# Experimental Methods for Moral Behaviour Analysis in Human-Robot Interaction

Francesco Perrone

*$2^{nd}$ Year, Progression Report*

*University of Glasgow, School of Computing Science*

**Type of study**: part-time
**Funding source**: self-funded
**Thesis submission**: 20/01/2024

**Report**

## 1. Introduction

In this follow-up progress report we outline the work done during the second year of my PhD. This work succeeds the preliminary evidence we gathered, in the first year, about the potential that experimental methodologies have as a new tool in the investigation of the autonomous moral behaviour of man-made machines, in the field of AI.

Withal, we outline two main research activities that we conducted this year, in the field of Machine Ethics (see page 2 for a definition):

a) **A deeper analysis of the literature** that suggests the emergence of the following: two related research themes in Machine Ethics, *i.e.*, Human-Machine Ethics and Computational Machine Ethics; the emergence of two distinct trends in Psychology and Philosophy, *i.e.* cognitive/affective models of moral judgments and rationalism/intuitionist approach to moral reasoning, that exert a deep influence on the research objectives and methodologies in Computational Machine Ethics;

b) **An experimental activity** about the interplay between the presence of social robots and human prosocial behaviour;

Furthermore, following the analysis in a) and evidences in b) we will argue in favour of the adoption of new research methodologies in Computational Machine Ethics that should follow recent experimental evidences in support of models of moral judgements as affect-laden intuitions (explained below). This model of moral reasoning has not yet been taken into consideration in any of the work done in Machine Ethics up to date.

The most interesting implications of such a turn for Computational Machine Ethics would arguably be the following:

1 The possibility to design experiments that quantify differences in moral attitudes through the measurable outcomes of decisions made by subjects at least in a controlled setting (*i.e.* experiments);

2 The possibility of analysing moral decisions through measuring behaviour, which in turn lends itself to the application of Social Signal Processing and Affective Computing methodologies to the investigation of moral reasoning, its analysis and automation.

On this account, for the rest of my PhD we will address the following two *research questions* which imply our intended *research statements*:

(Q1) Does the presence of social robots change the outcome of decisions made by humans?

(Q2) Do moral decision leave physical traces in terms of observable, machine detectable behavioural cues?

Q1 refers mainly to point 1, and will shows that it is possible to explore whether principles and laws underlying Moral Psychology apply to Computational Machine Ethics.

Q2 refers mainly to point 2, and will show that it is possible to apply existing social and psychological approaches for improving the investigation and validation of theories of human moral behaviour.

The rest of this report is organised as follows: Section 2 argues for a synthesis of two projects in Machine Ethics, and proposes a new organisation of the surveyed state-of-the-art, Section 3 shows experimental work aimed at addressing Q1 and point 1, Section 4 proposes a plan for the rest of my PhD and Section 5 draws some conclusions.

## 2. Multi-systems Machine Ethics

Machine Ethics is the subfield of Computer Science that develops methods and theories aimed at enabling machines to interact morally with their users in real-world scenarios. The attention that Machine Ethics has received from the scientific community has increased sharply in the past decade[1]. A central reason for this encouraging circumstance is an unprecedented interdisciplinarity: researchers in Machine Ethics are now capable of freely drawing on scientific resources and experimental data from well beyond the confines of their fields [**?** ], which can now be integrated into Artificial Intelligence (AI) technologies. Machine Ethics could be thought as a laboratory to verify and generalise, philosophical small-scale theories and thought experiments, which have heavily characterised and shape the work in this field up until now [**? ?** ].

We have shown in year one, how the approaches presented in the Machine Ethics literature could be divided into three major groups: namely foundational, procedural and psychological approaches.

Foundational work tries to define the field of Machine Ethics in terms of common objectives, methodologies, terminology and societal issues. We discussed the reason why Machine Ethics needs to define itself in these terms in year one, *i.e.*, due to a lack of common and clear objectives, consistent and rigorous terminology, to established research methodologies and standard metrics for the measurement of research performance.

Procedural approaches instead are based on the implementation of predefined rules expected to reproduce the logic behind moral decisions, as we have seen, and psychological approaches are those that intend to capture aspects of Ethics which are relevant to human-robot interactions, by means of psychological and sociological analyses, aiming to gauge the public opinion and perception towards robots in different contexts.

A deeper comparative review between the literature in Machine Ethics and allied academic fields such as, Philosophy and Moral Psychology, has revealed a deeper relationship between the three, a relationship that we used to outline, for the first time, two contrasting but interdependent research themes.

In fact, we believe that research in Machine Ethics can be classified further as being dominated by two predominant research projects. The first, which we call *Human-Machine Ethics* focuses on developing *ethics-for* humans, promoting discussions of proper and improper human behaviour concerning the utilisation of machines that implement and use AI technologies.

[1] On Google Scholar, the keyword 'Computational Morality' alone yields more than 39,000 results, while the keyword 'Machine Ethics' yields about 3,000,000 results.

The second, which we call *Computational Machine Ethics*, aims at developing *ethics-in* machines and therefore, it concerns possible ways to implement procedurally models of moral reasoning, so that machines can *function morally*, without human causal intervention (after they have been designed for a substantial portion of their behaviour). This classification is one of the main achievements we made this year, which we are expanding to target the publication of an article in the Journal of Social Robotics (top 10% of the Scimago Journal Ranking) and in the Journal of Ethics and Information Technology which we identified as the most suitable avenue on the Philosophy side.

### *2.1. Human-Machine Ethics*

Much of the pioneering work in Machine Ethics aimed at assessing the ethical implications that AI technologies might have and, applied classical moral concepts such as democracy, equality, fairness, and transparency to resolve ethical dilemmas presented to humans by the usage of more or less autonomous artificial agents.

Human-Machine Ethics can be thought as a branch of Applied Ethics[2], and its research is mostly entirely determined to the development of Ethics for human through discussions on ethical and legal questions that AI and Robotic Technologies may raise. Such questions might include liability or potentially biased in decision-making, data protection, digital rights and general ethical standards for a responsible driven utilisation of Robotics and AI technologies in human societies [**?** ]. This analysis is used to motivate several new questions facing the field of Computational Machine Ethics (discussed in Section 2.2, page 4) and to capture aspects of Ethics which are relevant to human interacting with machines, by means of psychological analyses, which we discuss next.

Commonly, psychological (but also sociological) approaches are also seen in Human-Machine Ethics are used to capture aspects of Ethics which are relevant to human-machine interactions, by means of empirical analyses. Physiologically, they investigate ethical aspects of human-machine interaction that hinge on human perception, such as the attribution of mental properties to machines [**?** ], [**?** ], [**?** ], *i.e.*, whether human perceive machines as being able to think, being able to reason, being able to have thought and emotions *etc.*, and furthermore, they investigate how individuals apply moral norms to non-human counterparts [**?** ], [**?** ], [**?** ].

Socially these approaches gauge the public opinion towards the adoption of robots and AI in human-centric domains by measuring perceptions, acceptance level, worries and reservations that people might have about robotics technologies and AI in different human-centric societal contexts [**?** **?** **?** **?** ]. They also investigates how human might express moral judgments such as, condemnation and blame, with regards to machines' actions in morally sensitive settings [**?** **?** **?** ] and combine the results to explored the possibility of defining common objectives concerning societal issues pressing the development of fully autonomous moral agents [**?** **?** **?** ].

Notably, this line of research brought about the institution of ongoing tasks forces such as the UK Commission on Artificial Intelligence, based at the Alan Turing Institute, which examine the social, ethical and legal implications of recent and future developments in AI [**?** ], and the European AI Alliance which gathers multi-stakeholders and international actors in the field to regulate the human ethical implications of AI and the use of big data for innovation [**?** ].

Lastly, Human-Machine Ethics expands its focus into the prospects of machines being able to make autonomous and morally relevant decisions in sensitive human-centric contexts [**?** ]. This analysis amis to provide design recommendations for the implementation of ethical controls suitable, for example, to constrain lethal actions, in military robotic systems, and more in general to make sure that the application of AI technologies falls within the bounds prescribed by the human laws [**?** **?** ].

[2]Applied Ethics refers to the practical application of moral considerations with respect to real-world actions such as Ethics of AI, Ethics of Nursing, Business Ethics *etc.*

Most of the psychological and sociological work we have reviewed argue against having artificial entities capable of truly autonomous moral behaviour [**?** ] on grounds that machines *must* only have meaning and significance in relation to a human components with which they collaborate and that have meaning only in relation to human beings [**?** ].

In conclusion, Human-Machine Ethics emphasises on the pervasiveness of AI technologies in modern societies and provides grounds for an intrinsic legitimacy and necessity in investigating how AI should be shaped to support the maintenance and strengthening of constitutional democracy [**?** ], health care [**?** ], and warfare [**?** ]. Human-Machine Ethics discloses various ethical problems that AI technologies give rise to, and relate them to discussion on responsible innovation and how Ethics should be carefully considered to develop technology that understands fully human social dynamics, moral norms and social behaviours [**?** ] [**?** ]. Modern AI technologies are programmed to make decisions in an autonomous way, with profound impact on our lives [**?** **?** **?** ] but they are still incapable of demonstrating any capability for moral reasoning processing [**?** **?** **?** ] . It is this necessity that brought about the legitimacy of a second research theme in Machine Ethics, which we call Computational Machine Ethics and discuss next.

### *2.2. Computational Machine Ethics and its objective*

*Moral reasoning* also known as *moral decision making*[3] is the cognitive process of choosing between 'judgements' we make in moral contexts *i.e.*, decisions we need to take on what is *"right or wrong" "good or bad" etc.*, and that we use as motive, purpose and direction for our conscious and practical behaviours.

Whether moral decisions can be made computable has long been the predominant question in Computational Machine Ethics (CME from now on). In contrast to Human-Machine Ethics (page 3), CME aims at developing *ethics-in* machines so that they can *function morally* without human causal intervention, after they have been designed for a substantial portion of their behaviour.

It should be mentioned that in the past the study of moral decision making has been a special province of Philosophy and Psychology, that investigate human functioning in moral context, making the following scenario to emerge:

- Philosophy has been largely speculative about the nature of moral reasoning and light on facts. It aimed, for the greatest part, to identify and justify the *structural* content of ethical frameworks (*e.g.* is has focused on questions such as:'is Utilitarianism right?' 'Is there any moral truth in nature?' What is right and what is wrong?' and so on).
- Psychological work on moral decision making instead, focused mostly on empirical and experimental activities, but have been light on theory [**?** ]. It aimed at investigating the psychological representation associated with those ethical frameworks that Philosophy provides.

We cannot exclude the possibility that the gaps between the two has given a spurious picture of what moral reasoning might be, and affected in turn the research in CME.

As we have discussed in year one, the great majority of approaches in CME focused on the implementation of predefined rules expected to reproduce the logic behind moral decisions (for example [**?** ], [**?** ], [**?** ], [**?** ]). In general, these approaches start from a moral thought experiment, *i.e.*, a hypothetical scenario that involves a challenging moral decision. A typical example is the 'trolley problem' [**?** ] that, in general terms, requires people to decide whether it is more morally acceptable to kill one individual or a group of several individuals in the case that one of these options *must* be realised (the literature proposes several variants of

[3]We will use the two terms interchangeably throughout this report although there are subtle differences between the two definitions on which we will expand in future work.

this base version see for example [**?** ]). Most research in CME was built on procedural approaches inspired by philosophical tradition probably because the psychological elements of moral decision making are more difficult to implement [**?** ], but researchers did not take into account that even in Philosophy thought experiments are not supposed to represent reality, but simply to stimulate discussion about the way we think about moral issues [**?** ].

This problem has been capture by the AI community since the foundation of Machine Ethics [**?** ] and have developed into more recent approaches that try to follow the same overall scheme but replacing rules and logic-based methodologies with Neural Networks and Machine Learning (se for example [**?** ], [**?** ], [**?** ]). None of these approaches result into methodologies that generalise easily, and programs that scale [**? ? ?** ].

### *2.3. The 'intuitionist' turn in Computational Machine Ethics*

Moral Psychology has proposed a division between *emotional* versus *cognitive* moral judgments, and between *automatic* versus *controlled* moral judgments [**?** ]. Furthermore, recent experimental data seems to confirm that emotions can motivate and impel us to act morally, even without consciously stepping into rational thinking, thoughts and norms endorsement, or the standard patterns of deductive and inductive argumentation and inference [**?** ].

This line of research has shown further that the *noticeable*[4] occurrence of emotions prior to more conventional (*i.e.* rational) moral evaluations, influences human moral responses [**?** ], suggesting therefore that the link between emotions and moral judgements is not merely correlational or epiphenomenal [**?** ] but moral emotions and moral reasoning work together in the creation of human morality [**?** ]. As a consequence of this evidence, Moral Psychology, once dominated by *rationalist* models of moral reasoning[5] adopted a rationalist model of morality which refers to the view that humans grasp moral truth not by process of ratiocination and reflection, but rather, by cognitive processes more akin to perception (for an excellent introduction refer to [**?** ]), in which one just sees without argument that their evaluations are, and must be true [**?** ], appear to give rise to models of moral judgements as automatic, rapid, and emotionally forceful intuitions.

Intuitionist approaches have been given recognition from most of the modern work in Moral Psychology, promoted by new findings in evolutionary psychology and primatology [**?** ] that began to point out that:

a) The origins of human morality might be in a set of emotions that make individuals care about the wellbeing of others (a new stand called the *Affective Revolution* which took place in the 1990s [**?** ] in [**?** ])

b) The study of moral reasoning should follow the new focus on *automaticity*, that in Psychology advocates the mind's ability to solve problems unconsciously and automatically.

From this a comprehensive model, the *social intuitionist model* [**?** ] brought together research on automaticity with findings in neuroscience [**?** ] suggesting that moral judgment is much like aesthetic judgment, where we have an instant feeling of approval or disapproval towards situations we perceive in moral contexts. These feelings are best thought of as affect-laden intuitions:

> "[...] they appear suddenly and effortlessly in consciousness, with an affective valence (good or bad), but without any feeling of having gone through steps of searching, weighing evidence, or inferring a conclusion[...] [**?** ] "

[4]Here with the term 'noticeable' we mean that the raising of such emotions is a fact with precise physical traces that can be detected and observed experimentally.

[5]mostly during the cognitive revolution of the 1950s and 1960s with Kohlberg [**?** ], Piaget [**?** ], and Turiel [**?** ])

Most importantly, works in Neuropsychology has revealed that moral and non-moral emotions can be *detected*, *observed experimentally*, and *distinguished* at theoretical level without any cognitive theory describing what emotions are in general and in moral contexts [**?** **?** ].

There exist on this topic a large number of theories (see for example [**?** ] or [**?** ] for good introductions), and we will not survey this debate here. However, what it seems to emerge from this collective work is a common agreement on two points:

a) There are emotions that promote morally good behaviour by orienting us towards other people needs;

b) Emotions have component features such as 'eliciting event', 'facial expression', 'physiological changes', 'phenomenological experience', and 'action tendency' (or *motivation*) that can be used to analyse and classify them.

Haidt in [**?** ] used two of these component features *i.e.*, *elicitors* and *action tendencies* in a groundbreaking article, showing that it is possible to creates a bi-dimensional space in which moral and non-moral emotions can be plotted and distinguished. By the same token, Haidt in [**?** ] identified four sets of moral emotions which are arranged in two large families:

- *other-condemning*, which includes: contempt, anger, and disgust, indignation, loathing,
- *self-conscious*, which includes: shame embarrassment and guilt.

and and two small families:

- *other-suffering*, which includes compassion, indignation, loathing.
- *other-praising:*, which includes: gratitude and elevation

Pinning out these emotions gives us an invaluable tool to potentially detect and observe experimentally physical traces of moral reasoning which can bring about new exciting prospective for both CME and Philosophy, via the application of Social Signal Processing and Affective Computing methodologies for the understanding, modelling and automation of human moral decision making.

In fact, Affecting Computing and Social Signal Processing [**?** ], [**?** ], [**?** ] can provide the necessary tools to sense and understand human social signals and potentially capture the use of emotions in moral contexts as we have seen for other domains of human society [**?** ] (some of the many interesting application can be seen in [**?** **?** **?** **?** **?** **?** **?** **?** ]).

In fact, since moral judgments appear to arise suddenly and effortlessly as automatic affective reactions,[**?** ], [**?** ], [**?** ], [**?** , chapters 4 and 6] induced by emotions occurring prior to conventional moral evaluations that influence human moral responses [**?** ], it might be possible to design experiments that a) quantify differences in moral attitudes through measurable outcomes of decisions made in a controlled setting and b) analyse moral decisions through measuring the component features of its emotions via the application of Social Signal Processing and Affective Computing methodologies.

Therefore, the question we ask is whether moral decision leave any physical traces in terms of observable, machine detectable behavioural cues, in the context of interplay between the presence of social robots and human prosocial behaviour. This question refers to Q1 in section 1 page 1 (Does the presence of social robots change the outcome of decisions made by humans?), and can be said to be, theoretically, a general case of Q2 (ibid). We believe that the possibility to design experiments that quantify differences in moral attitudes through the measurable outcomes of decisions made by subjects in, at least, a controlled setting (*i.e.* experiments) would suggest that the analysing moral decisions through measuring behaviour in CME is a general fact, which in turn lends itself to the application of Social Signal Processing and Affective Computing methodologies to the investigation of moral reasoning, its analysis and automation.

