The title

Francesco Perrone

Glasgow, UK

University of Glasgow, School of Computing Science

Contents

1. Forewords

2. Introduction

The research presented here examines the role of experimental methodologies as a new tool for investigating prosocial behaviour and moral deliberation in the field of Machine Ethics. These methodologies allow us to study how such behaviors manifest under the subtle yet controlled influence of robots coexisting with humans. In particular, we describe how the mere presence of a non-interactive machine is shown to subtly shape ethical deliberation and prosocial behavior through perceived observation under controlled experimental settings.

Withal, we outline two main research activities that we conducted, in the field of Machine Ethics:

a) **A comparative analysis of the literature** that suggests the emergence of the following: two related, but distinct, research themes in Machine Ethics which we call Human-Machine and 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 deliberation, its analysis and automation.

On this account, the following two questions were addressed during the course of this research, forming the basis of the 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 remainder of this work is organized as follows.

3. Section

This thesis investigates moral reasoning as it manifests under the subtle yet controlled influence of human-robot coexistence, an experimental terrain

where ethical principles encounter observable behavior, under robotic observation. While it does not directly engage with the philosophical history or psychological foundations of morality, it draws upon both to establish the conceptual framework and terminology necessary to understand how machines might influence human ethical decision-making through a deliberately non-interactive experimental setting that invites participants to confront ethical scenarios under robotic observation.

Central to this investigation are precise questions concerning the influence of human-robot interactions on moral reasoning: How do autonomous systems shape ethical decision-making in humans? What are the mechanisms through which robots alter perceptions of what is morally right or wrong? How can experimental methodologies illuminate these processes with empirical clarity? Addressing these questions requires a robust foundation in defining moral reasoning, first through a philosophical lens that considers consequentialist, deontological, and virtue ethics traditions, and then through the perspective of moral psychology, which examines the cognitive and emotional processes underlying ethical judgments. These frameworks collectively enable a systematic exploration of how human-robot interactions inform and reshape our understanding of morality.

Moral reasoning, as a species of practical reasoning, is the deliberative process directed towards deciding what to do, ultimately culminating in a judgment and, when successful, issuing in an intention. In this context, moral judgment represents the evaluative conclusion of reasoning, bridging deliberation and action. This thesis adopts this dual perspective, integrating philosophical frameworks such as deontology and virtue ethics with psychological theories of intuitive and deliberative reasoning, to examine how autonomous systems influence these processes in human decision-making.

While moral reasoning can be undertaken on another's behalf, it is paradigmatically an agent's first-personal (individual or collective) practical reasoning about what, morally, they ought to do. Philosophical examination of moral

reasoning faces both distinctive puzzles – about how we recognize moral considerations and cope with conflicts among them and about how they move us to act – and distinctive opportunities for gleaning insight about what we ought to do from how we reason about what we ought to do.

We can use the map shown in Figure 1 to captures foundational distinctions in the realm of judgments, particularly as they relate to moral philosophy and psychology. The map reflects central philosophical inquiries: What is true (factual)? What is right or wrong (moral)? What ought to be done (normative)?

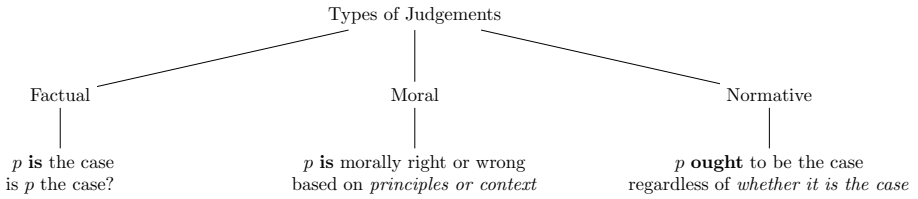

Figure 1: Diagram of Types of Judgements.

Paragraph.

Paragraph.

References

[1] S. T. T. Fiske, Social cognition: From brains to culture (2020).

[2] R. A. Baron, N. R. Branscombe, Social psychology, Pearson Education India, 2012.

[3] J. Haidt, The emotional dog and its rational tail: a social intuitionist approach to moral judgment., Psychological review 108 (4) (2001) 814.

