Brain-Computer Interfaces

ISSN: 2326-263X (Print) 2326-2621 (Online) Journal homepage: https://www.tandfonline.com/loi/tbci20

Using functional near infrared spectroscopy to measure moral decision-making: effects of agency, emotional value, and monetary incentive

Megan Strait & Matthias Scheutz

To cite this article: Megan Strait & Matthias Scheutz (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, DOI: 10.1080/2326263X.2014.912886

To link to this article: https://doi.org/10.1080/2326263X.2014.912886

Published online: 09 May 2014.

Brain-Computer Interfaces, 2014
Vol. 1, No. 2, 137–146, http://dx.doi.org/10.1080/2326263X.2014.912886

Taylor & Francis
Taylor & Francis Group

Using functional near infrared spectroscopy to measure moral decision-making: effects of agency, emotional value, and monetary incentive

Megan Strait* and Matthias Scheutz

Department of Computer Science, Tufts University, Medford, MA, USA

(*Received 22 December 2013; accepted 9 March 2014*)

The prefrontal cortex (PFC) has been investigated extensively with functional magnetic resonance imaging (fMRI) and identified as a neural substrate central to emotion regulation and decision-making, particularly in the context of utilitarian moral dilemmas. However, there are two important limitations to prior work: (1) fMRI imposes strict constraints on the physical environment of the participant and (2) experimental manipulations have yet to consider the role of agency and personal incentive on both brain-based and behavioral correlates. To address the first limitation, we investigated functional near infrared spectroscopy (NIRS), which showed it was a potential alternative to fMRI for observing the decision-making processes in a less-constrained environment [1]. To address the second, we examined the role of agency in deciding moral and non-moral dilemmas and whether the influences can be further modulated by way of monetary incentives. Our findings show that all three factors exert influences on both behavioral and neural metrics. In particular, emotional value increases, whereas incentive decreases, prefrontal hemodynamic activity. Moreover, agency interacts with both emotional value and incentive, further polarizing the behavioral and neural metrics with regard to human patients.

Keywords: functional near infrared spectroscopy (NIRS); decision-making; moral dilemma; agency; incentive

1. Introduction

Utilitarian dilemmas, which involve a conflict between competing imperatives, have long been used by philosophers and psychologists to study the cognitive processes involved in emotion regulation and decision-making. A standard example of a moral utilitarian dilemma is that of the 'trolley problem', which is formulated as follows:

> Suppose there is a runaway tram which can only be steered from one track on to another. Five people are working on one track and one person on the other; anyone on the track which is entered is bound to be killed.

The observer (the participant) must decide whether to exchange one person's life for the lives of five or to exchange five lives for one. The utilitarian view seeks to maximize welfare (or minimize harm), and as such, the morally preferred course of action is to steer the train to the track with only one person. However, an alternative view asserts moving to another track constitutes a participation in the moral wrong, making one partially responsible for the death when otherwise no one would be responsible.

Over the past decade this utilitarian dilemma has been employed in a multitude of neuroimaging studies to identify the psychological and neural substrates underlying emotion regulation and decision-making, highlighting, in particular, the role of the prefrontal cortex (e.g., [2–4]). However, this work is conducted within the confines of an fMRI, which imposes a number of constraints that limit experimental conditions to unrealistic settings. Furthermore, these neuroscientific studies largely ignore potential modulatory factors: in particular, the role of the perceived agency – or capacities (e.g., the capacity to feel pain) – of the dilemma's patients (those affected by the dilemma outcomes) as well as personal incentives towards one outcome or another.

In comparison to fMRI, functional near infrared spectroscopy (NIRS) is *relatively* portable and unobtrusive, making it a potential alternative for measuring hemodynamic activity in less constrained environments (e.g., [5]). Specifically, it is not limited to use in one location and restricts participants' movement to a lesser extent (tethered within range of the measuring device, as opposed to lying motionless in a noisy cylindrical tube). This avoids the imposition of highly unrealistic constraints on participants' natural behavior (i.e., moving) that fMRI poses, hence yielding better ability to conduct experiments in more relevant settings. As agency attribution has been shown to be a dynamic process subject to experience and interaction (e.g., [6,7]), it is important to extend neuroimaging techniques to environments in which interaction with physical agents can occur. Increased mobility would thus allow participants to be

*Corresponding author. Email: megan.strait@tufts.edu

placed within the context of a dilemma, rather than presented with a hypothetical.

In this paper, we first summarize our investigation into the ability of NIRS as a technique for measuring decision-making in the standard fMRI paradigm of utilitarian moral dilemma. We then extend related work by investigating the influences of both agency ascription and personal incentive in moral dilemmas via a series of four experimental manipulations. We conclude with a discussion of the implications and limitations of NIRS for evaluating emotionally sensitive decision-making cognitive processes in human-agent interaction settings.

2. Related work

Below we describe related neuroscientific studies concerning the investigation of moral decision-making and its neural correlates, followed by evidence of agency ascription and incentive as important factors in decision-making.

2.1. Manipulation of emotional artifacts

Extensive work using lesion studies and transcranial magnetic stimulation (TMS) has identified the anterior prefrontal cortex (PFC) as central to moral decision-making.[2–4,8–12] Additional work has shown direct connections between the ventromedial prefrontal cortex (vmPFC) and the amygdala in emotion regulation (ER) tasks (e.g., [13,14]). As the standard utilitarian dilemma involves ER due to negative stimuli (e.g., [15], vmPFC activity is also observed in moral decision-making (e.g., [4]).