Psychologists have adopted two converging strategies to understand decision making [**?** ]:

1. statistical analysis of multiple decisions involving complex tasks;
2. experimental manipulation of simple decisions, looking at the elements that recur within these decisions.

Taking into consideration our two research statements (section 1 page 1 and above), and in view of what we have discussed thus far, strategy 2 above seems the best candidate to our objectives. To such a purpose, working closely with my supervisor, we have designed an experiment in which the participants are asked to fill in psychometric questionnaires and then they are given the opportunity to donate part of the compensation they receive for carrying out the experiment to a charity, to investigating whether there is an association between the presence of robots and the outcome of moral decisions made by humans, which is described next.

## 3. Experiment

The study we present here aimed at investigating whether associations between the presence of robots and the outcomes of 'moral decisions' made by humans exist. In particular, we hypothesised that changes in the outcomes of moral decisions made by individuals in experimental settings we designed might be observed if a robot were to be place in their environment during a behavioural tasks.

### *3.1. Design*

To this extent we designed an experiment to test if the presence of a robot (NAO), programmed to appear *passively watching* [**?** ] participants working towards a decoy experimental task, could have been associated with differences in their charitable giving behaviour *i.e.*: *whether donations made to a charity tended to be higher/lower when participants shared the room with a robot than when they were alone, for the duration of the experiment.*

We used mixed methodologies consisting of psychometric tests administered to determine a) participants' personality characteristics and b) their brain types, this together with the development of a behavioural task, to assess participants moral decision making.

The behavioural task we designed to test whether the presence of a robot could be associated with differences in the outcome of individual moral decision processes, was adapted from the *watching eye paradigm*, an experimental model in which adults tend to displayed greater prosocial behaviour in the presence of observation cues.

### *3.2. The watching eye effect*

The watching eye paradigm is a fine-tuned experimental setup that allows the investigation of a particular class of phenomena that results from the effects of *observation cues* on human cognition, known as *watching-eye effect*.

The watching-eye effect suggests that:

> "[...] just feeling watched may be enough to make us modify our actions independent of deliberative, explicit, conscious, evaluation [...] [**?** ] ".

In other words, the perception of direct gaze causes in individuals a sudden heightened processing of incoming stimuli in relation to the self, which leads to the enhancement of self-awareness and memory, together with the activation of positive appraisal of others and prosocial behaviour [**?** ]. Particularly relevant to us was that experimental work has shown how individuals (strategically) modify their behaviour when being observed by others towards acting more prosocially [**?** ], and conform to local norms, to gain

valuable reputation and avoid the possibility of sanctions by potential observers [**?** **?** ]. Will this apply when the observer is a robot?

A recent experiment [**?** ] found that even subtle cues such as stylised eyespots on a computer background, increased the amount of money that was offered in a dictator game, as well as increased the odds of donating something rather than nothing to the other players in the same settings [**?** ]. Similarly, an image of a pair of eyes increased money contributions to an honesty box used to collect money for drinks in a university lounge [**?** ], and a simple intervention of displaying signs featuring images of watching eyes and a verbal message about being watched was associated with a large reduction of bicycle thefts [**?** ].

It should be noticed that experimental work on the watching eye effect confirmed that observation cues do not need to be explicit. It has been shown in [**?** ] that priming in individuals the presence of supernatural, omnipresent entities can activated implicitly increased prosocial behaviour, even in situation when the behaviour was anonymous and directed toward strangers. Whether explicit or not, cues of being watch seem to be sufficient to affect different facets of human prosocial behaviours [**?** ], including *prosociality* in those situations where the behaviour cannot directly be traced back to the actors, by any potential observer [**?** ].

### *3.3. Methods*

We advertised the experiment as a study to gather data on human personality traits for a representative sample of population, requesting the following three questionnaires be filled in. The Empathizing and the Systemizing Quotients (ESQ) [**?** ]: a diagnostic questionnaire designed to measure the expression of neurological types in an individual by his or her own subjective self-assessment; and the Big-Five Inventory 10 (BFI-10) [**?** ]: a shorter version of the Big Five Inventory which assess people personality traits. Participants ($n = 73$) were drawn from two sources, which correspond to two populations sub-groups described below.

The *Computing Group* (CS) comprised 30 adults, 23 males and 7 females, taken from the undergraduate students population studying Computing Science at University of Glasgow. CS had a mean age of $x = 20.7$ with $s.d. = 4.6$. Participants needed to meet following two inclusion criteria to be hired for the experiment: being 17 years of age or above, and being British (passports where requested and checked on the day of the experiment). Although the sex ratio might appear to be disproportionately affected by a particularly high number of males ($\approx$ 3:1,m:f), this ratio was in line with the UCAS data on STEM undergraduate students in UK [**?** ] at the time of recruitment. When we looked at the data published by UCAS, on a total of 24,090 students studying Computing Sciences in UK, 19% were female, and 81% of the students were male. **The difference between the ratio 'male:female' nationwide and that of our sample was not statistically significant** ($p = 0.55$ according to a $\chi^2$-test on sex ratio) with 23% of the students drawn being female, and 77% male.

The *Psychology Group* (PS) comprised 43 adults, 15 males and 28 females recruited through a subject-pool database provided by the School of Psychology of the University of Glasgow[6]. PS had mean age of $x = 24.9$ and $s.d. = 9.15$, From a total of 361 booking requests, we randomly selected those individuals who fulfilled the two same inclusion criteria define for the CS group (above), and if a booking was made by students, we only selected those studying a degree other than Computing Science.

The two groups combined (*i.e.* the population sample which comprised of *CS + PS*) had mean age $x = 23.53$ and $s.d. = 7.23$ with a sex ratio of approximately 1:1 (m:f) in line to that found in similar experimental samples (see for example [**?** ]). Each participant in the population sample was randomly assigned to one

[6]This is a database of volunteers set up at the School of Psychology and Institute of Neuroscience and Psychology at the University of Glasgow, which allows members of the public to register their details and take part in a wide range of studies. The database gathers individuals from the general public in UK, and represents a wide mixture of cultural backgrounds and occupations including clerical and manual workers, professionals, undergraduate and postgraduate students, see https://participants.psy.gla.ac.uk.

of two experimental conditions we designed (in line with the watching eye paradigm). The two conditions were *Control* and *Robot* discussed next.

In *Robot*, a humanoid robot (NAO) is placed in the setting (experiment room) in autonomous life [**?** ], a configuration provided by the robot's manufacturer that makes NAO look alive through the simulation of breathing. Furthermore, the robot can track people with its head, thus conveying the impression of *observing them* (such a process is activated only if a person establishes eye contact with the robot [**?** ]).

In *Control*, the setting is the same as in Robot, but without NAO. A pictorial representation of the two settings is given in figure below. Both conditions are set in the same room hence, with exception made for the presence or absence of NAO, all participants sit the experiment in the same environment.

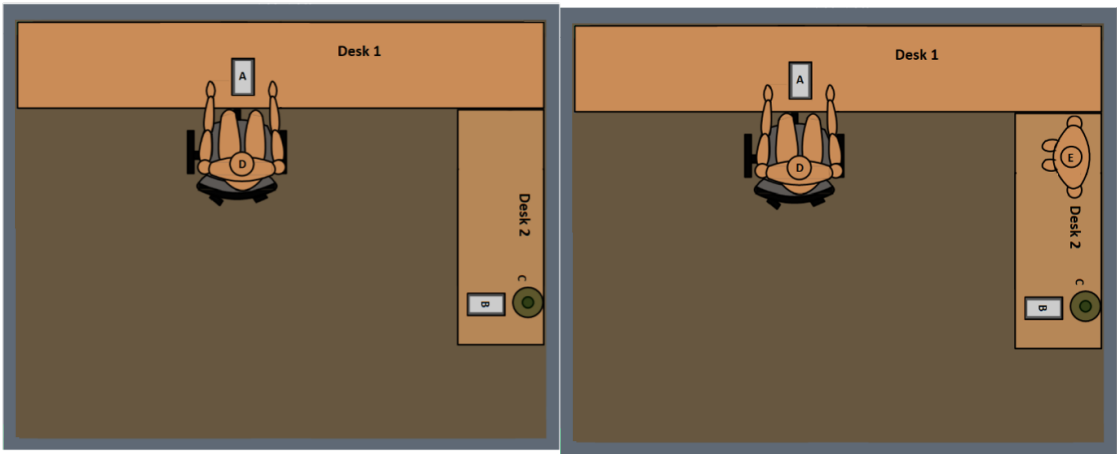


Figure 1: A pictorial representation of the experiment settings for both the "control" (left) and "robot" (right) conditions. In the picture: A) questionnaires, B) consent form and payment, C) the charity box, D) human subject, E) robot (NAO).

The *Robot* condition comprised 38 adults randomly selected from the population sample, 20 males and 18 females with mean age of $x = 25.4$ and $s.d. = 8.3$. The *Control* condition comprised 35 adults randomly selected from the population sample, 18 males and 17 females with mean age of $x = 20.7$ and $s.d. = 4.1$.

| **Sub-groups** | **Control** | **Robot** | **Total** | **%** |
|---|---|---|---|---|
| **Male CS** | 12 | 11 | 23 | 31.5% |
| **Male PS** | 6 | 9 | 15 | 20.5% |
| **Female CS** | 3 | 4 | 7 | 9.5% |
| **Female PS** | 14 | 14 | 28 | 38.3% |
| **Total** | 35 | 38 | 30 | 100% |

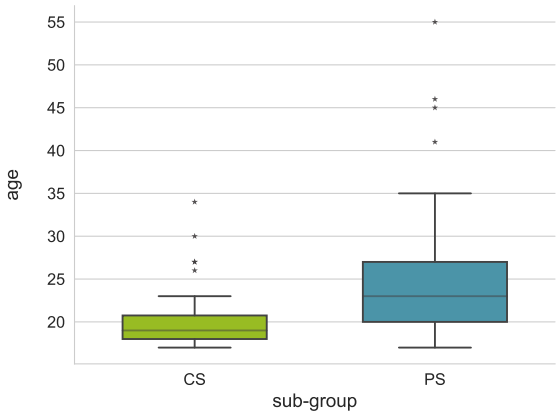


Figure 2: Population sample is given for each experimental conditions - *Control* and *Robot*- arranged by sub-groups: *CS* and *PS*. In CS most students were aged between 18 and 21 with median $m = 19$, in PS participants had a greater age variability and larger outliers. Participants in PS were aged between 20 and 27 with with median $m = 23$.

*3.4. Protocol*

On the day of the experiment, each participant had individual access to the experiment room (see Figure 1) for the duration of one hour. Alone in the room, they were requested to answer personality questionnaires, collect a compensation of £10 for their participation (made of 10 £1 coins), and decide whether to donate part of it to a charity before living the room (the moral decision case).

To each participant we allocated a date and time to sit the experiment via email, hence no frontal contact with the participants was made before their allocated slot. On the day of the experiment, the participant would have arrived at the foyer of the School of Computing Science where a laboratory assistant would have been waiting for them. Then, the following scripted steps were always performed:

- The participant is welcomed, the passport is checked. Conversation is kept to a bare minimum.
- The participant is walked to the experiment room via a fixed route.
- Upon arrival, the experiment setting is describe to the participant from outside the room (the room is kept closed for the participant to open). The description of the setting is limited to how to find the experiment instructions.
- The participant is asked to open the door when ready, enter the room and start the experiment by reading the instructions.
- Once the questionnaires have been filled, the participant would sign a consent form and gather the compensation.
- The experiment finishes once the participant has left the room, closed the door and called the laboratory assistant to be accompanied to the closest exit.
- Once the participant has left the building, we gather the filled questionnaires, consent form, and counted any money that were left, in any, to the charity by opening the charity box, collect the money if any, and closes the box back.
- We then prepared the room for the next participant, scan the filled questionnaires, and takes note of the amount of money donated (including any 0-case).

*Decision case:.* Before leaving the room, each participant had to decide whether to make a donation using a green charity box placed in the room (desk 2 in Figure 1, page 9).

**It is important to notice that**: participants did not know the experiment task before entering alone the room and getting access to the instructions, neither have they been in the experiment room before. This means that participants from both groups had no indication that a robot could have been in the room, nor did they know before experiment took place that they would have an opportunity to make a donation.

*3.5. Results*

The dataset we created is made of 73 decision cases, each describing the amount of money that participants *decided* to donated to charity. Overall, the amount donated by participants in the Control condition (£66) higher than the amount donated by participants under in the Robot condition (£44.35).

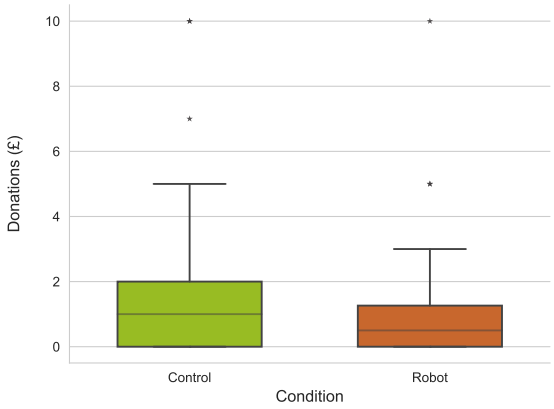


Figure 3: A comparison of the donations made in the two experimental conditions reveals a difference in the amount donated between Control and Robot.

This difference in donations observed was statistically significant ($p = 0.01$ after FDR, according to a $\chi^2$-test on donations observed) **suggesting an association between the presence (or absence) of the robot and differences in moral decisions made by participants** (measured in terms of donation made to charity).

### *3.6. Confounding factors*

We tested for other possible confounding factors that might be responsible for a spurious association between the donation behaviour observed and the presence or absence of a robot in the participants' settings. In particular we tested four confounding factors: a) differences in the donations observed across age, gender and educational background; b) differences in the donations observed across brain types; c) differences in the donations observed across personality traits, d) Differences in the donations observed across educational background

#### *3.6.1. Age, gender and educational background*

Bakker [**?** ] reports that a large academic literature on charitable giving seems to suggest that the relationship between age and charitable donations is a positive one, although there are no work up to now that clarify the exact age at which the age gradient becomes weaker. However, studies with a large proportion of respondents of 'similar age', and studies on 'specific populations' seem to suggest a negative relationship between age and the likelihood of giving to charity, although findings seem to vary from study to study. In particular, what type of relationship is found between gender and giving seem to depend on the other variables included in the experimental settings: the more socioeconomic variables, such as income and educational level, are included in the models examining charitable giving, the smaller the reported gender differences in giving are.

Therefore, being our population sample quite specific, with very sparse range of socioeconomic variables, care should be therefore taken to make any assumption regarding the role that age played in the we donation observed [**?** ]. Furthermore, Bekkers reports in [**?** ] that gender is also relevant in the majority of studies on charitable giving analysed. More interestingly women have on average stronger prosocial values than men[7], including concern and responsibility for the wellbeing of others. Hence, we assumed that both gender and age across the two experimental conditions might provide an alternative explanation for the difference in the donation behaviour observed.

[7] A fact that could be connected to the E-S theory which we introduce in in the next section.

However, the difference along these factors were not statistically significant ($p = 0.85$ after FDR, according to a $\chi^2$ test on gender, $p = 0.17$ after FDR, according to a $t$-test on age, and $p = 0.83$ after FDR, according to a $\chi^2$ test on educational background *i.e.*, across CS and PS) and, therefore, **they could not be used as an alternative explanation for the donation observed.**

*3.6.2. Differences in the donations observed across brain types*

According to Baron-Cohen each person - whether male or female - has a particular *brain type*. Gender cannot tell us what brain type a person is, the central claim of Baron-Cohen's theory is in fact that all we can say at best is that *on average*, more males than females have a brain of a precise type (type S), and more females than males have a brain of another kind (type E) (for more details refer to [**?** ]). There are three common brain types according to how strong one person's *empathising* and *systemising* skills are. The *routes to morality* we take is a result of the ratio between the two independent psychological processes *i.e.*, systemizing and empathizing, that our brain is made of [**?** ], since both cognitive and emotional processes play a crucial but *mutually competitive* roles in motivating individuals to act prosocially [**?** ][**?** ] [**?** ]. In particular:

*Systemizing* is defined as a person's capability to analyse or construct systems when experiencing life. When we 'systemize' we are trying to identify regularities and rules that govern our environment, in order to predict how it will behave [**?** ].

*Empathising* is define as a person's capacity to identify another person's emotions and thoughts, and to respond to these with an appropriate emotion. When we empathize we are having an emotional reaction, an affect-based intuition about how people are feeling, and how to treat people with care and sensitivity, that compels us to act.

Empathy-based morality leads seemingly selfless acts of altruism impelled by processing stimuli of a certain kind (e.g. a child eyes filled with terror) suddenly and effortlessly in consciousness, but without any feeling of having gone through steps of weighing evidence. A system-based moral response relies on culturally derived or logically inferred rules to decide how one ought to act in accordance of good behaviour. People stronger on empathic skills have a "female brain type" (type E), individuals stronger in systemising have a "male brain type" (type S), individuals equally strong in their systemising and empathising are called "balanced brain" (type B) [**?** ].

The SEQ can test which brain type (E, S, or B) individuals are. Even though we are not interested in assessing the distribution of the individual brain types among the experimental settings, **differences of empathising and systemising might potentially explain the difference observed in the donations made**.

A t-test was used to examine the significance of the difference between the means of the two samples, which suggested that there are no statistically significant differences in in empathic and systematic brain types across the two experimental conditions ($p = 0.17$ for EQ and $p = 0.18$ for SQ after FDR, according to a $t$-test on EQ and SQ observed) and, therefore, **brain type cannot be use as an alternative explanation for the difference in donations we have observed**.

*3.6.3. Differences in the donations observed across personality traits*

Personality is an enduring construct that comprises an individual's unique adjustment to life [**?** ], and therefore the resulting pattern/s of habitual behaviours, cognitions, and emotional responses an individual might have [**?** ]. Personality traits have been used to predict behaviour defining a trait as "that which defines what a person will do when faced with a defined situation" [**?** ]. Differences in individual personality traits might therefore provided an alternative explanation for the donation behaviour we have observed. For this reason, personality scores have been collected for each subject by dispensing to them the BFI-10 which includes the following five broad dimensions that describe their personality traits (for a complete description of these traits see [**?** ] and [**?** ]):

Openness (O): Artistic, Curious, Imaginative, etc. People scoring low on Openness tend to be conventional in behaviour and conservative in outlook. Conscientiousness (C): Efficient, organised, thorough, etc. The conscientious person is purposeful, strong-willed and determined. Low scorers in C predicts a less exact application of moral principles. Extraversion (E): includes traits such as sociability, assertiveness, activity and talkativeness. Extraverts are energetic and optimistic. E is characterised by positive feelings and experiences and is therefore seen as a positive affect. Agreeableness (A): appreciative, kind, generous, etc. An agreeable person is fundamentally altruistic, sympathetic to others and eager to help them. Neuroticism (N):anxious, self-pitying, tense, etc. A high Neuroticism score indicates that a person is prone to having irrational ideas, being less able to control impulses, and coping poorly with stress.

The BFI-10 is a taxonomy for personality traits grouping developed from the 1990s onwards in psychological trait theory [**?** ] and supported by strong experimental evidence showing how the same traits appear with regularity across a wide spectrum of situations and cultures, hence corresponding to psychologically salient phenomena that can be measured and assessed [**?** ]. The BFI-10 is today one of the most influential paradigm in personality research, widely accepted in the computing community as well [**?** ]. To the best of our knowledge, no other theories were ever adopted in computing oriented research.

Although testing for differences of C and A would seem most relevant in our experiment, here we made no assumption on what trait/s could be more important or relevant to the altruistic behaviour we wanted to measure. Hence a t-test was used to examine the significance of the difference between the traits means of the two samples, which suggested that there were no statistically significant differences in personality traits across the two experimental conditions ($p = 0.11$ for O, $p = 0.14$ for C, $p = 0.24$ for E, $p = 0.28$ for A and $p = 0.25$ for N after FDR, according to a $t$-test on OCEAN observed) and, therefore, **personality traits cannot be use as an alternative explanation for the difference in donations behaviour we have observed**.

## 4. Discussions and research plans

### *4.1. Experimental results*

The lack of statistically significant effects in terms of age, gender, educational background, brain types, and personality traits suggests that **the presence (or absence) of the robot in the experimental settings is the best available explanation of the donation differences observed across the two experimental conditions**, at least in our sample.

This result seems to support our hypothesis that the presence of social robots can affect the outcome of moral decisions made by humans, which answers positively the first research statement, *i.e.* Q1 in Section 1, page 1. By the same token, these findings suggest that it might be possible to design experiments that quantify differences in moral attitudes through the measurable outcomes of decisions made by subjects at least in a controlled settings which provides a positive answer to Point 1 in Section 1, page 1.

On the other hand, this study has been unable to demonstrate that the watching-eye effect applies in our experiment in the same way as it does in other similar settings (see Section 3.2 page 7). Our findings are contrary the previous studies we surveyed, which have suggested that the perception of direct gaze and subtle cues such as stylised eyespots on a computer background, increased the amount of money that was offered in a dictator game, as well as increased the odds of donating something rather than nothing to the other players in the same settings [**? ? ? ? ? ? ? ?** ]. **To the best of our knowledge, our result has not previously been described**.

However, this result may be explained with reference to some other works in the literature according to which greater prosociality is observed when subjects are exposed to eyes (or eye-like images), compared to other non-social objects (e.g., flowers or geometric shapes) [**?** ]. This could perhaps give an explanation, although it would be highly inconsistent with NAO's anthropomorphism and intended design.

In fact, NAO is designed to be a social object. NAO is well establish social robotics platform used in different fields including care in autism spectrum disorders, which is designed to expand social and communication skills in subjects interacting with it [**?** ], or even as home-based healthcare robot to help older adults with mild cognitive impairment (MCI) or early dementia [**?** ] to enhance social inclusion. Furthermore, it has been observed that NAO is capable of eliciting heightened processing of prosocial behaviour in a real-life HRI [**?** ]. Additional uncertainty arises from the presence of a poster above the charity box which we used to advertise the donations in the experiment settings. This poster depicts a child affected by cleft lip perhaps reinforcing the conclusions in [**?** ] which, recall, suggests greater pro-sociality when subjects are exposed or eye-like images such us the child in our poster, compared to other non-social objects, like NAO might be.

There might be other possible explanations in the characteristics of the sample we have collected, and/or in other properties of the experimental settings which we intend to explore further. Therefore, further research work is needed including additional data analysis to establish whether a *directionlity* of the effect observed in the donation behaviour can be predicted. In other words, further work is required to establish what characteristics of the experiment can be used, if any, to predict whether or not subjects will donate money to charity.

In conclusion, the present result is significant in at least two major respects. It confirms that social robots might be associated with changes in the outcome of moral decisions made by humans (Q1) and suggests that it might be possible to design experiments that quantify differences in moral attitudes through the measurable outcomes of decisions made by subjects at least in a controlled setting (point 1), opening up the possibility for Computational Machine Ethics to analyse moral decisions via the application of Social Signal Processing and Affective Computing methodologies.

Therefore, we have set the following objectives for the period up to March 2022 (point 1), and up until the viva (points 2 and 3) with regards to the experimental results obtained:

1. To target a conference paper in the coming IEEE RO-MAN 2022 (submission deadline in march 15, 2022).
2. To target the publications of these results with an article in the Journal of Social Robotics (top 10% of the Scimago Journal Ranking)

3 Further experimental research to investigate what properties of the experimental settings and/or population sample might be associated with differences in the donation behaviour observed.

Overall, risks are reasonably low for points 1 and 2, given that the content of the article has been largely drafted already. Some risks might be involved with point 3 since variations of the experiment will possibly be considered, in particular more interactivity on the robot side, to test and verify the potential of the moral psychology and its social intuitionism approach as a new theoretical platform to study the role of social cues in modelling moral responses automatically, and given the ongoing COVID restrictions on social distancing.

### *4.2. Computational Machine Ethics: new prospectives*

One of the more significant findings to emerge from the comparative review we did this year, is the emergence of two related research themes: Human-Machine Ethics and Computational Machine Ethics. This new classification helped to outline more precisely two different objectives in this field and provide evidences in favour of analysing moral reasoning through measuring behaviour and processing of emotions.

These new classification warrants further investigation within a larger theoretical span. In fact, models of moral reasoning can be distinguished further into those which concentrate on *what it is right*, and others that seem to give priority to a definition of *what it is good*. These are two well distinct philosophical concepts

(Figure 4) that define another division in CME, *i.e.*, the division between *deontological* and *teleological* theories of moral judgements, which in turn have deeply determined different approaches in the research methodology in the field. This is another important issue for future research.

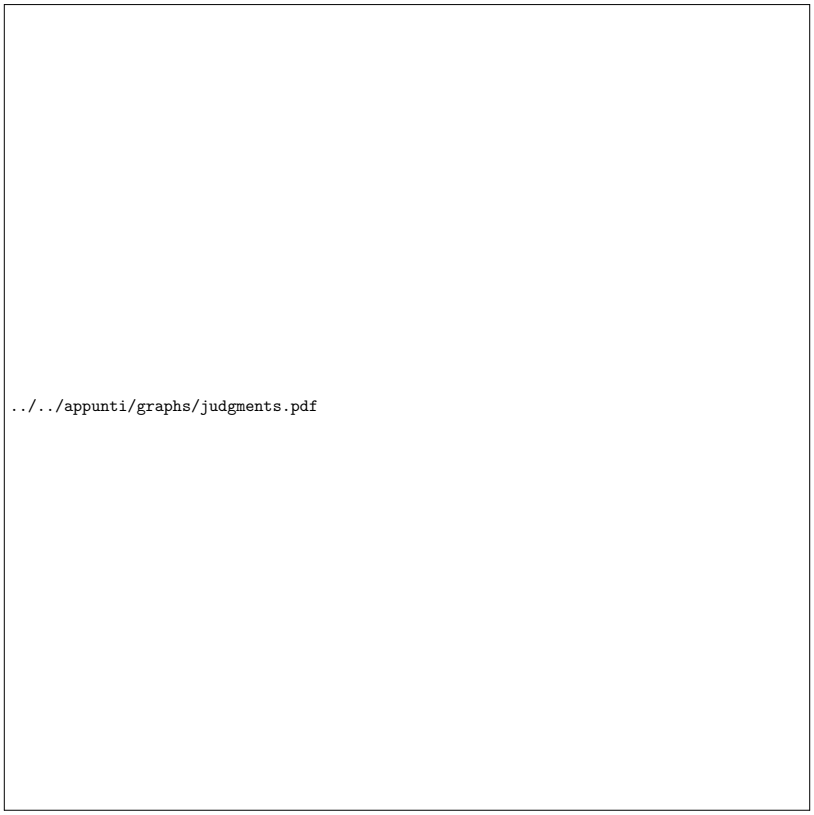


Figure 4: This distinction presuppose a sufficient prior understanding of the relevant uses of *is* and *ought* (or *should*) which will not discuss in details here. It is important to notice that, the presence of such marks as is neither a sufficient nor necessary criterion for the distinction we make, due to the striking variability of the relevant uses of the two words in every day language. For example, the sentence 'copper should be a metal' is not intended to be normative, and 'murder is evil' is not meant to be factual. Some philosophical theories claim that moral judgements lack of some desirable properties that factual statements have such as *objectivity* or *truth-apt*.

In any case, this classification is one of the main achievements we made this year, which we are expanding to target a publication of another article in the Journal of Social Robotics (top 10% of the Scimago Journal Ranking) or in the Journal of Ethics and Information Technology.

In conclusion, we provided solid experimental evidences which seem to confirms that social robots can affect the outcome of moral decisions made by humans. We have also seen that the link between emotions and moral judgements is not merely correlational or epiphenomenal [**?** ] but moral emotions and moral reasoning work together in the creation of human morality [**?** ]. Furthermore, we have seen that it is possible to pin out emotions arising in moral contexts giving us an invaluable tool to potentially detect and observe experimentally physical traces of moral reasoning. This refers to question Q2, *i.e.* whether it is possible to analyse moral decisions through measuring behaviour via the application of Social Signal Processing and Affective Computing methodologies, which will be the last research objective for the remaining time of my PhD.

## 5. Conclusion

This report has first provided a new system of classification of the state-of-the-art in Machine Ethics and argued in favour of the adoption of new research methodologies that should follow a model of moral judgements as affect-laden intuitions. The most interesting implications of such a turn for Machine Ethics,

would arguably be the possibility of analysing moral decisions through measuring behavioural cues via the application of Social Signal Processing and Affective Computing methodologies, to the investigation of moral reasoning, its analysis and automation.

In favour of this position, we have provided experimental evidences showing that such an avenue is suitable confirming that experiments like the one presented in Section 3 can successfully quantify differences in moral attitudes through measurable outcomes of decisions made in a controlled setting. Finally, we discussed these results and outlined the main research activities left until the next viva.

**References**