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. We show that 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 can 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) **An experimental activity** about the interplay between the presence of social robots and human prosocial behaviour.
b) **A comparative analysis of the literature** that suggests the emergence of the following:

   I. Two related, but distinct, research themes in Machine Ethics which we call Human-Machine Ethics and Computational Machine Ethics.

   II. The emergence of two distinct trends in Psychology and Philosophy, *i.e.* cognitive/affective models of moral judgments and rationalism/intuitionist approaches to moral reasoning, that exert a deep influence on the research objectives and methodologies in Computational Machine Ethics.

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 (a subfield of 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. Moral Judgement

4. Machine Ethics

Over the past decade, the field of Machine Ethics—an area of inquiry concerned with the moral dimensions of artificial systems—has garnered significant attention from the scientific community.

FP: Good, now the problem is the start and structure of the next portion of text but one step at time! Keep nextdiv for next portion. It's good to have a division like this. Remember the focus is to introduce ME as soon as we open.

The growing prominence of Machine Ethics within the scientific community is evidenced by multiple indicators. A bibliometric analysis of AI and ethics publications revealed a significant rise in research output, with over 1,500 papers published by mid-2021 and a marked acceleration from 2014 onward [1]. Foundational works, such as Anderson and Anderson's Machine Ethics (2011), have played a pivotal role in catalyzing this interest by providing early frameworks for exploring the moral capabilities of artificial systems [2].

This trend is further underscored by the institutionalization of AI ethics as a recognized academic discipline. A keyword analysis spanning two decades identifies 2014 as a pivotal year, marking a surge in academic engagement with terms related to Machine Ethics and the ethical dimensions of AI research [3]. The increasing focus on these issues is reflected in the growing body of literature addressing topics like transparency, accountability, and human oversight in autonomous systems [5].

Finally, recent meta-analyses highlight the expansion of Machine Ethics as a domain of inquiry, capturing the attention of researchers and practitioners alike. For example, Otterbacher *et al.* (2023) discuss the heightened focus on ethical challenges posed by AI, particularly over the past decade [4]. Together, these developments illustrate a clear trajectory of growth and consolidation, establishing Machine Ethics as an area of significant and sustained interest within the scientific community.

Truth be told, Machine Ethics, as a term and filed, encompasses a wide range of ethical considerations related to machines, often leading to conflation of distinct issues. To address this, the domain of Machine Ethics is redefined through the integration of philosophical traditions and insights from Moral Psychology.

This redefinition establishes a clear vocabulary and experimental framework for investigating how the mere presence of robots influences human moral deliberation and prosocial behavior. Grounded in empirical methodologies, this approach positions Machine Ethics as a critical intersection of ethical theory and moral psychology, advancing the understanding of the nuanced dynamics between humans and intelligent systems

A detailed comparative analysis of the literature on Machine Ethics and related academic disciplines, such as Philosophy and Moral Psychology, highlights a relationship that forms the basis for classifying two distinct research paradigms within Machine Ethics.

Machine Ethics can be divided into two primary areas of focus. The first, termed *Human-Machine Ethics*, examines ethical considerations for humans, centering on behavior in relation to AI and robotic systems. This area addresses societal responsibilities, ethical dilemmas, and legal implications surrounding the deployment and use of AI technologies. The second area, referred to as *Computational Machine Ethics*, focuses on designing systems capable of autonomous moral reasoning, enabling them to operate within predefined ethical parameters.

In most cases, the ethics of AI focuses on the socio-economic, and legal impacts of AI and, the moral and ethical issues surrounding the use of these systems.

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 [2, 7, 8, 9]. For some, *the new field* of Machine Ethics is concearned with giving machines ethical procedures for discovering ways to resolve the ethical problems that *they might encounter*, thus enabling them to function through their own ethical decision [2, pp. i - iv]. Unlike Computer and Information Ethics, which refer to the application of various ethical thoeries (*e.g.* utilitarianism, Kantianism, virtue ethics) to cases that significantly involve computers and computer networks, or to ethical issues surrounding human use of machines [10], the ultimate goal of Machine Ethics, is to create machines that

follow an ideal ethical principle or a set of principles; that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take.

Machine Ethics concears the behaviours of complex , autonomus systems, towards humans [4]

While this work may be classified as technical, I assume readers possess a basic understanding of certain non-technical concepts, such as artificial intelligence and autonomous intelligent systems. For technical terminology mentioned throughout, I provide concise definitions to clarify their context.

A central reason for this trend 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 [2], 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 [11, 12]

FP: this should be part of the closing remarks to introduce the experiment.

In Floridi's view [13], machines that exibit such capabilities for moral deliberaiton, needs to be moral agents in the sense that they should be capable of performing actions with moral significance whcih in turn is define by two primary criteria.

(P1) **Autonomy**. A moral agent must have the ability to act independently and make choices that are not entirely determined by external forces. Autonomy entails intentionality, which means the agent acts with a purpose or goal in mind. A moral agent must be capable of being held accountable for its actions.

(P2) **Responsibility** This accountability implies that the agent understands the consequences of its actions and can justify them within a moral framework. robots or algorithms, as they can act in ways

that significantly affect the moral landscape. However, he acknowledges that their "agency" is limited and derivative because they lack intrinsic moral intentionality, functioning instead within parameters set by human designers.