To further explore the PFC and to what extent various factors implicit in utilitarian dilemmas might have, additional studies have been conducted. These include, in particular, artifacts hypothesized to elicit increased emotional engagement (e.g., [16–20]), such as:

- *Action immediacy*: participant performs the action versus the participant tells a surrogate to perform the action.[21]
- *Personal force*: participant performs direct harm (i.e. pushes a person in front of a train) versus indirect harm (i.e. switches the train tracks) (e.g., [4,22]).
- *Visual immediacy*: the participant imagines (versus views pictures of) the scenario.[23]

Other manipulations have investigated the effects of cognitive load,[24] honesty,[25] intent,[26] and stereotype.[27]

However, this work is limited in two key aspects: first, the experimental conditions are highly constrained and thus not representative of realistic conditions or contexts surrounding moral dilemmas and, second, they do not consider the role of agency of the patients (those being affected by the dilemma outcomes) in the dilemma. That is, dilemmas employed in related work concern *only human patients*, as opposed to patients theoretically attributed less agency such as animals (e.g., cats and dogs). Thus it remains unaddressed whether it is the patient's ascribed agency or the emotional context or both that elicits the corresponding hemodynamic changes and how that may affect participants' behavioral decisions.

2.2. Evidence for the role of agency

Recent work in human-robot interaction studies suggests that agency factors into decision-making. A two-part study mimicking the Stanley Milgram experiments showed that perceptions of agency in robotic artifacts play a role in moral decision-making.[28] It first demonstrated that people have less concern for robotic agents than human counterparts, and proceeded to show that humans had more willingness and enthusiasm to destroy robots of lower perceived agency. Another recent study found effects of perceived agency on how successfully a robot could dissuade a human participant from completing an emotionally sensitive task.[29] Although these studies focus on robots and are not of the standard utilitarian dilemma employed in the above imaging studies, the observed differences in behavioral outcomes suggest the level of agency ascribed influences the processes involved in decision-making. Given that lower perceptions of agency manifested behaviorally as less concern for patients, we expected that less agency would also manifest as lower activation of the prefrontal cortex (and conversely, greater perceptions of agency would yield greater activation).

2.3. Influence of personal incentive

In addition to the potential influences of agency on moral decision-making, the physical constraints of fMRI-based investigations place implicit distance between participants and the utilitarian dilemmas, limiting the degree to which personal incentives can be examined for influencing both brain and behavior in moral decision-making. Specifically, the constraints require the presentation of dilemmas as hypotheticals (e.g., '*imagine/suppose* there is a runaway tram …'), which severely limits the personal involvement of participants. Results of such investigations thus rely on participants' individual incentives or willingness to fully consider the dilemmas.

Given the overlapping neural substrates of incentive-based and utilitarian-based processing (e.g., [30–32]), it is thus possible that personal incentive may thus be a

confounding factor in the degree of neural activity underlying moral judgments. In particular, we hypothesized that increasing personal involvement in the decisions (e.g., by placing participants within the context of the dilemmas) would be reflected as increased prefrontal activity. Hence, it is of interest as to whether personally incentivized decision-making results in differentiable effects on both brain and behavior correlates. In order to investigate such influences, we introduced a personal incentive to bias towards one outcome. As the manipulation would reduce the personal incentive to consider the dilemma fully (thus reducing the difficulty of the decisions), we expected that providing a personal incentive to bias towards one outcome would result in reduced neural activity.

In the following sections, we attempt to address the aforementioned limitations, namely the physical limitations that fMRI faces, by employing NIRS for measuring hemodynamic activity in more realistic settings. We secondly attempt to address (1) the role of agency in decision-making by varying the agent types (robot versus canine versus human) involved in utilitarian decision-making tasks and (2) the influence of personal involvement by varying the degree of incentive (none versus a $5 monetary bonus) in an extension of Strait et al.[1]

3. Preliminary investigation using NIRS to measure moral decision-making

We previously investigated whether NIRS can measure the hemodynamic activity in the prefrontal cortex associated with moral decision-making. In Strait et al.[1] we reported a within-subjects experiment based on the standard moral decision-making paradigms used with fMRI. Here we summarize the details and findings of that experiment, which leads into the follow-up, four-part investigation to distinguish the relative effects of agency, emotional value, and monetary incentive.

3.1. Paradigm

We constructed a set of 16 standard utilitarian moral dilemmas and a corresponding scenario to explain the task to the participants. Participants were instructed that they would be managing 'emergency evacuations' to transport endangered patients to safety. For example, given an office fire, participants were presented with a dilemma formulated as follows.

> You come across a room with a man trapped underneath fallen debris. You can hear several people crying for help in the room next door because the door is jammed. The man trapped underneath the debris sees you and cries for you to help him. There is not enough time to help both the man and the people next door.

Participants were then prompted to answer the following question: 'What do you do? *Save the* man or *leave him* to save the people next door?' and were shown two buttons with the labels 'save' and 'leave' to select from. Additionally, we instructed participants to have a goal of *evacuating* as many patients as possible. While this presents a limitation of the paradigm (potentially reducing the emotional value of the dilemmas), we chose to do so, so as to avoid having participants randomly choosing between the two outcomes.

3.2. Stimuli

We designed two treatment conditions, a *control* condition to serve as a baseline comparison for our *test* condition, as we had a limited region of measurement (only the PFC). In the *control* condition, the participant evacuated eight nonliving, inanimate patients (glass-blown objects), and in the *test* condition the participant evacuated eight standard human patients (see Figure 1). A set of eight images of blown-glass objects and eight random images of people were collected and paired with the 16 dilemmas. Conditions were administered in blocks (of eight trials, randomized) preceded by a 30-s resting sample (for the conversion to hemoglobin) and instructions (i.e. 'there is a fire in office x, evacuate as many people as possible'), and counterbalanced. A trial was composed of four parts:

- A pre-dilemma period (30 s) consisting of a simple counting task to ensure the participant's attention to the computer screen.
- A textual description (15 s) of a moral dilemma accompanied by a randomized photo of an agent. The type of the agent displayed (either glass or human) was in accordance with the condition (control or test, respectively). In the *control* condition, the dilemmas also employed physical impediments to evacuating the objects in a timely fashion (i.e., instead of 'a boy in a wheelchair' one might have

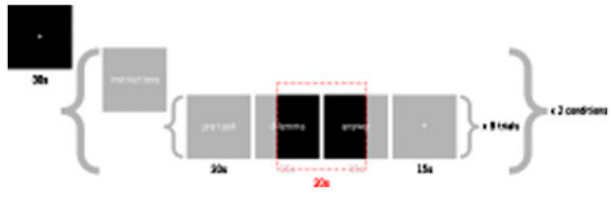

Figure 1. (Color online) Experimental protocol: a pre-dilemma task signaled the participant's attention and a fixation point signaled a resting period. The parts indicated in red contained a moral dilemma stimulus and decision, during which the participant's hemodynamic activity was sampled. The ordering of conditions was counterbalanced between-subjects and the ordering of trials was randomized by subject.

a blown-glass object too large or too heavy, etc. to evacuate quickly).
- An answer period (15 s) during which the participant selected an action option (save or leave) in response to the dilemma.
- A rest period (15 s) where the participant focused on a fixation point and relaxed.

3.3. Population and procedure

Ten subjects (five male), ages 19 to 33 (M = 22.0, SD = 3.8) were recruited and provided informed, written consent. To avoid effects from task confusion, learning, or difficulty, the participants first completed four practice trials. Following that, the participants completed the two conditions (eight trials per condition) in succession.

3.4. Data acquisition

A two-probe (two-channel), ISS OxiplexTS near infrared tissue oximeter was used to record hemodynamic activity in the PFC at a temporal resolution of 6.25 Hz. An elastic, black headband was used to securely fit the NIRS probes in place on the subject's forehead to sample the left and right PFC respectively. The subject was seated in a generic office chair (Zoom Seating, model SP46105).

3.5. Signal processing

The ISS OxiplexTS records relative absorption and scattering coefficients of the sampled tissue, which require additional processing prior to statistical analysis. The raw measures are first converted to hemoglobin values using the modified Beer Lambert Law (MBLL). We used the MBLL implementation from NIRS-SPM, a publicly available processing and analysis package for NIRS data.[33] This yields a measure of deoxygenated (Hb) and oxygenated (HbO_2) hemoglobin for each probe (left and right). We first applied a high-pass filter to remove low-frequency artifacts such as Mayer waves, followed by a low-pass filter to smooth artifacts arising from cardiac pulsations and respiration.[1] Lastly, we applied a correlation-based signal correction (CBSI) [34] and then averaged over trial repetitions, to reduce the effects of task-unrelated artifacts.

Since the CBSI correction is calculated from the correlation between Hb and HbO_2 measures, the deoxygenated (Hb) measures become redundant and are thus discarded from statistical consideration at this point. We then truncate the HbO_2 signals to a 20-s window after the delivery of the dilemma instruction (5–25 s, exclusive). This 20-s truncated signal maximizes chances of capturing the peak of the hemodynamic response even if the peak of the response varies temporally as a function of condition. As a result, we have two 20-s HbO_2 signals (left and right PFC) per condition.

3.6. Statistical analysis

As the investigation was a fully within-subjects design with one independent variable (standard moral dilemma with human patients versus baseline) and two dependent variables (left and right prefrontal hemodynamic activity), we inferred significance using two, two-tailed matched-pair *t*-tests. Specifically, to determine whether the standard moral dilemma with human patients elicited significant task-related hemodynamic changes, we conducted the two-tailed matched-pair *t*-tests on the difference in mean oxygenated hemoglobin (HbO_2) between the two conditions (see Figure 2) with the null hypothesis that the difference was zero. Here we represented the difference in HbO_2 between the two conditions using an area-under-the-curve (AUC) summary statistic. The AUC statistic is calculated by summing the signal change (condition hemoglobin concentration − baseline hemoglobin concentration) and taking the average of the 20-s truncated signal. This results in one mean AUC value for the left and for the right PFC, for each participant, for a total sample size of $N = 10$.

3.7. Results

As expected, the difference in activity between conditions showed that significantly greater activation was elicited by the human condition compared to the baseline in both the left ($t_{obs} = 5.1108$, $t_{crit}(9) = 4.7809$, $p < 0.0010$) and right ($t_{obs} = 4.9942$) prefrontal cortex. Specifically, the difference in mean HbO_2 (computed by subtracting baseline activity from the test condition activity) in both hemispheres was found to be significantly greater than zero. The significant differences observed are thus suggestive that NIRS is capable of capturing activity related to moral decision-making in the PFC. Hence we next proceeded with our follow-up experiments to investigate the relative roles of agency and incentive in decision-making.

4. Effect of agency, moral value, and personal incentive on decision-making

Via a four-part extension, we examined the effects of agency, moral value, and incentive on both behavioral and neural indices of decision-making. To evaluate their relative effects, we designed four protocols (between-subjects), where we constructed a set of both moral and non-moral dilemmas crossed with four agent manipulations (within-subjects). A preliminary report of the methods and results of the moral versus non-moral protocol

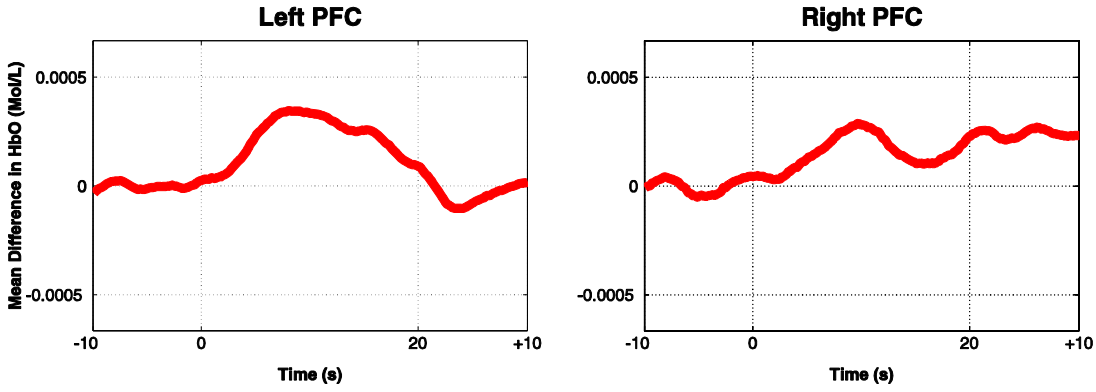

Figure 2. (Color online) Mean difference in oxygenated hemoglobin between the two conditions (human versus glass) across participants ($N = 10$). In the plots, the dilemma presentation begins at time $t = -5$ s and the answer period terminates at $t = 25$ s. The signal is truncated from 0 to 20 s to capture the maximum change in hemoglobin.

manipulation are published in Strait et al.[1] Below we extend both the analysis and discussion of the three manipulations.

4.1. *Agent manipulation*

To investigate the role of the agency in decision-making (emotional and non-emotional), we introduced two agent categories – dog and robot – in addition to the human and glass (control) categories used in our preliminary investigation. These four categories were selected to represent four hypothetical levels of agency (see Figure 3). We selected robots as prior work suggested lesser agency is ascribed to robots in comparison with humans (e.g., [28,35]), and specifically used the Aldebaran Nao for its humanoid appearance. We additionally chose the canonical canine pet based on Gray et al. [7] which showed canines were ascribed more agency than robots, but less than humans.

Hence we had four agent types (glass, robot, canine, and human) – where the agent type represented the patient (e.g., person facing life/death) in the dilemma – to evaluate the effects of agency. Agent categories were administered in blocks of six trials (randomized) and preceded by a 30-s fixation period (for post hoc conversion of raw NIRS data to hemoglobin). Ordering of the agent conditions was counterbalanced.

4.2. *Moral and non-moral protocols*

To disentangle the contributions from the emotional value of the decision (e.g., moral dilemma) versus value of the agent involved (e.g., canine), we designed two protocols to manipulate the moral value of the decision. Thus we constructed a set of 24 *moral* utilitarian dilemmas (six dilemmas per each of the four agent conditions) and 24 *non-moral* utilitarian dilemmas. The number of trials was reduced from eight (used originally in the preliminary investigation) to six in order to avoid significantly extending the total session time.

In the non-moral protocol, the evacuation scenario was modified to be a *relocation* scenario, where the

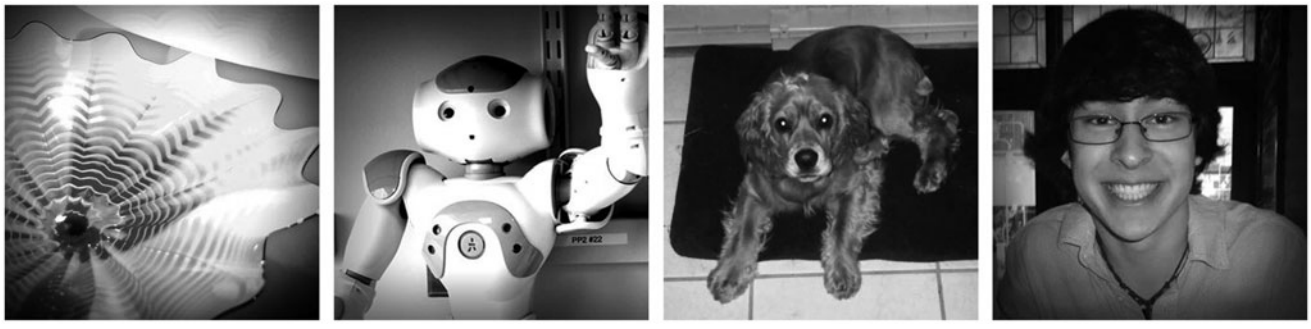

Figure 3. Agent categories: an inanimate glass object (far left) and human (far right) types were used in all experiments. The robot (Aldebaran Nao), middle left, and canine, middle right, agent types were added in the follow-up experimental series to evaluate effects of perceived agency.

goal was to *relocate* as many patients (agents) as possible. As in the moral protocol, we constructed a set of 24 non-moral utilitarian dilemmas (with one additional modification to the answer stimulus: instead of the option to 'save', the participant now had the option to 'take'). As an example of a non-moral dilemma (in the instance of a blown-glass object) the dilemma read as too large or too heavy, etc. to *transport* (instead of *evacuate*) quickly.

Based on the expectation that exposure to both protocols would bias participants, the protocol manipulation was administered between-subjects (as opposed to the within-subjects agent manipulation).

4.3. Reward stimulus

Following the 30-s fixation period and prior to the first of the four agent blocks, a reward stimulus was presented (see Figure 4). The stimulus offered $5 bonus for good performance, where good performance was defined as *evacuating* (in the moral protocol) or as *transporting* (non-moral protocol) as many agents as possible (i.e., making the most utilitarian – sacrificing one person for the lives of many – judgments possible). We expected this manipulation to bias participants towards the utilitarian decision, thus reducing the difficulty (and hence, expected prefrontal activation) of the decisions.

4.4. Measures

Two behavioral metrics, (1) *agency ascription* and (2) *dilemma outcome*, as well as the NIRS-based indirect measure of (3) *neural activity* were sampled. Participants' ascriptions of agency were operationalized by five dimensions:

- capacity to make decisions
- capacity to feel physical/emotional pain
- having of desires/preferences/intentions/goals
- having of common sense, and
- having of free will.

The agency measure was collected via a post-questionnaire using a 5-point Likert scale to rate each dimension (see Figure 5). A measure of dilemma outcome between −1 (corresponding to *kill* in the moral protocol and *leave* in the non-moral protocol) to 1 (*save*/*take*) was calculated by summing and then averaging each participant's decisions across the six trials for each agent type. Neural activity was recorded in the left and right prefrontal cortex again using the two-channel ISS OxiplexTS near infrared tissue oximeter (with a temporal resolution of 6.25 Hz). As we did previously, the raw data were then converted, filtered, and truncated prior to statistical analysis.

4.5. Population

Forty Tufts University students and staff were recruited to participate (10 per condition): 10 (three male) in the moral/non-incentivized protocol, ages 18 to 31 (M = 20.8, SD = 3.6); 10 (also three male), ages 19 to 22 (M = 20.2, SD = 1.3), in the non-moral/non-incentivized protocol; 10 (four male), ages 18 to 22 (M = 19.9, SD = 1.5), for the moral/incentivized protocol; and 10 (three male), ages 18 to 23 (M = 20.0, SD = 2.1), for the non-moral/incentivized protocol. All participants reported being right-handed, with no history of brain trauma. Participants provided informed, written consent and were paid $10 for their involvement. Participants again completed four practice trials prior to the experimental conditions and conditions were counterbalanced.

4.6. Statistical analysis

The four dependent variables – the two behavioral metrics (agency ratings and dilemma outcomes) and the two hemispheres of NIRS-based activity – were analyzed with a mixed-methods ANOVA model with the following independent variables: agent category (glass vs. robot vs. canine vs. human; within-subjects), emotional value (moral vs. non-moral; between-subjects), and incentive (incentivized vs. non-incentivized; between) and all related two-way and three-way interaction effects of those factors.

4.7. Results

4.7.1. Agency ascription

As expected, there was a main effect of agent category on subjective ratings of agency, confirming our

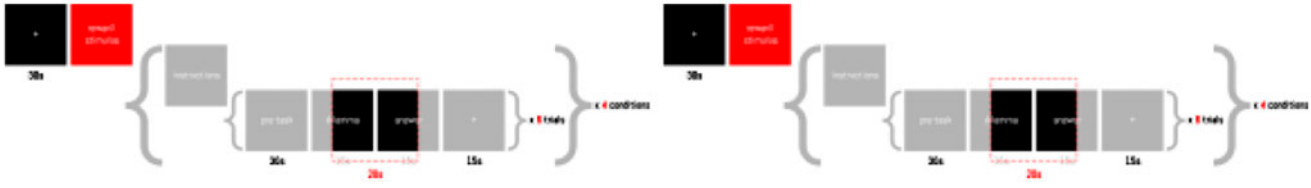

Figure 4. (Color online) Incentivized protocol. A reward stimulus is presented after the 30-s baseline acquisition and prior to receiving the dilemma stimuli. There were four agent categories manipulated within-subjects (six trials of each category).

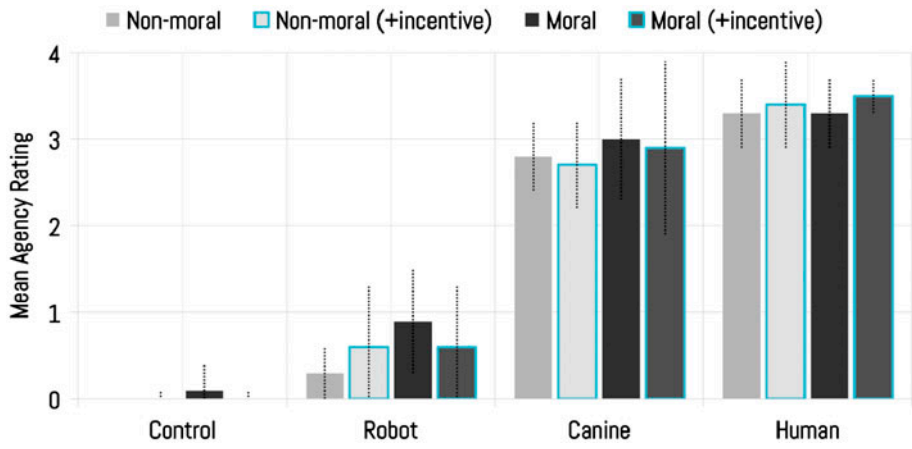

Figure 5. (Color online) Mean agency ascription across subjects (N = 10) for each agent category (averaged across the five dimensions of mental-state attribution). Ratings were normalized prior to averaging across subjects. Error bars represent standard deviation.

experimental manipulation of agent type. Specifically, human agents were attributed the highest level of agency, followed by dogs, then robots, and lowest, the control agent (glass): $F(3159) = 434.31$, $p < 0.0001$. Post-hoc multiple comparisons revealed that between living and non-living agent categories, all pair-wise comparisons were significant (i.e., human agents and canine agents were attributed significantly more agency than both robot and glass categories). However, within those categories (i.e., robot versus glass, human versus canine) – with the exception of the moral protocol, where the agency ascribed to robots was significantly greater than that of the control (glass) category – all other comparisons were non-significant (e.g., agency ratings between human and canines was not significant).

Moral value (moral vs. non) also showed a main effect on subjective ratings of agency ($F(1159) = 4.11$, $p = 0.0444$), with significantly higher ratings observed in the moral protocol. Surprisingly, however – counter to our hypotheses – no significant effects were observed due to monetary incentivization ($F(1159) = 0$, $p = 0.9496$).

4.7.2. Dilemma outcomes

Similarly, there was a main effect of agent category on the dilemma outcomes ($F(3159) = 9.68$, $p < 0.0001$). However, post hoc multiple comparisons showed only the human agent type was significantly affected (i.e., no statistically significant differences between glass vs. robot vs. canine), with the highest likelihood of saving/transporting (see Figure 6).

There was also, again, an effect due to the emotional value ($F(1159) = 7.62$, $p = 0.0065$), as well as an interaction effect between agent category and emotional value ($F(3159) = 14.71$, $p < 0.0001$). Multiple comparisons showed participants overall more likely to leave an agent (in exchange for the lives of more) in the moral protocol, whereas in the non-moral protocol they were not more likely to transport or to leave the agents. However, participants were significantly more likely to save living agents (human and canine) than non-living agents (glass and robot). Yet, again there was no overall significant effect due to the reward incentive ($F(1159) = 1.22$, $p = 0.2713$).

4.7.3. Neural activity

Statistical analysis with our ANOVA model revealed, regarding the *left* anterior prefrontal cortex, there was no main effect of either agent type or of incentive on the hemodynamic activity observed. There was, however, a significant main effect due to moral value ($F(1159) = 5.74$, $p = 0.0179$), with the moral protocol eliciting greater increases in oxyhemoglobin than the non-moral protocol.

Regarding the *right* PFC, there was a main effect of both emotional value ($F(1159) = 7.22$, $p = 0.0080$) and of incentive ($F(1159) = 12.94$, $p = 0.0004$). As in the left PFC, again the moral protocol showed greater hemodynamic change than the non-moral protocol. The effect of the reward stimulus, however, reduced the change in oxyhemoglobin in comparison to the non-incentivized protocols (see Figure 7) in line with our hypothesis. In addition to the main effects of moral value and incentive, there was also an interaction effect between agency and moral value, $F(3159) = 2.93$, $p = 0.0358$ (see Figure 8), and a trend towards a significant interaction effect with incentive, $F(3159) = 2.53$, $p = 0.0651$).

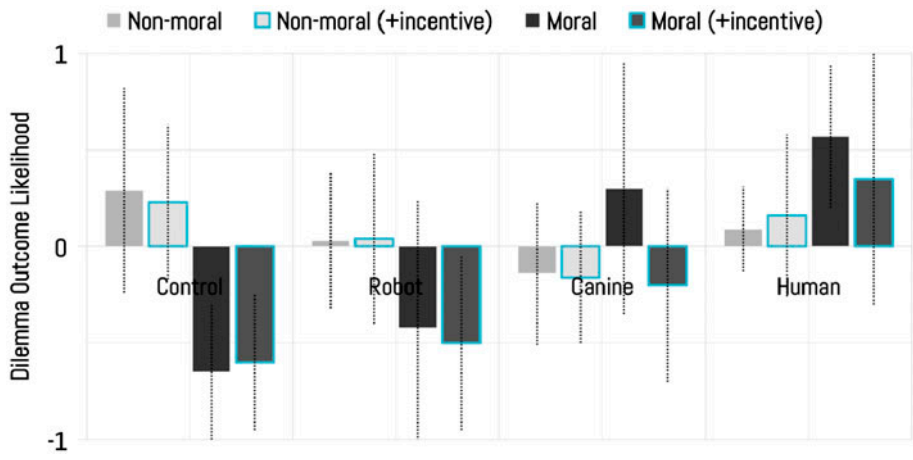

Figure 6. (Color online) Mean likelihood of behavioral outcomes across subjects (N = 10). Outcomes correspond to kill/leave (−1) and save/take (1).

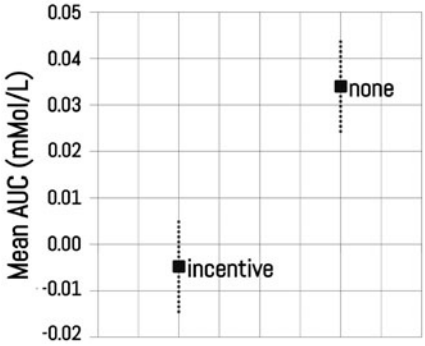

Figure 7. Main effect of incentive (left) and emotional value (right). The vertical axis depicts the condition, with the mean AUC (+/− standard deviation, unit is mmol/L) depicted along the horizontal axis.

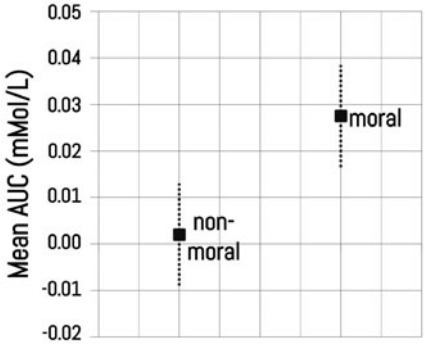

Figure 8. Interaction effect of incentive with agent category. The monetary stimulus actually reduces the change in oxyhemoglobin (bringing it closer to zero) across agent categories. Depicted is mean AUC, with units of mmol/L.

4.8. Discussion

The aim of this series of studies was to investigate the relative effects of agency, moral value, and incentive on both participants' behavior and prefrontal hemodynamics in utilitarian decision-making.

Consistent with prior work (e.g., [35]), we found main effects of both agent category on perceptions of agency and of moral value on behavior in the dilemma outcomes. However, while both canine and human agent categories were attributed significantly greater agency than the non-living categories (glass and robot), only human patients were significantly more likely to be saved than any other patient type. Moreover, contrary to our expectations, the monetary reward stimulus did not significantly influence either of the behavioral measures.

There was, however, a strongly significant main effect of the incentive on participants' hemodynamic activity in the PFC, showing in particular that the presence of the incentive substantially reduces prefrontal activity. This result suggests a diminished level of cognitive appraisal of the situation and, while there were no significant effects on participants' behaviors, it is still possible that with a greater reward (e.g., with $10 instead of $5) the behavior may indeed be influenced. Currently, the results suggest the effect of incentive (at this level of incentivization) is perhaps more readily observable using NIRS. Hence, monitoring the prefrontal cortex for reward-based neural activity may serve as a useful correlate in further experimentation.

Regarding the effects of moral value and agency manipulations on prefrontal activity, the moral manipulation showed a clear, bilateral effect while agency showed only interaction effects. Although the effects of agency on neural activity were non-significant, the observation of an interaction effect between moral value and agency

suggests a degree of importance in deciding the moral dilemma. Moreover, it is possible, given the limited presence of the agents (specifically, participants only viewed images of the agents), that by increasing the presence (such as placing participants within the same room) a main effect of agency may become more apparent from the NIRS data (e.g., [36]).

In terms of emotionally sensitive behavioral outcomes, there is a clear divide between living and nonliving agents. Specifically, living agents (human and canines) are more likely to be saved – suggesting life-like rather than human-like qualities exert more influence on decision-making in these contexts. That is, despite the human-like appearance of the Nao robot, it was no more likely to be saved than the non-human-like blown-glass object. However, it is possible that the Aldebaran Nao is not sufficiently human-like and that humanoids with greater human-likeness might show different behavioral outcomes (e.g., an android robot might be more likely to be saved than the Nao or glass objects due a highly human-like appearance).

4.9. Limitations

It is important to note that this series of investigations were conducted in a still very controlled fashion. Most importantly, the task was purely hypothetical and no interaction with physical agents took place, thus it remains to be investigated whether these results hold in realistic settings (e.g., [37]). Moreover, as precise coordinates for the placement of the NIRS probes were not used (and thus may result in alignment error), between-subjects analyses of the neural data are qualitative in nature and, moreover, comparisons with previous work in fMRI may be error-prone given the lack of confirmation as to the exact sampling region of the brain.

Additionally, it is important that the metrics used here may not have fully captured the influence of agency, emotional value, and incentive. While the monetary incentive showed no effects on the measured behavioral indices in the contexts of this series of experimentation, there may be other effects not captured by the particular behavioral metrics used here or effects that may present in other scenarios or with other agents. For instance, while we did not measure reaction times of responses to the dilemma prompts, they may further confirm whether the incentive and moral manipulations make it more or less difficult to come to a decision (despite the overall dilemma outcomes being unaffected). Moreover, the metrics used here did not assess the participant's workload or perceptions of the task difficulty, which may underlie the effects of these three factors, as each of the factors seems to increase (e.g., moral value) or decrease (e.g., monetary incentive) the difficulty of the dilemmas. Hence, to understand the mechanisms responsible for differences in prefrontal activation, further experimentation should attempt to disentangle or illuminate the relationship of workload to the factors under investigation here.

5. Conclusions

The first preliminary investigation in this series of studies supports NIRS as a potential alternative to fMRI for measuring neural processes recruited in moral-dilemma scenarios, thus allowing for a multitude of more realistic investigations on emotionally sensitive decision-making tasks. However, this study was conducted still in a very controlled fashion, as participants were instructed to minimize their physical movement (e.g., avoid scratching, stretching, etc.). It would require further investigation to validate both (1) whether NIRS would be suitable for realistic, let alone 'in the wild', investigations, and (2) whether the activity measurable in this protocol with NIRS fully corresponds to that measured using fMRI; nevertheless, it demonstrates the evaluation of decision-making processes in more realistic settings than what is currently possible with fMRI.

The second of the two suggests that all three factors play a role in decision-making. Specifically, moral value significantly increases the likelihood of living patients (humans and canines) being saved, as well as the corresponding prefrontal hemodynamic activity, whereas, surprisingly, the incentive shows only a significant influence of monetary reward on hemodynamics (reducing the observed activity) but *not* behavioral metrics. This result in particular suggests that effects of incentive are perhaps more readily observable using NIRS. Moreover, agency interacts with both factors regarding hemodynamic activity in response to human patients, with moral value further increasing the corresponding activity and with incentive further decreasing the activity.

Although replication using fMRI is necessary to confirm NIRS as a valid alternative, as well as TMS to pinpoint the prefrontal substrates of the underlying processes, this paper provides a preliminary evaluation of NIRS for studying decision-making processes in the presence of emotion and agent-based artifacts, as well as the influences of personal engagement (rather, disengagement via the $5 monetary incentive). While the technology and results are both limited in scope and applicability, we hope they may serve as a basis for further investigation of agency in emotional and non-emotional decision-making.

Notes on Contributors

Megan Strait is a graduate student in the joint PhD program in Cognitive Science at Tufts University, with research interests in human-robot interaction and brain-computer interfaces.

Matthias Scheutz is a professor of Cognitive and Computer Science and director of the Human-Robot Interaction laboratory at Tufts University. His research interests span the fields of artificial intelligence, cognitive science, philosophy, and robotics.

References

[1] Strait M, Briggs G, Scheutz M. Some correlates of agency ascription and emotional value, and their effects on decision-making. Proc Aff Comput Intell Interact. 2013;505–510.

[2] Casebeer WD. Moral cognition and its neural constituents. Nat Rev Neuroscience. 2003;4:840–846.

[3] Koenigs M, Young L, Adolphs R, Tranel D, Cushman F, Hauser M, Damasio A. Damage to the prefrontal cortex increases utilitarian moral judgements. Nature. 2007;446:908–911.

[4] Forbes CE, Grafman J. The role of the human prefrontal cortex in social cognition and moral judgment. Ann Rev Neuroscience. 2010;33:299–324.

[5] Canning C, Scheutz M. Functional Near-Infrared Spectroscopy in Human-Robot Interaction. J Human-Robot Interact. 2013;2:62–84.

[6] Epley N, Waytz A, Cacioppo JT. On seeing human: A three-factor theory of anthropomorphism. Psychol Rev. 2007;114:864–886.

[7] Gray HM, Gray K, Wegner DM. Dimensions of mind perception. Science. 2007;315:619.

[8] Ciaramelli E, Muccioli M, Ladavas E, di Pellegrino G. Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. Soc Cog Aff Neuroscience. 2007;2:84–92.

[9] Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD. An fMRI investigation of emotional engagement in moral judgment. Science. 2001;293:2105–2108.

[10] Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD. The neural bases of cognitive conflict and control in moral judgment. Neuron. 2004;44:389–400.

[11] Heekeren HR, Wartenburger I, Schmidt H, Schwintowski H, Villringer A. An fmri study of simple ethical decision-making. NeuroReport. 2003;14:1215–1219.

[12] Young L, Bechara A, Tranel D, Damasio H, Hauser M, Damasio A. Damage to the ventromedial prefrontal cortex impairs judgment of harmful intent. Neuron. 2010;65: 845–851.

[13] Urry HL, van Reekum CM, Johnstone T, Kalin NH, Thurow ME, Schaefer HS, Jackson CA, Frye CJ, Greischar LL, Alexander AL, Davidson RJ. Amygdala and ventromedial prefrontal cortex are inversely coupled during regulation of negative affect and predict the diurnal pattern of cortisol secretion among older adults. J Neurosci. 2006;26:4415–4425.

[14] Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN. Prefrontal-subcortical pathways mediating successful emotion regulation. Neuron. 2008;59:1037–1050.

[15] Cushman F, Gray K, Gaffey A, Mendes WB. Simulting murder: The aversion to harmful action. Emotion. 2012;12:2–7.

[16] Cushman F, Greene JD. Finding faults: How moral dilemmas illuminate cognitive structure. Soc Neurosci. 2012;7:269–279.

[17] Greene J, Haidt J. How (and where) does moral judgment work? TiCS. 2002;6:517–523.

[18] Moll J, Zahn R, Oliveira-Souza R, Kreeger F, Grafman J. The neural basis of human moral cognition. Nat Rev Neurosci. 2005;6:799–809.

[19] Moll J, de Oliveira-Souza R. Moral judgements, emotions and the utilitarian brain. TiCS. 2007;11:319–321.

[20] Valdesolo P, DeSteno D. Manipulations of emotional context shape moral judgment. Psychol Sci. 2006;17: 476–477.

[21] Paharia N, Kassam KS, Greene JD, Bazerman MH. Dirty work, clean hands: The moral psychology of indirect agency. Org Behav Human Dec Proc. 2009;109:134–141.

[22] Heekeren HR, Wartenburger I, Schmidt H, Prehn K, Schwintowski H-P, Villringer A. Influence of bodily harm on neural correlates of semantic and moral decision-making. NeuroImage. 2005;24:887–897.

[23] Amit E, Greene JD. You see, the ends don't justify the means: Visual imagery and moral judgment. Psychol Sci. 2012;23:861–868.

[24] Greene JD, Morelli SA, Lowenberg K, Nystrom LE, Cohen JD. Cognitive load selectively interferes with utilitarian moral judgment. Cognition. 2008;107:1144–1154.

[25] Greene JD and Paxton JM (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *PNAS*, 106.

[26] Greene JD, Cushman FA, Stewart LE, Lowenberg K, Nystrom LE, Cohen JD. Pushing moral buttons: the interaction between personal force and intention in moral judgment. Cognition. 2009;111:364–371.

[27] Forbes CE, Cox CL, Schmader T, and Ryan L.Negative stereotype activation alters interaction between neural correlates of arousal, inhibition and cognitive control. *Soc Cog Aff Neurosci* 2011.

[28] Bartneck C, Hu J. Exploring the abuse of robots. Interact Studies. 2008;9:415–433.

[29] Briggs G, Scheutz M. Investigating the effects of robotic displays of protest and distress. Soc Robot. 2012; 238–247.

[30] Henri-Bhargava A, Simioni A, Fellows LK. Ventromedial frontal lobe damage disrupted the accuracy, but not speed, of value-based preference judgments. Neuropyschologia. 2012;50:1536–1542.

[31] McClure SM, Liabson DI, Loewenstein G, Cohen JD. Separate neural systems value immediate and delayed monetary rewards. Science. 2004;306:503–507.

[32] Rushworth M, Noonan M, Boorman E, Walton M, Behrens T. Frontal cortex and reward-guided learning and decision-making. Neuron. 2011;70:1054–1069.

[33] Ye JC, Tak S, Jang KE, Jung JW, Jang JD. NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy. NeuroImage. 2009;44:428–447.

[34] Cui X, Bray S, Reiss AL. Functional near infrared spectroscopy (NIRS) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. NeuroImage. 2010;49:3039–3046.

[35] Bartneck C, Kanda T, Mubin O, Mahmud AA. Does the design of a robot influence its animacy and perceived intelligence? Inter J Soc Robot. 2009;1:195–204.

[36] Strait M, Canning C, and Scheutz M. Let me tell you! Investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality, and distance. *Proc Human-Robot Interact (HRI)* 2014.

[37] Strait M, Canning C, and Scheutz M. Limitations of NIRS-based BCI for realistic applications in human-computer interaction. *Proc Brain-Computer Inter Meeting* 2013.