Notes

Robot Details and Role. The experiment involved the NAO humanoid robot in an autonomous life setting, specifically simulating a breathing animation. This design aimed to give participants the impression of being observed subtly. This was core to the experiment, as it replicated the watching eye effect without direct interaction.

Impact of Robot Presence. The robot's presence in the room was intended to explore its influence on prosocial behavior, particularly in the context of moral decision-making, measured by charitable donations. The participants were unaware beforehand that a robot might be present or that donations would be part of the taks.

Social Cognition. Social Cognition is a field of study that investigates how individuals perceive, interpret, and respond to social stimuli and interaction *i.e.*, events, actions, or signals in a social environment that influence individuals' behaviors and responses [1, 2], encompassing the processes by which people understand themselves and others. It emphasises both the automatic and deliberate aspects of social interactions, with a focus on how cognitive processes operate in social contexts.

In early 2000, Jonathan Haidt in [3] laid the foundation of the Social Intuitionist Model (SIM) understanding how moral judgments are primarily driven by quick, automatic intuitions rather than deliberate reasoning processes. SIM contracts traditional models of moral reasoning, empathising the subconscious and socially intuitive nature of moral behaviour. This shift was important in highlighting the intuitive nature of moral reasoning which contrasted the established belief that reasoning was the main driver of moral-decision making. After Haidt's theoretical introduction, a series of empirical studies supported and expanded upon his claims.

Greene et al. conducted fMRI studies to explore the neurological basis of moral decision-making, providing empirical support for Haidt's model. Their findings

demonstrated that emotional regions of the brain were more active during moral dilemmas involving personal engagement (e.g., in "trolley problems"). Greene's 2001 paper, "An fMRI Investigation of Emotional Engagement in Moral Judgment," found that emotionally engaging dilemmas activated brain regions linked to emotion, reinforcing Haidt's idea that emotions, rather than rational deliberation, often drive moral judgments.

In 2004, Greene expanded this work with the study, "The Neural Bases of Cognitive Conflict and Control in Moral Judgment," demonstrating that rational control is often secondary to emotional intuitions in moral scenarios. Greene's later works, such as his 2008 paper, integrated a dual-process theory that further solidified Haidt's ideas by showing that moral judgment is influenced by both intuitive/emotional and rational/cognitive processes. This model helps reconcile instances where rational deliberation plays a role, complementing Haidt's initial SIM by illustrating a spectrum between automatic intuitions and deliberate reasoning.

Other researchers contributed by showing how automatic, unconscious processes play a central role in moral judgment, which supported Haidt's position. Studies on priming effects in moral decision-making illustrated how subtle cues could shift moral judgments without individuals being consciously aware of these influences.

One of the main contributions to SIM comes from provided by modern Social Cognition that provides a vast experimental evidence that human inferences leading to the execution of certain behaviors occur without precise, clear, a priori conscious decision-making—i.e., without the involvement of conscious and situated rational deliberation processes.

Research in psychopathology, particularly in the context of schizophrenia, further supports this view by demonstrating that critical aspects of social cognition, such as emotion perception and theory of mind (ToM), operate at a largely subconscious level. Penn et al. (2008) and Green et al. (2008, 2015) describe

these processes not as deliberate, conscious judgments, but rather as automatic responses to social cues. For example, the concept of motor resonance—where the observation of another's behavior triggers neural activation similar to performing that behavior oneself—illustrates the intuitive and automatic nature of social understanding. These findings reinforce Haidt's SIM by showing that social cognitive processes, like emotion perception and mentalizing, are largely pre-reflective and automatic, further underlining the dominance of intuition over reasoning in shaping moral judgments.

This aligns strongly with Jonathan Haidt's SIM, which posits that moral judgments are primarily the result of automatic, intuitive processes rather than explicit reasoning, highlighting how these social cognitive functions usually operate largely outside of conscious awareness. By showing that even in altered psychological states these processes remain largely automatic, the argument for the automatic and intuitive nature of social cognition, as posited by SIM, becomes more compelling. The use of psychopathology provides a contrasting scenario that highlights the essential, subconscious operation of these processes in normal functioning.