A moral patient, in contrast, is an entity that can be affected by the actions of a moral agent and therefore deserves moral consideration. Floridi emphasizes:

(P1) **(Intrinsic Worth)**Moral patients have value in and of themselves and must not be treated merely as a means to an end (echoing Kantian ethics).

(P2) **Vulnerability** Moral patients are characterized by their capacity to be harmed or benefited. This includes humans, animals, ecosystems, and increasingly, informational entities (e.g., data or digital environments).

(P3) **Inclusiveness** Floridi's framework expands traditional moral boundaries. He argues for considering entities like artificial agents and digital ecosystems as potential moral patients, based on their capacity to be affected within the infosphere.

Floridi's theory is particularly relevant in the context of the infosphere, a term he uses to describe the informational environment we live in, including digital and physical realms. He introduces the concept of distributed morality:

a) Actions are no longer confined to individual moral agents but are distributed across networks of human and non-human actors.

b) For instance, a decision made by an algorithm (moral agent) might impact a user (moral patient) in ways that require ethical scrutiny.

Floridi moves beyond anthropocentric (human-centered) ethics to advocate for ontocentric (being-centered) ethics, where all entities with intrinsic worth, including informational and digital entities, are considered. By redefining moral agents and patients, Floridi expands moral responsibility to include designers

and users of technology, artificial agents with decision-making capabilities, and vulnerable systems in the infosphere. Floridi's theory is grounded in information ethics, where the fundamental moral value is the "flourishing of the infosphere." Harm is understood as any action that degrades the informational integrity or flourishing of an entity. The implications are Floridi's framework suggests that as AI becomes more integrated into society, its role as a moral agent and its impact on moral patients (users, ecosystems, etc.) must be carefully managed. The inclusion of ecosystems and informational entities as moral patients aligns with broader discussions on environmental ethics and sustainability. By emphasizing distributed morality, Floridi's theory calls for collaborative responsibility in technology design, governance, and ethical AI deployment. In summary, Floridi's theory redefines moral agents and patients to address the complexities of modern, interconnected environments, emphasizing responsibility, inclusiveness, and the flourishing of all entities within the infosphere.

5. 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)?

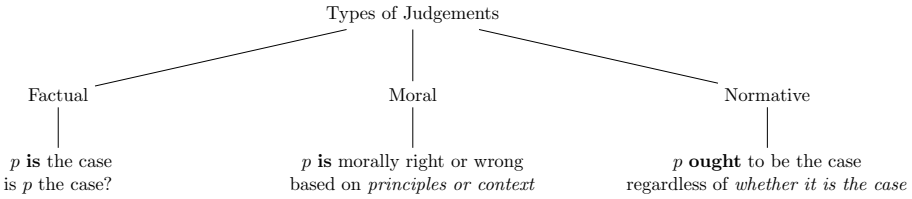

Figure 1: Diagram of Types of Judgements.

Paragraph.

Paragraph.

References

[1] C.-W. Chuang, A. Chang, M. Chen, M. J. P. Selvamani, B.-C. Shia, A worldwide bibliometric analysis of publications on artificial intelligence and ethics in the past seven decades, Sustainability 14 (18) (2021) 11125. URL https://www.mdpi.com/2071-1050/14/18/11125

[2] M. Anderson, S. L. Anderson, Machine ethics, Cambridge University Press, 2011.

[3] D. K. Gao, A. Haverly, S. Mittal, J. Wu, J. Chen, Ai ethics: A bibliometric analysis, critical issues, and key gaps, International Journal of Business Analytics (IJBAN) 11 (1) (2024) 1–19.

[4] J. Otterbacher, Y. Manolopoulos, Machine ethics research: Promises and potential pitfalls, IEEE Intelligent Systems 38 (4) (2023) 62–68.

[5] W. Wallach, C. Allen, Moral machines: Teaching robots right from wrong, Oxford University Press, 2008.

[6] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: Survey of an emerging domain, Image and vision computing 27 (12) (2009) 1743–1759.

[7] V. Nallur, Landscape of machine implemented ethics, Science and engineering ethics 26 (5) (2020) 2381–2399.

[8] L. M. Pereira, A. B. Lopes, Machine ethics: from machine morals to the machinery of morality, Springer, 2020.

[9] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, A. Bernstein, Implementations in machine ethics: A survey, ACM Computing Surveys (CSUR) 53 (6) (2020) 1–38.

[10] T. Bynum, Computer and information ethics, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, summer 2018 Edition, Metaphysics

Research Lab, Stanford University, 2018, p. n/a.
URL https://plato.stanford.edu/archives/sum2018/entries/ethics-computer/

[11] C. Allen, W. Wallach, I. Smit, Why machine ethics?, IEEE Intelligent Systems 21 (4) (2006) 12–17.

[12] C. Allen, W. Wallach, Moral machines: contradiction in terms or abdication of human responsibility, Robot ethics: The ethical and social implications of robotics (2012) 55–68.

[13] L. Floridi, J. W. Sanders, On the morality of artificial agents, Minds and machines 14 (3) (2004) 349–379.

[14] 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 , 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.

FP: Fiske2020, Baron2012

In early 2000, Jonathan Haidt in [14] 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.