876	.001	.180	.674	.006	.071	.792	.002
Argument	.270	.607	.009	.179	.286	.038	.020	.888	.001
Dilemma x Argument	1.955	.172	.061	.208	.651	.007	.383	.546	.013
Dilemma x Argument (Around answer)									
Dilemma	1.441	.239	.046	2.605	.117	.080	.420	.522	.014
Argument	.001	.969	.000	2.977	.095	.090	1.017	.321	.033
Dilemma x Argument	2.371	.134	.073	1.904	.178	.060	.000	.982	.000

Paired t-tests	Left ROI			Rostral ROI			Right ROI		
	t	*df*	*p*	*t*	*df*	*p*	*t*	*df*	*p*
Dilemma									
Time-Locked to Dilemma	.336	30	.739	.012	30	.192	- .186	30	.854
Argument Nature									
Time-Locked to Argument	1.232	30	.228	1.064	30	.296	- .045	30	.964
Before Response	- 1.330	30	.193	.967	30	.341	- .277	30	.783
Around Response	- .043	30	.966	- .584	30	.563	- .045	30	.964

Table 2. Summary of the statistical analyses performed on the *f*NIRS measurements.

- AI-advisors can influence human beings' moral decisions.
- Receiving deontological arguments from AI-advisors decreased participants' appraisal of the affective response to the dilemmas.
- The influence AI-advisors depends on human advisees' susceptibility to obey and to fear non-human agents.
- These results confirm the necessity to define clear rules to regulate the way artificial moral advisors are trained and used in everyday life.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: