Experimental Methods for Moral Behaviour Analysis in Human-Robot Interaction

Francesco Perrone

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

February 2023

This work has been partially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the grant "*Socially Competent Robots*" (EP/N035305/1).

Abstract

Abstract text goes here.

Contents

List of Tables

List of Figures

Acknowledgements

Acknowledgements text goes here.

Declaration

With the exception of chapters 1, 2 and 3, which contain introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated.

1. Introduction

Moral decision making, is the cognitive process of choosing between competing moral judgments *i.e.*, mutually exclusive evaluations we make on what is right or wrong, good or bad, and that we use as motive, purpose and direction for our conscious, and practical behaviour.

a) **Cognitive Process** This term refers to the mental actions or operations that individuals use to acquire knowledge and understanding. It includes processes such as perception, memory, reasoning, decision-making, and problem-solving. Cognitive processes are essential for interpreting and interacting with the world;

b) **Behaviours:** In academic terms, behaviours are the observable actions or reactions of an individual in response to external or internal stimuli. These actions can be voluntary or involuntary and are influenced by various factors, including cognitive processes, emotions, and environmental conditions.

Moral decision making is the intricate cognitive process of choosing between competing moral judgments; these are mutually exclusive evaluations we make regarding what is right or wrong, good or bad. These judgments serve as the motive, purpose, and direction for our conscious and practical behaviour. This process involves an array of cognitive functions such as perception, memory, reasoning, and problem-solving, which collectively inform our moral evaluations and decisions. Moreover, these cognitive processes translate into behaviours, which are the observable manifestations of our moral choices. These behaviours, whether conscious or subconscious, reflect our internal moral deliberations and are influenced by a complex interplay of cognitive functions, emotions, and contextual factors. Hence, moral decision making encompasses both the mental operations that guide our judgments and the resultant actions that embody our moral principles in the practical realm.

The perception of direct gaze, that is, of other individual gaze directed at the observer, is know to influence a wide range of cognitive processes and behaviours.

1.1 Machines' 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 role of Machine Ethics has received increased attention across a number of academic disciplines, in the past few years

[1].

A central reason for this encouraging circumstance is an unprecedented inter-

disciplinarity: researchers in Machine Ethics are now capable of freely drawing on scientific resources from well beyond the confines of their fields, a scientifically robust data that can now be integrated and used as a laboratory to verify e generalise more qualitatively philosophical outsets which were common of its foundational work [1, 6].

The broad concept of "artificial intelligence" (AI) encapsulates any form of synthetic computational mechanism that exhibits intelligent actions, which are complicated actions conducive to achieving objectives. We aim to refrain from confining "intelligence" strictly to tasks requiring human intellect, contrary to Minsky's proposal [25]. Thus, we include a wide array of machines, encompassing "technical AI" systems that demonstrate only limited learning or reasoning skills but excel in task automation, and "general AI" systems designed to establish a universally intelligent agent. AI tends to intertwine more with our existence than other technologies, hence the emergence of the "philosophy of AI". Possibly, this arises from the AI's endeavour to fabricate machines that possess attributes that we humans perceive as vital to our identity, such as the ability to feel, think, and show intelligence. The primary roles of an AI agent likely involve sensing, modelling, planning, and execution, but current applications extend to perception, text scrutiny, natural language processing (NLP), logical deduction, gameplaying, decision-making aids, data analysis, predictive analytics, along with self-operating vehicles and other robotic manifestations [34].

AI might employ various computational strategies to achieve these goals, like classic symbol-manipulating AI, cognitive inspired processes, or machine learning through neural networks [20, 33]. It's important to acknowledge that historically, the term "AI" was used as previously mentioned roughly between 1950-1975, followed by a period of skepticism during the "AI winter", approximately from 1975-1995, and was subsequently constrained. Consequently, areas like "machine learning", "natural language processing", and "data science" were typically not categorized as "AI". Around 2010, the usage expanded again, with at times nearly all of computer science and even high-tech being consolidated under "AI". Presently, it has transformed into a prestigious moniker, a thriving sector with substantial capital investment [32], and is on the brink of resurging hype. As Erik Brynjolfsson pointed out, it might empower us to virtually eliminate global poverty, massively reduce disease, and provide superior education to almost every person on earth [2].

While AI can solely be software-based, **robots are tangible machines capable of movement**. Robots are subject to physical effects, primarily via "sensors", and exert physical force onto the environment, typically through "actuators", such as a gripper or a rotating wheel. Therefore, autonomous vehicles or aircrafts are robots, and only a tiny fraction of robots are "Humanoid" (human-resembling), as depicted in films. Some robots employ AI, while others do not: Standard industrial robots rigidly adhere to fully defined scripts with minimal sensory input and devoid of learning or reasoning (approximately 500,000 such new industrial robots are deployed each year [23]). It is likely appropriate to state that although robotic systems incite more apprehension among the public, AI systems are more likely to significantly influence humanity. Moreover, AI or robotic systems designed for a narrow range of tasks are less likely to pose new

challenges than more flexible and independent systems. Hence, robotics and AI can be visualized as encompassing two intersecting categories of systems: those that are solely AI, those that are strictly robotic, and those that are a combination of both. Our interest spans all three; the focus of this article encompasses not just the intersection, but the amalgamation, of both categories. In the rapidly progressing domains of artificial intelligence (AI) and social robotics, the necessity of ethical deliberation and moral agency is paramount. As these technologies become increasingly sophisticated and entrenched in our everyday lives, timeless philosophical queries concerning purpose, potentiality, and morality gain renewed relevance. Ancient Greek philosophers endeavoured to delineate and comprehend human moral agency, a task that now confronts us in the context of AI and robotics. Drawing on the profound insights of philosophers like Aristotle, we can navigate and address the unique ethical conundrums raised by these technologies. However, it is crucial to recognise a prevalent shortcoming in the discourse on AI and robotics. Academics and authors in the field frequently employ terms such as "moral and morality", "ethics", "intentionality and agency", yet these concepts often lack a deep philosophical grounding [26]. This absence of philosophical understanding can lead to misconceptions and flawed assumptions, particularly in a field as nuanced as AI [10]. For instance, the application of "moral agency" to AI systems can be contentious, given that traditional interpretations of the term presuppose qualities like consciousness and intentionality that machines do not possess [17]. Similarly, there can be a tendency to anthropomorphise AI systems when discussing their 'ethics,' which can obfuscate the fact that their 'ethical' behaviours are entirely human-programmed [41]. In this paper, we strive not only to draw insightful parallels between ancient philosophy and contemporary ethical discussions in AI and social robotics but also to illuminate and correct potential misconceptions caused by a lack of philosophical understanding. By grounding our discussions in solid philosophical foundations, we hope to foster a more nuanced, accurate, and productive discourse on AI ethics.

Aristotle's teleological view of existence, as detailed in his collective works [7], interprets the universe as inherently intentional. He advocates that potentiality is in service of actuality, asserting that matter's essence lies in the prospect of adopting form[41], paralleling how an organism is endowed with sight for the purpose of perception. In this vein, every entity bears unique potentialities that spring from its form. Drawing upon this, a serpent, due to its form, possesses the capacity to undulate, implying it's naturally inclined towards this movement. The fulfilment of potential is directly tied to the realisation of its intended purpose.

This teleological paradigm serves as the foundation of Aristotle's ethical philosophy [35]. The form of humans confers upon them certain abilities. Hence, their purpose is intertwined with the proficient and complete utilisation of these capacities.

Transitioning to computational morality and robotics, Aristotle's teleological framework presents a compelling lens for analysis. Analogously, robots, initially devoid of purpose, derive their purpose from their programmed tasks and abilities. In a manner similar to Aristotle's view of matter waiting to receive form, a raw computational canvas exists to embrace coding and programming[40]. Mirroring an organism's sight intended for seeing, a robot is equipped with sensors

designed to interact with its environment [31].

Each robot, through its specific programming or "form," carries certain capabilities. For instance, an autonomous vehicle, due to its form, has the ability to navigate, implying that it is programmed to do so. The extent to which a robot actualises its potential mirrors the success it achieves in fulfilling its designed purpose.

When Aristotle's teleological worldview is applied to computational morality in AI systems, it generates intriguing considerations. AI systems, due to their 'form' or programming, are vested with certain abilities, such as learning, analysing, and decision-making based on intricate algorithms [**?**]. Therefore, their 'purpose' can be seen as the maximal and effective application of these abilities, aiming to reach ethical decisions that align with their programmed ethical framework [1].

Aristotle's teleological views weren't formed in a vacuum, and they can be further contextualised within the larger discourse among Ancient Greek philosophers. For instance, Plato, Aristotle's mentor, maintained a theory of forms, emphasising an immaterial world of 'perfect' forms separate from our everyday world. Yet, Aristotle rejected this dualism, proposing instead that forms existed in objects and, crucially, it was this form that gave objects their purpose.

Aristotle's emphasis on the form and potentiality of a being can be intriguingly juxtaposed with the concept of "Levels of Abstraction" (LoA) proposed by Luciano Floridi [19]. Floridi suggests that understanding a system requires viewing it at the appropriate LoA, a conceptual lens that filters out unnecessary details and focuses on the information needed to understand or interact with the system. In computational terms, the 'form' of an AI system would correspond to its designed LoA. Just as Aristotle sees a being's form as key to understanding its purpose and potentiality, Floridi sees an AI's LoA as critical to understanding its function and capabilities. This highlights the parallels between ancient philosophical thought and contemporary information philosophy. This connection further emphasises the relevance of Aristotle's teleology to computational morality.

If we take the AI's designed LoA as its 'form', then the purpose of the AI system becomes fulfilling the functions and potentialities set out at this level. This mirrors the Aristotelian notion that an entity's purpose is tied to fulfilling its potentialities as dictated by its form. A complete understanding of computational morality, therefore, requires an appreciation of the designed LoA of the AI system. Just as Aristotle advocated for a nuanced understanding of an entity's form, so too does Floridi's framework encourage us to consider the appropriate LoA when grappling with moral issues in AI and robotics.

Aristotle serves as a starting point for this exploration due to his pivotal role in laying the groundwork of Western philosophical thought. His concept of teleology, or the purposefulness of all things and actions, has significantly influenced subsequent understandings of ethics and morality.

Moreover, his views on Actuality and Potentiality provide a useful lens through which to consider the capabilities and purpose of artificial intelligence. Nevertheless, it is crucial to appreciate that Aristotle's perspective is only the first of many that we will engage with in this investigation. As we traverse the historical

landscape of philosophical thought on morality and ethics, we will encounter a rich tapestry of ideas that each contribute uniquely to our modern grappling with these concepts in the context of AI and social robotics.

Within the realm of formal logic, the precision of definitions constitutes a bedrock. For instance, the rigorous delineation of a proposition as a statement with a definitive truth value - either true or false, but never both nor neither underpins all ensuing discourse. Logical connectives, such as 'and', 'or', and 'not', gain their operational power from the meticulously prescribed relationships they signify between propositions. The process of formulating complex logical rules and inferences becomes an orchestrated composition, owing its harmony to the preciseness of these core definitions [24]. In mathematics, the emphasis on defining primitive entities is equally profound. For example, in set theory, which provides a foundation for virtually all of mathematics, the concept of a set is primitive and left undefined. Instead, the properties and operations of sets are described by axioms, such as those proposed by Zermelo and Fraenkel [37]. In number theory, the definition of what constitutes a number has evolved over time, from the natural numbers to the inclusion of zero, negative numbers, rational numbers, real numbers, and complex numbers, each expansion necessitating a precise definition to avoid ambiguity and contradiction [30]. The rigorous defining of terms is far from a simple formality; it facilitates clear communication, reduces ambiguity, and enhances the richness of academic discourse. The vast terrain of interdisciplinary fields like AI and Social Robotics demands a similar level of precision and clarity in the definitions of often philosophically loaded terms like 'morality', 'ethics', and 'agency', especially given their diverse interpretations across various contexts [26].

Notes

[1] A search for the keyword '*Computational Morality*' alone on Google Scholar yielded an astonishing number of more than 39,000 results as of October 2021. However, as of today, this figure has significantly grown to about 86,200 results, indicating a substantial increase in literature on the subject over the past year. Furthermore, a search for the keyword '*Machine Ethics*' on Google Scholar produced an already staggering number of approximately 3,000,000 results as of October 2021. However, the figure has seen a remarkable growth, now standing at about 3,230,000 results, emphasising the continued expansion of research and scholarly engagement with the ethical aspects of artificial intelligence. These notable increases and changes in the figures for both 'Computational Morality' and 'Machine Ethics' highlight the growing prominence and visibility of these fields within the academic community. They signify the escalating interest among researchers, scholars, and ethicists in investigating the ethical dimensions of computational systems and the moral implications of their actions *at the least*. The significant growth in literature not only reflects a broader understanding of the ethical challenges posed by advancing technologies but also underscores the pressing need to address and discuss the ethical considerations associated with the design, deployment, and impact of computational systems in our society. It is worth noting that the figures provided here are based on a search conducted on Google Scholar as of November 18, 2025. Due to the dynamic nature of online databases, the exact figures may vary over time. Nonetheless, the substantial increase in publications on computational morality and machine ethics signifies the continuous expansion and significance of these fields in the realm of ethical inquiry. The rapid growth of research in the field of computational morality and machine ethics highlights its paramount importance in our increasingly technologically-driven world. As computational systems and artificial intelligence become more integrated into various aspects of society, it is crucial to explore the ethical implications of their actions [**we are not doing this here**]. Understanding and addressing the moral dimensions

of these systems is vital to ensure their responsible development, deployment, and impact on individuals and communities. The remarkable expansion of literature in computational morality is a testament to the urgency and significance of this research area. In fact, the rate of growth in this field often surpasses that of many other scientific and computer science-related disciplines, illustrating the heightened attention and recognition it receives. This exponential rise underscores the interdisciplinary nature of computational morality, drawing insights from philosophy, computer science, sociology, and other fields. It highlights the recognition among scholars, researchers, and practitioners that the ethical considerations and social implications of computational systems are integral to the advancement of technology and the well-being of society as a whole. By delving into computational morality, we pave the way for a future in which ethical principles guide the design, implementation, and use of intelligent systems, ensuring that they align with human values and promote the greater good.

2.

2.1

The term *morality* and its morphosyntactic transformations, are frequently employed in public discourse, policymaking, and interdisciplinary research with significant conceptual ambiguity, often lacking the precision characteristic of academic subjucts such as moral philosophy, normative ethics and psychology [1]. This imprecision results in interpretative inconsistencies and epistemic distortions, particularly when the term morality is operationalised as an *empirical construct* [2, 3]. While moral philosophy has long sought to formalise the principles underlying moral judgment and moral decision-making, its transposition across disciplines remains inconsistent, leading to methodological and theoretical fragmentation. Such conceptual instability is especially problematic in empirical research, where rigorous analysis demands a precise and operationally sound definition of key terms. This is particularly evident in Machine Ethics, Social Signal Processing, Affective Computing, and Emotion AI, wherein the imperative should extend beyond mere operational viability to the instantiation of autonomous moral intelligence. These domains necessitate a formalisation of moral

[1] Morphosyntactic transformation refers to the systematic modification of a word's form and syntactic function within a sentence, altering its grammatical category while preserving or adapting its semantic role. In the case of *morality* and *moral*, this transformation involves the transition from a noun denoting a conceptual system (*morality*) to an adjective (*moral*) that qualifies specific judgments or decisions within that system. For the purposes of this discussion, we assume that what applies to "morality" (*i.e.* a system) also applies to "moral" (*i.e.* some system's instances), recognising that a deeper linguistic and semantic differentiation exists but is beyond the scope of our analysis. However, if the analysis involves computational modeling, normative theory, or conceptual precision, the distinction should not be abstracted away easily, as it could obscure critical methodological differences in how moral reasoning is implemented in machines. More importantly, the shift from "morality" (a noun denoting a conceptual system of ethical principles) to "moral" (an adjective qualifying specific judgments and decisions) occurs implicitly in this passage. While "morality" refers to a broader framework of ethical norms, "moral" in "moral judgment" and "moral decision-making" functions as a qualifier, indicating that these cognitive and behavioral processes are evaluated within such a framework. This transition can be expressed in propositional logic as follows:

$$(\forall x)(M(x) \rightarrow \exists P(J_P(x) \vee D_P(x)))$$

where $M(x)$ denotes that x is a system of moral principles ("morality"), $J_P(x)$ denotes that x is a moral judgment, and $D_P(x)$ denotes that x is a moral decision. This formula can be red in plain English as "any system of morality necessarily enables at least one principle that applies either to moral judgment or to moral decision-making". If implemented computationally, this formulation could inform the design of algorithms that map $\mathcal{M}(x)$ to structured decision models, ensuring that any operationalisation of morality in artificial systems maintains a formally derivable connection to at least one P guiding $J_P(x)$ or $D_P(x)$, thereby preventing arbitrary or ad-hoc moral classifications. The existential quantification over P indicates *that 'moral' as a qualifier derives its meaning from the underlying framework of morality.* This transition is often assumed in our discussion, but making it explicit here might help clarify how abstract ethical theories are applied to concrete evaluative acts.

agency that is neither reducible to heuristic approximations nor susceptible to ad-hoc normative interpolations. The articulation of moral cognition within artificial systems demands an ontologically rigorous and epistemically defensible framework, ensuring that computational architectures are not merely reactive to competing normative imperatives but are endowed with an inferential structure capable of adjudicating between morally salient *stimuli* and *cues*. In the absence of such precision, purportedly moral outputs risk devolving into algorithmic simulacra—an eventuality that has already materialised in contemporary research, where systems masquerading as ethically attuned frequently exhibit inconsistencies, biases, and spurious normativity [4, 5, 6, 7, 8]. The absence of a well-defined theoretical substrate has, in practice, led to models that conflate statistical correlation with moral discernment, engendering a proliferation of mechanised ethical facades devoid of genuine evaluative coherence [9, 10, 11, 12, 13].

This epistemic and methodological deficiency is, at its core, symptomatic of the deeper conceptual misapprehension and misapplication of the term *morality* itself, invoked with significant conceptual ambiguity across interdisciplinary research, often devoid of the precision characteristic of moral philosophy and psychology. This lack of terminological and conceptual rigour has engendered a cascade of methodological distortions, wherein computational models inherit the equivocations of their foundational premises. In failing to acknowledge the epistemic constraints that govern moral reasoning within philosophical and psychological discourse, contemporary approaches in Machine Ethics and related fields risk formalising ethical architectures upon indeterminate and incoherent abstractions, rendering their claims to moral agency not only theoretically untenable but practically defective.

2.1.1

Most of the foundational work in machine ethics (2011) exhibits this terminological imprecision, failing to clearly distinguish between moral cognition, ethical norms, and the technical implementation of moral decision-making. Most of the published work in machine ethics have repeatedly fails to delineate *morality* as a cognitive phenomenon from *ethics* as a normative, prescriptive framework. This is evident in Moor's paper [14] where *ethical agents* are described without addressing whether these agents operate under descriptive moral cognition (human-like decision-making processes) or prescriptive ethical norms (principle-driven reasoning). For example, the term *ethical-impact agents* is used to describe machines that influence ethical considerations but are not intrinsically ethical in nature. The very framing of this category is problematic, as it suggests that the ethical implications of an action are equivalent to an agent possessing ethical reasoning capabilities. Such an assumption, if left unchecked, leads to the misguided notion that any AI system with ethical consequences is *moral* or *ethical*. Furthermore, in works as Moor's, classification of machines into implicit ethical agents, explicit ethical agents, and full ethical agents lacks conceptual clarity due to the unexamined interchangeability between ethics and morality. Implicit ethical agents are described as systems whose behaviors *promote ethical behavior* through design constraints. However, these are not *ethical* agents but rather engineered control systems designed to minimize undesirable behaviors. The label *ethical* in this context wrongly implies an inherent moral capability, whereas the systems merely

FP: why did I put "For example" in here?

encode regulatory constraints. Explicit ethical agents, as described by Moor, operate on symbolic ethical reasoning, but again, this classification presupposes that rule-based reasoning is equivalent to moral cognition. This is a categorical mistake, as *an agent following ethical constraints does not mean it possesses moral understanding*. The effect of this imprecise terminology is the erosion of theoretical rigor in Machine Ethics. If computing scientists adopt these flawed distinctions, the field risks overstating the moral capacities of AI systems while neglecting the fundamental difference between ethics as an engineered property and morality as a cognitive phenomenon. Most of these studies explicitly or implicitly ask: "Can a machine be a full ethical agent? [15]" without first providing a structured definitional framework for what *full ethical agency* entails. The argument proceeds without clearly distinguishing between:

1) Moral agency (the ability to autonomously form and revise moral judgments)

2) Ethical compliance (the ability to follow externally imposed ethical principles)

3) Ethical justification (the ability to offer defensible normative reasoning)

The failure to differentiate these reinforces the mistaken belief that moral capabilities can be put into machines without first clarifying what it means for a machine to *possess* such ability. If full ethical agents are simply machines with complex rule-based ethical architectures, then the term loses its normative weight. This blurs the boundary between human moral cognition and computational ethical rule-following, which could mislead engineers into assuming that sophisticated AI systems possess moral insight, rather than simply executing rule-based approximations of ethical principles.

Publications that concentrate on autonomous moral reasoning and decision-making in artificial agents have frequently exhibit substantial terminological imprecision and a conflation of moral and ethical concepts, undermining efforts to establish a robust metaethical foundation. Metaethics, in its proper and most rigorous sense, is not concerned with the mere prescription of conduct nor the delineation of duties, but with the fundamental nature of moral thought itself. It inquires into the meaning of moral terms, the objectivity or subjectivity of ethical claims, and the epistemic status of moral knowledge. To speak of 'goodness' or 'duty' without a prior account of whether such notions are objective or conventional, whether they track facts or express attitudes

FP: This links to our work directly

, is to mistake the foundations of morality for its superstructure [2]. Any discourse

[2]In this context, 'superstructure' denotes the derivative and contingent domain of ethical frameworks, normative systems, and prescriptive theories that rest upon deeper metaethical inquiries. The term originates in structuralist and Marxist discourse, where it designates the ideological, legal, and cultural formations built upon an economic base. Here, it signifies the theoretical and procedural structures guiding ethical reasoning, which presuppose prior commitments regarding the nature, justification, and epistemic status of moral claims. Ethics, as the study of principles governing conduct, operates at the level of this superstructure; metaethics, by contrast, concerns itself with the conceptual and ontological conditions that render ethical

on the moral agency of artificial systems that neglects these inquiries is, at best, premature, and, at worst, methodologically incoherent.

2.2

While the author recognizes that machines may influence ethical considerations, the discussion fails to clearly distinguish between moral cognition, ethical norms, and the technical implementation of ethical decision-making. The consequences of this imprecision are non-trivial, as they introduce ambiguities.

Even more so, if emerging research fields such as Machine Ethics, attempt to imbue machines with capabilities for autnomous moral decision-making. Testimony of this inconsistency is perhaps the most important effort made by the AI community to define a common agenda for such purpose. The 2011 volume Machine Ethics [16], edited by Michael Anderson and Susan Leigh Anderson, is widely recognized as a seminal work that has significantly influenced subsequent discourse in the field of machine ethics and related fields. This collection of essays represents foundational efforts by philosophers and artificial intelligence researchers to address the necessity of integrating ethical dimensions into autonomous machines, exploring the philosophical and practical challenges inherent in this endeavor. The impact of this work is evident in its extensive citation across scholarly literature. For instance, the volume has been referenced in discussions about the formalisation and scalability of ethical principles in intelligent autonomous systems, highlighting its role in shaping contemporary approaches to embedding ethics within AI design.

Prior to its publication, discussions on ethical decision making in artificial agents were largely scattered across disciplines such as philosophy, cognitive science, and artificial intelligence. While previous studies, including Moor 1985 and 2006 [17, 14], Moor (2011) [15] and Allen, Wallach, and Smit (2005) [6], had outlined the conceptual foundations of ethical machines, the Andersons' volume consolidated these perspectives into a structured research agenda. By assembling a diverse range of theoretical and applied contributions, it provided a unifying framework that distinguished Machine Ethics from broader discussions on Computer Ethics, emphasizing the moral agency of artificial systems rather than the ethical responsibilities of human operators. The volume is widely cited as a seminal reference and has significantly shaped subsequent discourse on the subject. However, despite its foundational role, terminological ambiguity persists within the field. The frequent interchangeability of terms such as morality and ethics, often treated as synonymous despite their conceptual distinctions, has contributed to an ongoing lack of precision in defining the normative constraints and decision-making frameworks applicable not only to artificial agents and related fields.

Establishing a rigorous conceptual framework is thus a prerequisite for ensuring that empirical approaches to moral pattern detection are rooted in analytical coherence rather than contingent or culturally idiosyncratic interpretations.

discourse possible. Any attempt to construct ethical decision-making architectures in artificial agents without clarifying these metaethical presuppositions risks mistaking the contingent for the necessary, yielding systems that simulate ethical deliberation without securing its intelligibility or coherence.

A brief etymological analysis of the term is a necessary preliminary step, as it provides critical insights into its conceptual foundations and historical semantic shifts. By tracing the term's linguistic evolution, we can clarify its epistemic underpinnings and mitigate the risk of importing anachronistic or culturally contingent assumptions into our theoretical framework.

It should be acknowledged that a thorough etymological investigation of the term is an extensive undertaking—arguably the subject of an independent doctoral inquiry—which lies beyond the scope of this work. Instead, we will draw upon existing scholarly analyses and established linguistic research to extract the necessary insights for our purposes.

Our goal is to illuminate the term's meaning in a manner that is both conceptually rigorous and functionally relevant to our investigation into machine-detectable moral cues, as well as to enhance our understanding of the meaning and usage of technical terms such as *Moral Judgment* and *Moral Decision-Making*. Given that empirical analyses in this domain require a clearly delineated and operationally sound definition of *morality*, this clarification is not only essential for our study but also valuable for researchers in Computing Science and related sub-fields engaging with similar questions.

2.3 Justification for Etymological Analysis in Understanding Moral Concepts

While alternative methodologies—such as conceptual analysis [18], discourse analysis [19], or even experimental cognitive studies—offer valuable insights into how moral terms function in contemporary discourse, they do not account for the historical trajectories that have given these terms their present semantic and normative weight.

2.4 Linguistic Evolution of "Morality"

In justifying an etymological analysis as a methodological tool for generating understanding—while maintaining consistency with the empirical nature of this work—it is worth noting that direct empirical studies explicitly linking etymology to enhanced comprehension remain limited. However, scholarly literature in psychology and philosophy supports the claim that examining a word's etymology can deepen our understanding of its meaning and function by elucidating its conceptual evolution and cognitive associations [20]

Philosophically, hermeneutics—the study of interpretation—emphasizes the significance of historical and contextual analysis in understanding texts and language [21]. Etymological inquiry serves as a methodological tool within this tradition, uncovering layers of meaning shaped by cultural and historical contingencies, thereby refining our comprehension of a term's contemporary usage [22]. This approach aligns with the hermeneutic principle of the fusion of horizons, wherein understanding emerges through the dynamic interplay between a text's historical context and the interpreter's conceptual framework [23].

Psychologically, the concept of *apperception* involves the process by which new

experiences are assimilated into existing cognitive frameworks. Understanding a word's etymology allows individuals to connect new linguistic information to prior knowledge, facilitating deeper comprehension and retention. This process underscores the role of etymology in shaping our cognitive structures and enhancing our understanding of language [20, 24].

Thus, an etymological examination of the term morality serves as both a hermeneutic and cognitive tool, offering deeper epistemic insight into its conceptual foundations and historical evolution. By uncovering the linguistic, cultural, and philosophical contexts that have shaped its meaning over time, this analysis refines our theoretical understanding and enhances the clarity of its contemporary usage.

2.4.1 On the Origin of the Word

The term itself, morality, is a product of Latin intellectual efforts to translate and adapt Greek ethical concepts, yet the ideas it denotes predate this linguistic shift. The history of moral thought cannot be fully understood without tracing its origins to Greek philosophy, where early conceptions of virtue, conduct, and ethical reasoning took form. While Aristotle (384–322 BCE) is often regarded as the first philosopher to systematically explore morality, his work did not emerge in a vacuum. The intellectual landscape of moral thought was already shaped by pre-Socratic traditions, particularly the Pythagorean school (c. 6th century BCE), whose teachings wove together ethical precepts, metaphysical principles, and a structured way of life [25].

Sidgwick notes [25] that the Pythagorean tradition was less a school of ethical philosophy in the modern sense and more a moral and religious order, built on a cosmological vision that sought to align human conduct with mathematical and harmonic principles. This view extended even to justice, which they conceived as a "square number," reflecting the idea of proportional retribution [25, p. 12]. Their emphasis on self-discipline, moderation, and purification was closely tied to their belief in metempsychosis (the transmigration of souls), suggesting an early framework in which moral conduct had direct consequences for the fate of the soul [25, p. 10].

This connection between early Greek ethical reflection and its later Aristotelian articulation is crucial for understanding the historical development of moral philosophy. If we acknowledge that Greek ethics was expressed in terms of *êthos* (ἦθος) and *êthikê* (ἠθική), rather than through the Latin-derived "morality", we can better appreciate the conceptual shifts that occurred in both language and thought. Sidgwick points out that while Aristotle ((384–322 BCE) brought moral philosophy into a systematic form, he was inheriting a legacy of moral reflection that had already developed within pre-Socratic traditions (c. 6th–5th century BCE) like Pythagoreanism (6th century BCE onward), Heraclitean thought (c. 535–475 BCE), and Democritean ethics (c. 460–370 BCE) [25, p. 18]. The transition from Greek to Latin terminology was not merely a matter of translation; it also signified a shift in emphasis—from a philosophical discourse concerned with the formation of character and virtue to one increasingly embedded in social norms and customs. The following discussion will explore how the Pythagorean

tradition (c. 6th century BCE) contributed to the foundations of moral thought, preceding Aristotle and influencing the trajectory of ethics in Western philosophy.

The word "morality" has a Latin etymology. It derives from the term *moralis*, coined by Cicero to translate the Greek êthikos (ἠθικός), which in turn derives from êthos (ἦθος), meaning "custom," "character," or "habit." In other words, the concept of morality has Greek origins, but the term we use today is of Latin derivation.

Why the Term "Morality"?

If the earliest discussions on morality can indeed be traced back to the Pythagoreans and later Aristotle, it is also true that the Greek language expressed these concepts with ἦτηος (êthos, ἦθος) and ἠτηικῆ (êthikê, ἠθική). However, in the transition from Greek to Latin, an equivalent term became necessary. Cicero, in his attempt to adapt Greek philosophical vocabulary to the Latin language, employed *moralis* as a calque of ἠτηικος (êthikos, ἠθικός), deriving it from *mores* (the customs, habits, and traditions of a people).

Indeed, Latin already had words to indicate virtuous behavior or righteous conduct, such as *virtus* (virtue) and *honestas*. However, *moralis* served to translate directly the Greek concept of ἠτηικῆ (êthikê, ἠθική). From this root, *moralitas* later emerged, which, during the medieval period, became established as the term we use today to denote the set of principles regulating human action in relation to good and evil.

An Interesting Detail

Aristotle uses ἦτηος (êthos, ἦθος) in two distinct senses:

1) ἦτηος (êthos, ἦθος) with a long eta → denotes character and moral dispositions.

2) ἔτηος (éthos, ἔθος) with a short epsilon → denotes habit or custom.

This distinction suggests that the Greeks conceived moral behavior both as an innate disposition of character and as something shaped through habit and practice. The Latin *mores*, however, emphasizes above all the social dimension of morality, that is, the set of shared norms within a community. From this perspective, the Latin term implies a more normative conception compared to Aristotle's, which was more closely linked to ethics as a form of *areté* (areté, ἀρετή) (personal excellence).

Thus, the term "morality" is the result of a linguistic evolution that originates from Greek philosophy but takes its definitive shape in the Latin tradition, thanks to Cicero and later medieval scholars who solidified it within Western ethical discourse.

2.5 Defining Moral

Understanding morality as a philosophical concept requires engaging with a discourse that spans millennia, drawing on diverse traditions, frameworks, and interpretations. This section does not attempt to provide an exhaustive account of all perspectives on morality—such an endeavor would constitute an entire research work of its own—but instead introduces key philosophical milestones that have shaped the discourse. By referencing fundamental contributions, this review serves as a guide to some of the cornerstone works necessary to understand moral philosophy, without claiming to cover the full breadth of perspectives. Despite the depth and complexity of these contributions, there remains no single, universally accepted definition of morality. The term is used in at least two primary senses: descriptively, to refer to moral norms adhered to by societies or individuals, and normatively, to refer to principles that rational agents would endorse under ideal conditions [26]. This inherent ambiguity has led some scholars to question whether morality can be meaningfully defined at all, with some arguing that it lacks a unified essence and is best understood as a historically contingent and philosophically contested construct [27]. Given this conceptual landscape, the present discussion will focus on how foundational philosophical works have contributed to shaping contemporary understandings of morality, while recognizing that, for the purpose of this work on moral decision-making in human-robot interaction, operational definitions will be employed in a way that is internally consistent but not universally binding. To trace the intellectual development of morality, we begin with Aristotle, whose virtue ethics provides one of the earliest structured accounts of moral thought.

Aristotle (350 BCE) in Nicomachean Ethics [28], laid the foundation for moral philosophy by defining morality in terms of character and virtue, emphasizing that the highest human good is eudaimonia, or human flourishing, which is achieved through habituation and the exercise of practical wisdom (phronesis). Centuries later, Thomas Hobbes (1651) in Leviathan [29] introduced a radically different approach by defining morality as a social construct emerging from self-interest, necessary for societal stability and preventing the "state of nature", which he described as a war of all against all. Jean-Jacques Rousseau (1762), in The Social Contract [30], responded to Hobbes by proposing that morality derives from an innate human goodness that is later corrupted by society, distinguishing between natural compassion and the moral norms constructed within political communities. Around the same time, David Hume (1739), in A Treatise of Human Nature [31], argued that morality is fundamentally grounded in human emotions rather than pure reason, contending that moral judgments stem from sentiments like sympathy rather than logical deductions. Immanuel Kant's Groundwork of the Metaphysics of Morals [32] (1785) provided a rationalist counterpoint, rejecting Humean sentimentalism and asserting that moral principles must be derived from reason alone, introducing the categorical imperative, which dictates that moral laws should be universalizable and rooted in respect for human dignity. Shortly after, John Stuart Mill (1861), in Utilitarianism [33], introduced a consequentialist approach, defining morality as the maximization of happiness and minimization of suffering, advocating that the rightness of actions is determined by their outcomes in relation to the greatest happiness principle. Friedrich Ni-

etzsche (1887), in On the Genealogy of Morality [34], radically challenged traditional moral frameworks, particularly those influenced by Christianity and Kantian ethics, arguing that moral norms are historically contingent and rooted in power dynamics. He distinguished between master morality, characterized by strength and self-affirmation, and slave morality, which he associated with humility, guilt, and compassion, calling for a re-evaluation of values. These foundational works collectively shape contemporary moral thought, integrating virtue ethics, deontological principles, consequentialism, sentimentalism, social contract theory, and historical critique into a comprehensive framework for understanding morality.

If we assume that there is such a thing as a linear chronological continuum in the development of a universally accepted definition of *morality*, we could define the term as follows:

Definition 1 (Morality)**.** *the system of principles, values, and norms that guide human conduct, balancing reason, emotion, and social cooperation to determine right and wrong. It is grounded in the pursuit of human flourishing (Aristotle), the rational application of universal duties (Kant), the maximization of collective well-being (Mill), the influence of human sentiment (Hume), the historical and cultural re-evaluation of values (Nietzsche), and the necessity of social cohesion (Hobbes and Rousseau).*

This definition acknowledges morality as a multi-faceted construct, shaped by ethical reasoning, human emotions, social agreements, and evolving historical contexts. It integrates the core elements of virtue ethics [28, 35], deontology [32, 36], consequentialism [33, 37], sentimentalism [31, 38], critique of values [34], and social contract theory [39, 40], making it a synthesis of the most enduring contributions to moral philosophy.

This comprehensive synthesis, however, does not imply the existence of a single, universally accepted definition of morality. As noted by Gert and Gert [26], the concept of morality is inherently ambiguous, employed in at least two distinct senses: a descriptive sense, referring to the codes of conduct endorsed by societies, groups, or individuals, and a normative sense, which attempts to establish a universal standard that rational agents would accept under ideal conditions. This *duality* or *dicotomy*, alongside historical and cultural variations in moral thought, prevents any definitive, uncontested characterization of morality. Some philosophers, such as Sinnott-Armstrong [27], have even argued that morality itself is not a unified domain, making any attempt at a singular definition inherently problematic. Furthermore, while moral philosophers have long engaged in theorizing about moral principles and ethical frameworks, explicit definitions of morality remain scarce, often giving way to discussions about moral judgment instead [41, 42]. This ongoing conceptual fragmentation underscores that morality is not a fixed, objectively defined entity, but rather a historically contingent and philosophically contested construct, one that continues to evolve in response to new ethical, social, and scientific challenges.

Morality is a system of guiding principles that regulate behavior by distinguishing between right and wrong actions [26, 43]. It exists in two broad forms: *a)* as

a descriptive system, which refers to the moral codes followed by societies or individuals [36], and *b)* as a normative system, which refers to the principles that would be endorsed by all rational agents under specified conditions [44]. While moral norms vary across cultures and contexts, their function is generally to facilitate social cooperation [45], reduce harm [46], and promote fairness.

FP: Portion about moral judjments here. Also by now there should be two dicotomies and relative figures.

The study of morality requires not only an examination of specific moral judgments but also an understanding of how moral systems function as a whole. Philosophers attempt to systematically account for morality, treating it as a normative system with defined principles and reasoning structures. At its most basic, morality consists of norms and principles that regulate human actions, particularly in relation to others. These norms are not arbitrary but are seen as carrying a special kind of weight or authority, meaning they are not merely preferences but obligations that structure human interaction [47].

Beyond this minimal definition, morality can also be understood as a system of moral reasons—either grounded in some more fundamental values or, alternatively, as the foundation upon which value is built [48]. This dual perspective raises an important question: Are moral norms universal, or do they vary based on cultural and situational factors?

A common view is that moral norms are universal—they apply to all rational agents in similar circumstances. This universality is supported by the notion that moral principles can be formulated without reliance on specific personal details, meaning they should apply consistently across cases rather than being tailored to individuals [49]. Furthermore, morality is commonly regarded as impartial, requiring that all individuals be considered equally in moral deliberation.

2.6 Moral Decision-Making

Moral decision-making [3] is a complex cognitive process rooted in the integration of reasoning and affective information within the nervous system [50, 51]. It engages a distributed brain network—including regions responsible for perspective-taking [52], emotional regulation [53, 54], and translating cognitive appraisals into behavior [55]—emphasising action rather than abstract moral deliberation, which collectively work to produce *actionable outcomes*.

By emphasising action rather than abstract moral deliberation, this process ensures that moral cognition is inherently directed toward navigating real-world societal challenges, translating abstract evaluations into practical behavior. The concept of actionable outcomes underscores the practical implications of moral

[3] In philosophy, compound modifiers like "decision-making" are typically hyphenated when used adjectivally to ensure clarity and precision. The hyphen explicitly ties "decision" to "making," emphasizing their joint function as a single concept. "Moral decision-making" aligns with established philosophical and ethical literature, where the term is used to describe the process of evaluating and choosing actions based on moral principles or values. In psychology, "decision-making" is widely recognized as a compound noun, referring to the cognitive process of selecting a course of action. The hyphen is consistently used in research contexts (e.g., "rational decision-making," "emotional decision-making"). Here "Moral decision-making" is preferred to avoid ambiguity, as "moral decision making" could suggest a looser connection between the act of deciding and the process of making, which could lead to interpretive issues.

decision-making processes, transcending abstract moral deliberation to yield measurable behaviors and tangible results. In this framework, the integration of cognitive and affective processes culminates in observable actions, forming a crucial bridge between moral psychology and computational analysis. This approach aligns directly with our second research question (Q2), which explores whether moral decisions leave discernible, machine-detectable behavioral cues. By investigating how moral choices manifest as physical traces, we extend the traditional boundaries of moral psychology into the domain of Computational Machine Ethics. This not only provides empirical grounding for the theoretical frameworks discussed but also establishes a novel methodological foundation for validating human moral behavior theories through technological means.

We should briefly point out that the term *moral cognition* above refers to the broader set of cognitive and affective processes involved in understanding, evaluating, and reasoning about moral concepts, dilemmas, and principles [53, 56]. In contrast, moral decision-making is the *specific process* by which moral cognition culminates in the selection of a course of action in response to a moral scenario [57]. While moral cognition encompasses abstract deliberation, perspective-taking, and the integration of emotional and rational inputs, moral decision-making functions as its *actionable output*, where abstract cognitive evaluations are transformed into concrete behaviors [58, 45].

Both processes stem from a broader psychological framework that governs moral thought and behavior. This framework can be understood through two distinct but interrelated constructs: *moral functioning* and *moral agency*.

Moral functioning refers to the cognitive, affective, and behavioral mechanisms that enable moral understanding and action[59, 60]. It represents the psychological architecture underlying moral cognition and decision-making, encompassing processes such as moral sensitivity, reasoning, and emotional regulation. In essence, moral functioning provides the internal structure that allows individuals to recognize moral dilemmas, evaluate their ethical dimensions, and formulate responses. However, moral functioning alone does not necessarily lead to moral action. Moral agency, on the other hand, is the capacity to act intentionally and responsibly in moral contexts[61, 62]. While moral functioning supplies the necessary cognitive and emotional structures for moral thought and behavior, moral agency actualises these capacities by incorporating self-awareness, autonomy, and accountability in decision-making. A morally functioning individual may recognize an ethical issue and deliberate on its implications, but moral agency is what allows them to translate this deliberation into socially and ethically accountable action.

The relationship between these constructs can be understood through an analogy: if moral functioning is the engine of a car, moral agency is the driver. The engine provides the necessary mechanisms for movement (cognition, affect, and decision-making), but it is the driver who exercises intentional control, making conscious decisions about direction, speed, and response to external conditions. Thus, moral functioning is a prerequisite for moral agency, providing the cognitive and emotional foundation upon which autonomous and accountable moral actions are built.

In this integrated framework, moral cognition provides the evaluative struc-

ture upon which moral decision-making operates, while moral functioning and moral agency ensure that these processes are both internally coherent and socially actionable [63]. At the intersection of moral cognition and moral decision-making lies moral agency, which enables individuals to engage in morally relevant reasoning and behavior. This interplay between cognitive structures, decision-making processes, and individual agency reflects the dynamic and contextual nature of human morality [64].

2.7 Moral Decision-Making: Neural Evidence for Its Practical Nature

Arguably, the term moral-decision making has often been misunderstood as an abstract exercise in philosophical reasoning outside neurospychological context, likely due to the connotations of the term "moral" that accompanies the tuple "decision-making" [65, 58]. Some published work show that this misconception often reduces the discussion around moral cognitive processes to a purely theoretical endeavor divorced from their practical nature [66]. However, moral decision-making is fundamentally rooted in cognitive processes that prioritise actionable outcomes. Its primary function is not merely to deliberate abstractly but to translate moral evaluations into behaviors that enable agents to navigate complex real-world scenarios.

Advancements in neuroscience, particularly the development of functional neuroimaging techniques like Functional Magnetic Resonance Imaging (fMRI) in the early 1990s [67, 68], have provided robust empirical evidence to support this perspective. By enabling non-invasive visualization of brain activity through blood-oxygen-level-dependent contrast, fMRI has revolutionised our understanding of brain functions in vivo [67]. Studies employing these methods can help identified significant overlaps between neural circuits involved in moral reasoning and those associated with practical, action-oriented decision-making. Several lines of evidence found in the published work suggest that there are key regions clearly implicated might include:

a) the medial prefrontal cortex,

b) the posterior cingulate cortex,

c) superior temporal sulcus (STS),

d) precuneus/posterior parietal cortex,

e) orbitofrontal cortex, and

f) the inferior parietal lobule

When analyzed collectively, these regions, which will be referred to by their Brodmann Area (BA) classifications, reveal a pattern: moral decision-making is not an abstract exercise but a cognitive process intrinsically oriented toward producing actions. Cross-analyses of their functional roles, the social pathologies arising from damage to these regions, and their integrative neural linkages might underscore this conclusion [69]. This growing empirical literature thus provides a compelling basis for reframing moral decision-making as a process designed for practical, societal applications rather than theoretical abstraction.

Medial prefrontal cortex

In particular, the medial prefrontal cortex (commonly associated with BA 9/10) might serves as a critical neural hub for integrating cognitive and emotional processes that underlie moral decision-making and guide practical behavior. Much of the literature since early 2000s have shown that this region supports higher-order cognitive functions, including decision-making, planning, and behavioral regulation, by integrating information about current goals, contextual cues, and potential future outcomes to facilitate adaptive, goal-directed behavior [70, 71]. Its role in maintaining and updating working memory enables individuals to weigh competing priorities and select optimal courses of action in real-world contexts [71]. Functional imaging studies further highlight its involvement in processing complex social information, such as inferring others' intentions and assessing the contextual appropriateness of responses—capabilities essential for coordinated, intentional actions [70]. Importantly, this region mediates the integration of abstract moral concerns into practical decision-making, directly linking moral reasoning to actionable, goal-directed behavior [69]. By bridging moral reasoning with behavior, the medial prefrontal cortex exemplifies how moral cognition operates as a process of practical reasoning.

Posterior cingulate cortex and STS

The posterior cingulate cortex and the posterior superior temporal sulcus (STS)/inferior parietal lobule have been widely recognized in the literature as playing critical roles in integrating memory, perception, and social cognition to support practical, goal-directed actions. Evidence from functional neuroimaging studies suggests that the posterior cingulate cortex facilitates the retrieval of episodic memories and integrates them with contextual information, enabling individuals to simulate and evaluate potential actions and their outcomes [72, 73]. As noted by Maddock [72], its connections with the hippocampal formation and parahippocampal cortex provide a foundation for translating memory-based evaluations into adaptive decision-making.

In parallel, the posterior STS and inferior parietal lobule have been identified as central to processing socially significant dynamic stimuli, such as biological motion and intentional actions, which are critical for inferring others' goals and guiding contextually appropriate motor responses [74]. Together, these regions are frequently described as integrating episodic memory, spatial reasoning, and social perception to generate informed, actionable behaviors. Greene [69] has emphasized how this network links evaluative processes with real-world decision-making, allowing individuals to simulate the social and ethical consequences of their actions. This body of research collectively establishes that these neural structures form a critical basis for moral decision-making, reinforcing its characterization as a practical, adaptive process.

The precuneus

The precuneus, also referred to as the posterior parietal cortex, has been extensively studied for its role in integrating spatial, sensory, and social information,

which are critical for action-oriented moral decision-making. Previous research has highlighted its involvement in visuospatial processing, mental imagery, and perspective-taking, functions that support the simulation and evaluation of actions within complex, real-world scenarios [69, 75]. For instance, Damasio [76] has shown that this region integrates self-referential thought and contextual information, enabling individuals to anticipate the social and ethical consequences of their actions.

Functional neuroimaging studies frequently report consistent activation of the precuneus during tasks involving moral reasoning and simulations of potential outcomes, as noted by Moll and colleagues [77, 75]. These studies suggest that its activation patterns overlap significantly with those observed in regions associated with social cognition, such as the superior temporal sulcus, indicating a broader network dedicated to integrating social and spatial cues into adaptive responses. Additionally, Damasio [76] emphasizes that this region supports attentional control, allowing individuals to evaluate competing priorities and align their behaviour with social norms. Notably, the precuneus appears to bridge abstract moral considerations with practical reasoning by integrating information from both memory systems and perceptual mechanisms, enabling dynamic decision-making across varying contexts.

Collectively, the literature positions the precuneus as a key node in the neural networks underlying moral reasoning, particularly in its ability to connect visuospatial cognition with moral deliberation and facilitate actionable, goal-directed responses [69, 78]. This capacity to integrate multiple streams of cognitive and affective input underscores its indispensable role in transforming moral reflection into practical, context-sensitive behaviour.

The orbitofrontal cortex

The orbitofrontal cortex, commonly associated with Brodmann Areas 10 and 11, has been consistently identified in the literature as a critical integrative hub for transforming abstract moral considerations into practical behavioural outcomes. This region plays a central role in real-world decision-making processes, particularly in reward-based decision-making and behavioural inhibition, which are essential for evaluating the social and ethical consequences of potential actions. As noted by Moll and colleagues [79, 77], the medial and lateral subdivisions of the orbitofrontal cortex are pivotal in linking actions to their respective social and environmental outcomes, with the medial subdivision specialising in moral and social dimensions of decisions. These functions are underpinned by neural representations of reward and punishment, which guide behaviour in complex, socially charged scenarios [69].

Functional imaging studies frequently highlight the orbitofrontal cortex's engagement in explicit moral judgment tasks and its involvement in autonomic and emotional regulation during morally relevant decision-making [80, 76]. This dual role of cognitive evaluation and emotional regulation reinforces its integrative function in aligning abstract moral reasoning with actionable behaviours. Moreover, evidence from clinical studies underscores the consequences of damage to this region, which often results in impaired social decision-making, diminished moral sensitivity, and inappropriate behaviours [76, 79]. Such findings provide

compelling support for its essential role in ensuring that ethical reasoning translates into socially appropriate behavioural outcomes.

Taken together, the literature positions the orbitofrontal cortex as a key neural structure in moral decision-making, enabling individuals to navigate ethical challenges within dynamic social environments [69]. Its ability to synthesise cognitive, emotional, and social inputs into adaptive, context-sensitive behaviours exemplifies the practical and action-oriented nature of moral cognition.

The superior temporal sulcus and inferior parietal lobule

FP: Minor overlaps with STS paragraph abpve. Stramline to solve.

Unlike the previous discussion, which focused on broader functional networks and integrative roles (see page 19), this section highlights the specific contributions of the superior temporal sulcus (STS) and inferior parietal lobule, frequently associated with Brodmann Area 39. These regions are examined together due to their complementary roles in processing dynamic social cues and integrating perceptual and conceptual information, both of which are central to linking moral reasoning with practical, action-oriented behaviour [75].

The superior temporal sulcus (STS) and inferior parietal lobule have been widely recognised in the literature for their pivotal roles in perceiving and interpreting dynamic social cues, such as biological motion, gaze direction, and intentionality [81, 82]. These regions are integral to the social cognition processes that underpin moral reasoning, as they enable the detection and interpretation of socially significant movements and intentions. As highlighted by Greene [69], the STS and inferior parietal lobule facilitate the direct linkage between moral judgments and action-relevant social cues, ensuring that moral evaluations inform contextually appropriate responses in real-world scenarios.

Functional neuroimaging studies reveal that the STS is uniquely positioned to process biological motion and goal-directed actions, acting as a computational hub for inferring others' intentions and mental states [81, 75]. Moll et al.[75] further demonstrate that the STS collaborates with regions such as the medial prefrontal cortex to translate social cues into moral evaluations, reinforcing its role in facilitating practical, socially guided behaviour. Simultaneously, the inferior parietal lobule, as emphasised by Decety[82], is crucial for representing intentional actions and integrating sensory and motor information, enabling moral reasoning to bridge abstract ethical considerations with concrete, action-oriented outcomes. This dual functionality ensures that both regions work in tandem to support the seamless transition from moral deliberation to behavioural execution.

Cross-analyses of neuroimaging studies and lesion-based evidence further highlight how these regions interact within a broader network underpinning moral cognition. For example, Greene et al.[69] observed that tasks requiring moral judgment consistently activate the inferior parietal lobule in concert with the STS, suggesting their joint involvement in integrating perceptual and conceptual information for adaptive decision-making. Additionally, evidence from Lane et al.[83] indicates that disruptions in these regions due to damage or pathology can impair social cognition, leading to deficits in empathy and moral sensitivity. Such findings underline the indispensable role of the STS and inferior parietal lobule in facilitating contextually appropriate moral decisions.

Together, these regions exemplify the practical nature of moral decision-making, as they allow individuals to process dynamic social environments and produce adaptive behaviours. Their capacity to synthesise sensory, social, and cognitive information not only supports the detection of morally salient cues but also ensures that moral reasoning leads to actionable outcomes in ethically complex scenarios. This integrative function situates the STS and inferior parietal lobule at the core of a neural network dedicated to the real-world applicability of moral cognition [69, 81, 82, 83, 75].

The collective insights derived from the literature reviewed foreground moral decision-making as a fundamentally social and action-oriented process, deeply embedded in practical reasoning. Rather than being purely reflective or rational, moral cognition emerges as an intuitive and dynamic mechanism, where evaluative processes are seamlessly integrated with contextual, emotional, and social cues to guide adaptive behaviours. Each neural region examined contributes to this framework by prioritising actionable outcomes over abstract deliberations, thereby aligning moral decision-making with its practical and context-sensitive nature.

The medial prefrontal cortex exemplifies this integrative function by harmonising affective and cognitive signals to support goal-directed actions. Similarly, the posterior cingulate cortex and superior temporal sulcus underscore the role of memory and social cognition in situating moral choices within a broader context of lived experience. These regions, in tandem with the precuneus, enable individuals to simulate potential actions and anticipate their outcomes, embedding moral deliberation within the spatial and social fabric of human interaction. Meanwhile, the orbitofrontal cortex transforms abstract ethical principles into behaviourally relevant responses, mediating between competing priorities and the demands of real-world scenarios. The superior temporal sulcus and inferior parietal lobule, critical for interpreting dynamic social cues, ensure that moral reasoning remains sensitive to the social and ethical intricacies of everyday life.

Crucially, this synthesis reinforces a social intuitionist perspective of moral reasoning, wherein decisions are not solely the product of deliberative rationality but are grounded in pre-reflective, affect-laden intuitions that are subsequently shaped by cognitive and social processes. This challenges traditional rationalist views, suggesting instead that moral reasoning is often a post hoc rationalisation of intuitions arising from evolved neural mechanisms designed to prioritise social cohesion and practical action.

Moreover, this framework sets the stage for exploring the dichotomy between teleological and deontological reasoning. The empirical evidence supports the view that moral decision-making is not limited to rigid rule-following (as in deontological ethics) but reflects the adaptive flexibility of teleological reasoning, where decisions are oriented toward achieving ethical outcomes in specific, context-dependent scenarios. This aligns with the Aristotelian notion of practical reason, where moral reasoning is inherently purposive, aiming to resolve ethical challenges in complex and dynamic social environments.

By bridging neuroscience with the normative theories of practical and teleological reasoning, this synthesis not only advances our understanding of moral cognition

but also highlights the fundamentally social and intuitive nature of moral decision-making. This perspective has significant implications for applied fields such as artificial intelligence, where designing morally intuitive systems necessitates prioritising action-oriented and socially adaptive ethical frameworks.

with little consensus about a precise definition [84, 85].

Differences in research approaches to moral decision-making, informed by various theories and perspectives, have led to a discrepancy in the definitions of its nature. This complexity arises from the interplay between cognitive, emotional, social, and cultural factors, each of which plays a significant role in shaping moral judgments [86]. Dual-process theories highlight this interplay, suggesting that moral decisions involve both rational, deliberative processes and fast, intuitive judgments. For example, the Social Intuitionist Model emphasizes the primacy of emotional responses, with reasoning often serving as a post hoc justification for decisions, rather than their origin.

Situational and contextual influences further complicate the definition of moral decision-making. Cultural norms, social pressures, and environmental factors significantly impact what is considered moral, making it challenging to create a universal framework for understanding these decisions. Additionally, individual differences, such as personal values, upbringing, and prior experiences, add to the variability in how morality is processed cognitively. This diversity is reflected in neuroscientific evidence, which shows that moral decision-making activates a distributed network of brain regions associated with reasoning, emotion, and social cognition, illustrating its integrative and context-dependent nature.

Moreover, developmental and evolutionary considerations demonstrate that morality is not static but evolves over time. Developmental theories, such as Kohlberg's stages of moral development, show how cognitive and emotional maturity interact to shape moral reasoning. From an evolutionary perspective, moral behaviors are thought to have developed as adaptive mechanisms for promoting cooperation and social cohesion, blending innate instincts with learned cultural norms. These perspectives highlight that moral decision-making cannot be reduced to a single cognitive domain but rather encompasses an interplay of diverse influences.

The inherent multidimensionality of moral decision-making also poses challenges for researchers attempting to measure and define it. The overlap of cognitive processes with emotional and social dimensions often results in conflicting definitions and findings across disciplines. As a result, moral decision-making remains a dynamic and multifaceted process that resists simplification, reflecting its deeply rooted integration within human cognition, emotion, and society.

This conceptual challenge becomes even more pronounced when examined through philosophical terms, where debates persist to date about the intricate relationship [85] between:

- *first-order moral truths*: in the realm of moral philosophy are propositions that directly govern or describe the ethical principles for example, statements such as "It is wrong to harm

others without justification" or "Justice demands equal treatment for all" are considered first-order moral truths because they articulate specific moral norms or values without delving into the underlying frameworks or theories that validate their authority

- *duties*, or values: obligations or responsibilities that arise from moral, legal, or social principles, guiding individuals to act in specific ways deemed necessary or appropriate, that ought to guide human action, where:
- *ought* refers to normative imperatives indicating how individuals are morally or rationally required to act, based on principles of ethical reasoning.

together with the metaphysics of morality, and the practical application of these in real-life situations.

Moreover, different philosophical, and psychological, schools of thought offer diverging perspectives on the role of key core concepts for defyingn precisily moral decision making such as:

- Moral facts: Some emphasise the direct perception of moral facts [87, 88, 89], while others argue for more complex reasoning processes in situations with novel perplexities and conflicting moral considerations [32, 90, 91].
- Moral principles: Some find moral principles essential for reasoning [32, 91], while others, like particularists, argue for the existence of moral reasons independent of any general principle [92, 93, 94].
- Moral psychology: Differing views on the role of emotions and motivations in shaping reasoning also contribute to the lack of a unitary definition.

Definition 2 (Rational Model)**.** *Moral decision making, is the cognitive process of choosing between competing moral judgments* i.e.*, mutually exclusive evaluations we make on what is right or wrong, good or bad, and that we use as motive, purpose and direction for our conscious, and practical behaviour.*

While widely varying definitions of the term moral have been suggested, a precise definition of moral decision making has proved elusive [84, 85]. The deficition we adopt here our own version as presented above. This version is a first particular case of a more general version of the definition which we will introduce at the end of this chapter, after discussing its components. Moral decision making embobies a multitude of which we will only introduce for completeness:

a) **Cognitive Process** This term refers to the mental actions or operations that individuals use to acquire knowledge and understanding. It includes processes such as perception, memory, reasoning, decision-making, and problem-solving. Cognitive processes are essential for interpreting and interacting

with the world [95, 96, 97, 98, 99];

b) **Behaviours:** In academic terms, behaviours are the observable actions or reactions of an individual in response to external or internal stimuli. These actions can be voluntary or involuntary and are influenced by various factors, including cognitive processes, emotions, and environmental conditions [100, 101, 102]

and more complex process *i.e.*,

Moral Reasoning: Theories in Philosopy

In an exclusively and formally philosophycal acceptance, provided *not* in any particular chronological order, Jonathan Dancy, a prominent moral particularist, argues that "reasons holism", the idea that a reason in one case may be different or absent in another, supports philosophical position known as Moral Particularism according to which moral reasoning does not rely on fixed principles but instead depends on the context-specific evaluation of reasons [103]. At its core, particularism rejects the idea that the morally perfect person is the "person of principle." Instead, it posits that moral sensitivity requires recognizing how features of a situation contribute to moral judgments in contextually variable ways. Dancy illustrates this with the "holism of reasons," which argues that a reason that counts in favor of an action in one situation might count against it or hold no relevance in another. For example, promising to do something is generally a reason to act, but this reason can be nullified or inverted in unique cases, such as when the promise itself is immoral or coerced [104]. The practical implications of particularism are significant. It suggests that moral agents must focus on the specific details of a situation rather than applying pre-set rules. This approach allows for greater flexibility and responsiveness to moral complexity. For example, a rule prohibiting lying might fail to account for scenarios where lying protects an innocent life, while a particularist framework would weigh the specific reasons at play and decide accordingly [92]

FP: ...but in a what that the computer science reader can ... or anything that glue the next portion of text

The key aspect of the traditional philosophycal discourse on the topic is to characterised moral reasoning as a specific form of practical reasoning, emphasizing its role in guiding moral actions and resolving conflicts, operating as the process through which agents figure out what they *morally ought to do* in concrete situations. Unlike theoretical reasoning, which seeks to understand principles or truths, moral reasoning directly addresses the question: "What should I do?" [105] and frames moral reasoning as practical reasoning because it resolves *real-life dilemmas* by weighing competing moral values and considerations. For instance, deciding whether to prioritize loyalty to family or a broader duty to society exemplifies moral reasoning in action. This deliberative process of moral reasoning is, indeed, *practical* in two essential respects. Firstly, it pertains to the *domain of action*, as its focus is on resolving questions related to what *should be done*. Secondly, it is practical in its *outcomes*, as the act of reflecting on matters of action inherently guides and motivates individuals toward acting [84]. A natural way to interpret this point of view is to contrast it with the standpoint of

theoretical reason.

Theoretical reasoning is concerned with resolving questions that are fundamentally theoretical rather than practical in nature. It focuses on explanation and prediction, retrospectively asking why events have occurred and prospectively determining what might happen in the future. Paradigmatic expressions of theoretical reasoning are found in the natural and social sciences, where causal relationships and empirical evidence are central [106, 107, 108, 109]. Beyond this, theoretical reasoning also extends to *non-causal explanations*, such as those explored in metaphysical, logical, and conceptual inquiries. Its focus on understanding matters of fact is distinct in its impersonal and *universally accessible approach*, which contrasts with the more situated and action-oriented focus of practical reasoning. Practical reasoning, by contrast, centers on deliberation about what one *ought to do*, providing guidance for action in specific circumstances. Unlike theoretical reasoning, which seeks to understand the world as it is, practical reasoning addresses how to navigate complex situations and achieve goals. It operates across a wide range of contexts, from moral obligations, such as promoting well-being or fulfilling promises, to professional domains like medicine, scientific experimentation, artistic creation, or athletic performance. Practical reasoning is also tied to how-to knowledge and technical skill, enabling individuals to adapt to dynamic environments and deliberate effectively about actions.

Theoretical reasoning, in this sense, reflects on what reasons justify accepting particular claims as true. Its focus is on evidential considerations that indicate the likelihood of propositions being correct. Practical reasoning, by contrast, deliberates on what makes actions desirable or worthy of choice. Its reasons are those that justify actions as worth performing. This divergence in subject matter also leads to different outcomes: theoretical reasoning modifies one's belief system by aligning it with truth, whereas practical reasoning culminates in action. As previously noted, practical reasoning is tied to action not only in its subject matter but also in its purpose and results.

Figure 2.1: Diagram illustrating the types of reasoning, distinguishing between theoretical and prescriptive reasoning.

Lastly, it should be said that despite their differences, theoretical and practical reasoning share an underlying structural unity. Both involve evaluating reasons, assessing justification, and adhering to principles of rationality, albeit in different domains. This shared framework is evident in our common vocabulary: we *speak* of what is "right to do" and "right to think," of being "justified" in both actions and beliefs, or of having reasons for what we choose or conclude. This structural unity reveals how the forms of reasoning mirror one another, even as their substantive concerns diverge—whether focusing on understanding facts, guiding action, or grounding belief.

Historically, the term moral reasoning as been used in philosophical discourse to describe any deliberative thinking that responsibly engages with moral considerations, to arrive at *actionable conclusions*. Unlike purely theoretical ethics,

moral reasoning directly influences practical outcomes, often navigating conflicts between competing moral principles or obligations [105]. This practical orientation ensures that moral reasoning is treated not merely as an abstract exercise but is tied to real-world actions and decisions. The philosophical importance of moral reasoning lies in its capacity to handle conflicts of values and integrate competing moral considerations into coherent courses of *action*.

Finally, the source questions whether moral reasoning is truly distinct from practical reasoning more generally. Different moral theories offer different answers. Aristotle, for example, saw a unified structure in practical reasoning, with the virtuous person simply having their sights set on the true good. Kant, however, saw moral reasoning as distinct, involving considerations of universalizability that aren't present in prudential reasoning. Given this diversity, the source suggests remaining agnostic on the relationship between moral and non-moral practical reasoning unless one assumes the correctness of a particular moral theory.

However, there are challenges. Empirical studies suggest we often reason poorly in moral situations. We can be 'dumbfounded' when asked to justify our moral intuitions and are susceptible to biases. This could lead to revisions in our norms of moral reasoning, a point the source acknowledges.

This process involves an array of cognitive functions such as perception, memory, reasoning, and problem-solving, which collectively inform our moral evaluations and decisions [58, 110, 111]. Moreover, these cognitive processes translate into behaviours, which are the *observable* manifestations of our moral choices. These behaviours, whether conscious or subconscious, reflect our internal moral deliberations and are influenced by a complex interplay of cognitive functions, emotions, and contextual factors. A number of recent experimental work have explored how cognitive attributions of intentionality influence moral judgments and subsequent behaviors, demonstrating the observable translation of moral cognition into action.

In [110], Antoine Bechara et.al., explores the neurobiological underpinnings of decision-making through the lens of the somatic marker hypothesis. This framework posits that decision-making is shaped by marker signals, which arise from bioregulatory processes—including emotions and feelings, and guide behavior by associating specific outcomes with somatic states. These processes are mediated by a network of cortical and subcortical structures, with the ventromedial prefrontal cortex (VMPC) playing a pivotal role.

Central to the hypothesis is the idea that decision-making is influenced by both conscious and unconscious processes. While reasoning and knowledge about options are critical, emotional responses provide biases that help narrow the decision-making space. These biases, often experienced as gut feelings or anticipatory emotions, act as alarm signals, highlighting advantageous or disadvantageous outcomes and allowing rapid, context-sensitive decisions.

The role of the VMPC in this system is particularly significant. Damage to this region disrupts the ability to generate somatic markers, leading to impairments in social and moral behavior. Patients with VMPC lesions exhibit an inability to

make advantageous decisions, particularly in complex, uncertain scenarios. The article discusses this through the "gambling task," a laboratory simulation that mimics real-life decision-making by factoring in reward and punishment. While healthy participants learn to select options with favorable long-term outcomes, VMPC patients persistently choose options offering immediate rewards despite long-term losses, demonstrating a "myopia for the future."

The inability of VMPC patients to generate anticipatory physiological responses, such as skin conductance changes, underscores the importance of somatic markers in decision-making. In healthy individuals, these anticipatory responses develop before conscious knowledge of advantageous choices, suggesting that unconscious emotional signals precede and guide cognitive reasoning. VMPC patients, in contrast, fail to produce these signals, even when they consciously understand the consequences of their choices. This dissociation highlights the critical role of emotion in shaping decisions, beyond purely rational calculations.

The article also distinguishes decision-making processes from related cognitive functions, such as working memory. While the dorsolateral prefrontal cortex (DLPFC) is essential for maintaining and manipulating information, the VMPC's primary function lies in linking knowledge about situations with emotional valence. This distinction is supported by evidence that VMPC patients with intact working memory still exhibit impaired decision-making, emphasizing the specialized contribution of emotional processing to ethical and practical deliberations.

From a broader perspective, the findings reinforce the argument that moral and practical decision-making behaviors are not solely products of rational deliberation but are profoundly influenced by the interplay of cognitive and emotional systems. The somatic marker hypothesis provides a compelling explanation for how unconscious emotional biases, shaped by prior experiences, dynamically interact with conscious reasoning to produce decisions. These processes align with the claim that moral deliberations and behaviors reflect a complex integration of cognitive functions, emotions, and contextual factors.

By identifying the neural substrates and mechanisms of these interactions, the article advances our understanding of decision-making as a holistic process. It underscores the necessity of considering emotional and somatic components in ethical theory, legal frameworks, and practical applications, such as treatment for decision-making deficits in clinical populations or the development of artificial intelligence systems capable of emulating human moral reasoning.

Churchland's work [111] supports the claim that moral deliberations and behaviors are influenced by a complex interplay of cognitive functions, emotions, and contextual factors. By grounding morality in neurobiological processes, Churchland demonstrates how moral decision-making reflects the integration of innate capacities with cultural and environmental influences. This perspective aligns with broader interdisciplinary efforts to understand morality as a dynamic and adaptive feature of human cognition, with implications for philosophy, psychology, and practical ethics.

There is a large volume of published empirical studies describing the guiding stenght that moral decision-making has on practial actions, as observable be-

havioural course of actions, from agents on their enviroment. In works such has [112, 113] Greene provides a detailed exploration of how moral decision-making operates as a practical function grounded in brain mechanisms, supported by empirical evidence and demonstrates that moral reasoning is not isolated to specific brain regions but emerges from the interplay of multiple systems tasked with value representation, cognitive control, and emotional regulation (largely discussed later on in this chapter).

FP: needs changes in the phrasing, and connection to the rest of the chapter. Maybe start using "guiding behaviour?

FP: This part relates to emotional dimension of moral decision-making.

The authors argue that moral decision-making is *inherently practical*, aimed at resolving conflicts between competing values and guiding behavior. This practicality is exemplified in dilemmas such as the trolley problem [114], where decisions require balancing the harm to one person against the benefit to others. Through functional magnetic resonance imaging (fMRI) studies, Greene and colleagues have shown that "personal" moral dilemmas (e.g., pushing someone off a bridge to save others) engage emotion-related areas like the amygdala and ventromedial prefrontal cortex (vmPFC). By contrast, "impersonal" dilemmas (e.g., pulling a lever) activate regions like the dorsolateral prefrontal cortex (DLPFC), associated with controlled reasoning and cost-benefit analysis.

The article emphasizes that moral decision-making extends beyond theoretical principles, manifesting in observable behaviors shaped by brain function. Altruistic actions, such as donating to charity, involve the frontostriatal pathway, which integrates social cues and rewards. Similarly, cooperative behaviors are supported by neural mechanisms that encode trust and reciprocity, demonstrating how moral reasoning translates into real-world practices.

FP: important for our case

Pharmacological studies also provide practical insights. For instance, increased serotonin levels, which amplify emotional reactivity, enhance deontological judgments, while decreased emotional engagement promotes utilitarian reasoning. These findings underscore the brain's adaptability in modulating moral behavior based on biological and environmental factors.The neuroscientific evidence supports the view that moral decision-making is not only about abstract reasoning but also deeply embedded in the practical, embodied realities of human life.

Hence, moral decision making encompasses both

1) the mental operations that guide our judgments, and
2) the resultant (observable) actions that embody our moral principles in the practical realm.

Here we are concerned with point 2 above, *i.e.*, the resultant *observable* actions that, as such, we hypotesize being a physical traces in terms of observable, machine detectable behavioural cues of an agent's moral decision.

The perception of direct gaze, that is, of other individual gaze directed at the observer, is know to influence a wide range of cognitive processes and behaviours.

Practical reason refers to the distinctively human capacity to determine, through reflective deliberation, the appropriate course of action in a given situation [28, 115, 116].

In this context, reflective deliberation is a cognitive function often characterised by conscious, effortful, and reason-guided evaluation of options, or courses of action. It usualy characterised in the literature as involving:

1) **Philosophical Perspective:** The capacity to critically assess one's desires, beliefs, and values, often engaging in second-order thinking to evaluate not just what one wants but whether those wants align with broader principles or long-term goals. Rooted in Kantian ethics and Aristotelian practical reasoning, it emphasizes autonomy and rationality [28, 117].
2) **Psychological Perspective:** A metacognitive process where individuals engage in controlled, systematic thinking to weigh evidence, consider alternatives, and predict outcomes. Reflective deliberation contrasts with automatic or heuristic-driven decision-making, drawing from dual-process theories of cognition [99, 118].

This deliberative process in practical reasoning is, indeed, *practical* in two essential respects. Firstly, it pertains to the *domain of action*, as its focus is on resolving questions related to what *should be done*. Secondly, it is practical in its *outcomes*, as the act of reflecting on matters of action inherently guides and motivates individuals toward acting [84].

People may act differently in public environments due to actual reputation concerns, or due to the mere presence of others. People often act differently when others are observing them, reputational concerns and signaling are widely theorized to be a driving mechanism explaining why people become more prosocial and moral when observed in public.

One of the two main objective of this work is to determine wheater the presence of social robots can affect the outcome of *moral decisions* made by humans in controlled, experimental settings.

In the discourse regarding evolution of moral theories across philosophy and modern psychology, it emerges a nuanced interconnection where *emotional* and *rational* elements not only diverge but also integrate in explaining how agents takes moral decisions. This interconnection reveals the intricate complexities inherent in the formulation of moral models, transcending a mere dichotomy to embrace a more holistic, undefined perspective.

Definition 3 (Emotional Model)**.** *Moral decision-making is an emotive process, wherein individuals navigate and choose between competing moral judgments* i.e.*, mutually exclusive evaluations we make on what is right or wrong, good or bad. This process is driven by emotional responses and intuitions, which guide and inform our conscious and practical behaviour, often preceding and shaping cognitive deliberation.*

Moral decision-making represents a cognitive exercise in the calculus of ethics. Within this framework of moral calculus, contemporary and classical scholars offer a spectrum of perspectives on the central role of emotional and cognitive faculties alike that incorporates both emotional and cognitive faculties.

It is worth delineating the concept of 'moral calculus' as distinct from *hedonistic*

calculus. While the latter term typically refers to Benthamite utility maximization, often quantified in terms of pleasure and pain, 'moral calculus' serves as a broader framework for ethical deliberation. Unlike hedonistic calculus, which is generally rooted in consequentialist traditions, moral calculus navigates the complexities of diverse ethical systems, be they deontological, virtue-based, or others.

The philosophical canon profoundly integrates the conceptual distinction between emotion-driven and reason-driven moral philosophies. This differentiation has seen considerable evolution over centuries, leading to a significant impact on contemporary psychological thinking, especially in the sphere of moral psychology. The nuanced separation of emotion-driven and reason-driven moral frameworks, deeply rooted in philosophical discourse, has evolved extensively over time, culminating in its marked influence on modern psychological studies, with a particular focus on moral psychology. This journey begins with the foundational works of ancient philosophy. In this era, thinkers like Plato in his 'Republic' delineate a clear preference for reason over emotion in guiding ethical conduct. Aristotle, in his 'Nicomachean Ethics,' echoes this sentiment to some extent by emphasizing rational virtues, yet he also acknowledges the significant role emotions play in ethical existence.

Moving forward into the Enlightenment, this conceptual distinction was further crystallized. A paradigmatic figure of this era, Immanuel Kant, championed a morality firmly rooted in reason and universal maxims in his works, such as 'Critique of Pure Reason' and 'Groundwork for the Metaphysics of Morals,' thereby relegating emotions to a subsidiary role. This period marked a significant shift toward a rationalist perspective in moral philosophy.

However, this shift was met with a counterpoint in the British empirical tradition. Figures like David Hume presented a challenge to the Kantian rationalism. In his 'A Treatise of Human Nature,' Hume provocatively posited that reason is subordinate to passions, thereby anchoring moral judgments in emotional responses. This perspective from British empiricists highlighted the importance of emotions, or 'sentiments', in moral considerations, offering a contrasting view to the prevailing rationalist approach.

Moral decision making, is a cognitive process of choosing between competing moral judgments- *i.e.*, mutually exclusive evaluations we make on what is right or wrong, good or bad, and that we use as motive, purpose and direction for our conscious, practical behaviour.

Moral decision-making is primarily an emotive process, wherein individuals navigate and choose between competing moral judgments *i.e.*, mutually exclusive evaluations we make on what is right or wrong, good or bad. This process is driven by emotional responses and intuitions, which guide and inform our conscious and practical behaviour, often preceding and shaping cognitive deliberation.

Hence, in the discourse of moral calculus, certain schools of thought, notably those propounded by Haidt (2012) and Greene (2007), assert with compelling vigour that the substratum of emotional faculties, rather than those of the cognitive domain, governs the architecture of ethical decision-making. This viewpoint

finds a harmonic resonance in the philosophical canon, corroborated by seminal treatises such as Nussbaum's 'Upheavals of Thought' (2001) and Damasio's 'Descartes' Error' (1994).

The theoretical edifice of moral calculus, while intellectually robust, gains tangible relevance when juxtaposed with empirical and phenomenological data. Bridging these domains allows for a more encompassing understanding of moral decision-making, marrying the abstract with *the* concrete, *the* theoretical with *the* experiential.

For the empirical aspect:

Neuroscientific research offers valuable empirical insight into the machinery of moral cognition. Studies have implicated regions like the prefrontal cortex and the amygdala in the ethical decision-making process (Greene et al., 2001; Decety & Cacioppo, 2012). These findings suggest that our 'moral calculus' may indeed have a tangible neurological substrate, grounding ethical theory in the biological realm.

For the phenomenological aspect:

Complementing these empirical observations, phenomenological accounts provide a subjective lens through which moral decision-making can be examined. Authors such as Sartre and Merleau-Ponty have explored the existential dimensions of choice, capturing the lived experience of moral deliberation (Sartre, 1943; Merleau-Ponty, 1945).

These works serve to enrich our understanding of 'moral calculus' by infusing it with the subjective quality of human experience.

Emotion-Centered Models: Some theories argue [**?**] that emotional processes, rather than cognitive ones, are at the core of moral decision making. Your definition may not adequately capture the emotive factors often considered essential. onathan Haidt's work in "The Righteous Mind" explores the role of emotional intuition in moral judgments, arguing that reasoning often follows, rather than guides, our moral intuitions [45]

Practical behaviour is a term widely used across philosophy and psychology, it's challenging to create an exhaustive chronological definition because the term does not correspond to a singular theory or concept that has evolved over time in a linear fashion. Instead, it has been interpreted and applied differently depending on the context, theoretical framework, or school of thought.

Practical behaviour in philosophy: has been interpreted in various ways across different philosophical schools of thought. 1) In Aristotelian philosophy, practical behaviour is associated with "praxis" or action guided by moral virtue aimed at the good life. Practical wisdom ("phronesis") is crucial here as it guides one's decisions and actions in accordance with moral virtue. 2) Immanuel Kant distinguished between theoretical reason (used to understand the natural world) and practical reason (used to govern behaviour and moral decision-making). For Kant, practical behaviour is guided by the categorical imperative, an absolute moral law. 3) In the late 19th and early 20th century, the pragmatists (like

William James and John Dewey) viewed practical behaviour as action informed by the effects that such behaviour would bring about.

Practical Behaviour in Psychology has been understood as observable actions and reactions to stimuli in behaviourism, deeply intertwined with internal cognitive processes during the cognitive revolution, and as a complex interplay of cognitive processes, emotional states, individual traits, and environmental influences in contemporary psychology. 1) In the behaviourist approach (Early 20th Century) pioneered by John Watson and B.F. Skinner, practical behaviour is understood in terms of observable actions and reactions to stimuli, often studied through conditioning processes. 2) With the cognitive revolution (Mid-20th Century), practical behaviour started to be seen as deeply intertwined with internal cognitive processes like problem-solving, decision-making, and planning. 3) Social-Cognitive Theory (Late 20th Century): Albert Bandura's social-cognitive theory emphasised the role of observational learning, self-efficacy, and goal setting in practical behaviour. **Modern times**: today, in ethics and action theory, practical behaviour typically refers to behaviour guided by practical reason, that is, reason concerned with action and decision-making. This involves deliberation about means and ends, moral obligations, and the values at stake in different courses of action. Similarly in Contemporary Psychology, practical behaviour is understood as a complex interplay of cognitive processes, emotional states, individual traits, and environmental influences. It is typically studied in context-specific terms, such as health behaviour, consumer behaviour, or prosocial behaviour.

This is quite an encompassing scope, but there are inevitable aspects of the broader discourse on moral decision making that we need to include in this purview.

— the following needs to be integrated in the text —

The term "practical behaviour" is a broad one, encompassing a wide range of actions that an individual might take in their everyday life. These can range from simple behaviours like brushing teeth or driving to work, to more complex ones like making a significant decision about one's career or personal life."Moral behaviour," on the other hand, is a subset of practical behaviour. It refers specifically to actions that involve moral or ethical considerations. In other words, all moral behaviours are practical behaviours, but not all practical behaviours are moral behaviours. The distinguishing feature is the presence of moral or ethical considerations in the motivations, implications, or consequences of the action. So, in response to your question, the key to understanding the difference between "practical behaviour" and "moral behaviour" does indeed lie in understanding the specific meaning and implications of "behaviour" in these contexts. However, it's also crucial to consider the specific nature and context of the action itself, including the intentions behind it and its potential consequences. the key to understanding the difference between "practical behaviour" and "moral behaviour" does indeed lie in understanding the specific meaning and implications of "behaviour" in these contexts. However, it's also crucial to consider the specific nature and context of the action itself, including the intentions behind it and its potential consequences. Moral domain: A behaviour typically falls within the moral domain when it pertains to questions of right and wrong, fairness,

justice, harm, and welfare. So, for instance, deciding to donate to charity falls within the moral domain because it involves considerations about the welfare of others. Playing the piano, on the other hand, would generally fall outside the moral domain because it's largely a personal interest or skill, not directly associated with the welfare or rights of others. Intention and motive: Moral behaviour often involves a level of intentionality, where the individual acts with a specific purpose or motive that is morally charged. An individual who donates to charity with the motive of helping others is engaging in moral behaviour. In contrast, an individual who plays the piano for personal enjoyment is engaging in a practical behaviour that isn't inherently moral or immoral. Consequences: The potential or actual impact of behaviour on others also plays a crucial role in determining its moral status. Behaviours with positive or negative impacts on others are often evaluated on a moral basis. **Additional notes**: 1) *Interaction of factors determining behaviour:* In both philosophy and psychology, behaviour is viewed as a result of a complex interplay of multiple factors, including cognitive processes, emotional states, individual traits, and environmental influences. 1) Cognitive processes: Cognitive processes, such as perception, memory, decision-making, and problem-solving, play a critical role in practical behaviour. For example, decision-making theories, such as the Dual Process Theory, suggest that people use both intuitive (automatic, fast, and emotional) and deliberative (slow, controlled, and logical) systems in guiding their behaviours [38]. 2) Emotional states: Emotions can also guide our behaviour. The James-Lange theory of emotion suggests that our emotional experiences are a response to our bodily reactions to a stimulus. For example, we don't run away because we're afraid; instead, we're afraid because we see ourselves running[39]. 3) Individual traits: Personality traits influence how individuals interpret and respond to their environment. The Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism) have been linked to various behavioural outcomes. For instance, high levels of conscientiousness have been associated with better job and academic performance[13]. 4) Environmental influences: Social and physical environments shape behaviour. Social Cognitive Theory emphasises the reciprocal nature of this relationship: our behaviour can both influence and be influenced by our environment. For example, observational learning suggests we learn behaviours by observing others, while self-efficacy can determine how we respond to challenges[9]. Modern psychology acknowledges that many behaviours are driven by processes outside of conscious awareness. For instance, implicit bias research shows that we often harbour unconscious biases that can influence our behaviour, including decision-making and interpersonal interactions[21].Decision-making is not always rational and is often influenced by cognitive biases. For example, the 'availability heuristic' suggests that people are more likely to consider information that's easily retrievable when making decisions, which may not always lead to accurate or optimal outcomes[42].This interdisciplinary field combines psychology and economics to understand decision-making and behaviour. For example, the concept of 'nudge theory' suggests that subtle changes in how choices are presented can significantly influence decisions and behaviour, a principle that has been applied in various domains like healthcare, finance, and public policy[43]. These insights suggest that our understanding of practical behaviour needs to be multifaceted, taking into account not only conscious, deliberate processes but also unconscious influences and the way cognitive biases and heuristics shape our

decisions and actions. They also underscore the importance of considering the individual within their social and environmental context

Moral decisions theories are often analysed into components features such as the model of judgment adopted- whether factual or normative, rational of affects laden- its causes, and the ethical outset it seems to follow. All three components happen to be useful for identifying and organise methods and work done in Computational Ethics since they are easily linked to different scientific approaches adopted in the field, and their basis deeply root into both philosophical and psychological theories which have deeply inspired and implicitly shaped the objectives set for this field in the past two decades.

In particular, most modern philosophers have frequently written about the conflict between factual and normative judgments [119], between reason and emotions [88], and between normative and motivating reasoning [26] three dichotomies.

2.7.1 Normative Non-Ethical agents

A moral decision is what we *judged* necessary to resolve conflicts with an explicitly moral dimension via special type of judgements which often called *normative* or *value judgements*: responses to stimuli with a moral dimension. Normative judgements assert or deny what *ought* to be the case whether or not it is *actually* the case (see figure 2.2, page 35), in contrast to factual judgements which assert or deny facts that *are the case* or a properly justified believe.

Factual judgements assert or deny facts that *are the case* or a properly justified believe, while normative judgements assert or deny what *ought* to be the case whether or not it *actually* is the case.

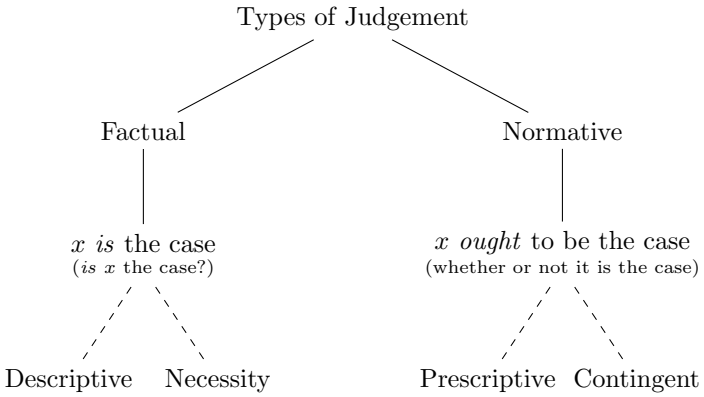


Figure 2.2: This distinction presuppose a sufficient prior understanding of the relevant uses of *is* and *ought* (or *should*) which will not discuss in details here. It is important to notice that, the presence of such marks as is neither a sufficient nor necessary criterion for the distinction we 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*.

Judgements such as 'copper is a metal' [119] or '$2 + 2 = 4$' express what is the case because they are *truth-apt* judgements which means that they are either true or false in there being some corresponding *fact* which settles the question of their truth value. In simple terms, we cannot have different opinions on '$2 + 2 = 4$' because there exists a system of mathematical principles that, if accepted, makes us committed to the believe that '$2 + 2 = 4$': there are corresponding *facts* that make these locutions true or false.

On the other hand, judgements such as 'innocents ought not to be punished' [119] or 'wrongful killing is always wrong' are judgments about what *ought* to be the case but that do not have any corresponding fact that makes it true or false. Can machines grasp the difference between the two? In contrast with other more empirical judgments, moral judgements seem to have an intrinsic connection to motivation and action, for they form in us a uniquely bonding intentions to perform a behaviour, and motivate us to act in accordance with it [120].

Moor in [15] was one of the fist to examine how this distinction is relevant for a predominate class of works in Machine Ethics. Moor noticed that ordinary computers are designed with a purpose in mind, they are *normative agents* in the sense that they perform something on our behalf, executing rule-based instructions of which efficacy can be assessed assessed. However, ethical agents are those that perform actions with an ethical impact (positive or negative), but not by being constrained by their designers as this would not count has ethical act by the the very same definition of Ethics.

2.7.2 Other

Hence, genuine moral decisions must have as 'end-product', actions or inactions. In The Language of Morals [41], R.M. Hare, one of the leading British moral philosopher of the twentieth century, gives this clear characterisation:

> If we were to ask of a person "What are his moral principles?" the way in which we could be most sure of a true answer would be by studying what he did. He might, to be sure, profess in his conversation all sorts of principles, which in his actions he completely disregarded; but it would be when, knowing all the relevant facts of a situation, he was faced with choices or decisions between alternative courses of action, between alternative answers to the question "What shall I do?", that he would reveal in what principles of conduct he really believed. The reason why actions are in a peculiar way revelatory of moral principles is that the function of moral principles is to guide conduct. ([41])

By the same token, the main objective of Machine Ethics is to develop implicit ethical agents that is to say, machines that have been programmed in a way that can decide on actions with an ethical impact on their environment. Machine Ethics revolves around a precise subset of decision-type, since not all decisions have a moral dimension, and therefore not all types of judgments are relevant to morality. For example, whether I should get a frosty cold drink on a hot day using my last pound is not a moral decision. Whether I should use my last pound to get a cold drink, or give it to the women begging for money, appears to be.

Both are instances of decision concerned with actions, they drive *goal-oriented behaviours* in which our perceptual and memory system support decisions that determine our *actions* [**?**].

2.8 Extended Types of Judgments

Typically, judgments are classified into two primary types: factual (descriptive or necessary) and normative (prescriptive or contingent). However, these categories, though fundamental, may not fully encompass the range of judgments humans engage with. Various disciplines, including philosophy, psychology, and mathematics, suggest other classifications or subcategories.

While factual and normative categories provide a foundational classification of judgments, the diversity and complexity of human thought suggest the utility of additional categories. The appropriateness of any specific set of categories, however, depends on the nature of the subject matter and the research questions at hand.

1. **Value Judgments:** Value judgments focus on the worth, importance, or intrinsic merit of a subject. As a subset of normative judgments, they often pertain to ethical or moral dimensions. However, their unique emphasis on 'value' might warrant a separate consideration.

2. **Aesthetic Judgments:** Aesthetic judgments concern beauty or other aesthetic attributes. Although they might be regarded as a form of value or normative judgments, the discipline of aesthetics often treats them as a distinct category due to their specialised focus.

3. **Prudential Judgments:** Prudential judgments, often used in economics, decision theory, and practical ethics, consider what is prudent or practically wise. These judgments typically involve an interplay of both descriptive and normative elements.

4. **Probabilistic Judgments:** Probabilistic judgments, prevalent in statistics, psychology, and decision theory, assess the likelihood or probability of a given event or condition. They often require a balance between empirical data and theoretical models.

5. **Counterfactual Judgments:** Counterfactual judgments, commonly used in philosophy and cognitive psychology, speculate on alternate realities or conditions. These judgments often hinge on the ability to imagine and reason about hypothetical situations.

6. **Analytic Judgments:** In Kantian philosophy, analytic judgments are those in which the predicate concept is included within the subject concept. These judgments are typically tautological and contrast with synthetic judgments.

7. **Synthetic Judgments:** Kant also proposed synthetic judgments, wherein the predicate concept is not contained within the subject concept. They can be classified further into a priori (based on reasoning independent of experience) and a posteriori (based on experience).

for example, the judgments made in physics, like those in other scientific disciplines, can be seen to fall into several categories depending on the specific context. Much of the work in physics involves making *descriptive (factual) judgments* about the nature of the physical world. These judgments are usually based on observation and experimentation and aim to accurately describe how the world is. While less common in physics than in other fields such as ethics, *prescriptive (normative) judgments* are sometimes made in the context of methodological rules about how to do physics. Physics often involves making *probabilistic judgments*. In quantum mechanics, the behaviour of particles is often described in terms of probabilities rather than definite outcomes. Physicists also frequently make *counterfactual judgments*, considering what would happen under different hypothetical scenarios. The distinction between *analytic and synthetic judgments* is also relevant in physics. An example of an analytic judgment in physics might be a mathematical truth that holds by definition within a certain model, while a synthetic judgment might be a statement about the physical world that is supported by empirical evidence.

The judgments made in physics, like those in other scientific disciplines, can be seen to fall into several categories depending on the specific context. Drawing upon Chalmers' work on the philosophy of science [14], we find that much of the work in physics involves making *descriptive (factual) judgments* about the nature of the physical world. These judgments are usually based on observation and experimentation and aim to accurately describe how the world is. While less common in physics than in other fields such as ethics, *prescriptive (normative) judgments* are sometimes made in the context of methodological rules about how to do physics, a concept explored by Laudan in his work on normative naturalism [15]. Physics often involves making *probabilistic judgments*. In quantum mechanics, the behaviour of particles is often described in terms of probabilities rather than definite outcomes [12]. Physicists also frequently make *counterfactual judgments*, considering what would happen under different hypothetical scenarios, a concept explored in the work of Woodward [16]. The distinction between *analytic and synthetic judgments* is also relevant in physics. Drawing on Bird's exploration of Kuhn's philosophy [11], we see that an example of an analytic judgment in physics might be a mathematical truth that holds by definition within a certain model, while a synthetic judgment might be a statement about the physical world that is supported by empirical evidence. In practice, many judgments in physics may involve a combination or an interplay of these types. The specific context and objectives of the work play a large role in determining which types of judgments are most relevant.

In practice, the types of judgments made in physics often involve a mixture of these categories. For instance, a descriptive judgment about the behaviour of a particle might be based on a combination of observation (a synthetic judgment) and mathematical reasoning (often involving analytic judgments). Thus, the understanding and classification of judgments in physics, like in other fields, benefit from a nuanced approach.

In a field such as Computer Science, a discipline that often intersects with logic, mathematics, and engineering, several types of judgments can be identified. Much of the work in computer science involves making *descriptive (factual) judgments*.

These often take the form of specifying the behaviour of algorithms or systems, such as a judgment about the time complexity of a particular sorting algorithm. *Prescriptive (normative) judgments* are also found in computer science, often relating to best practices for coding, architectural decisions in system design, or ethical considerations in AI development. *Analytic judgments*, where the predicate is contained within the subject, often emerge from logical deductions that follow from the definition of a concept or operation. *Synthetic judgments*, which refer to empirical findings that don't just follow from definitions, might involve observations about the performance of certain algorithms in specific contexts. Especially in areas like machine learning and algorithm analysis, computer scientists often make *probabilistic judgments*, like assessing the probability of a certain outcome given a set of inputs. In troubleshooting, system design, or in planning the development process, *counterfactual judgments* often play a significant role as computer scientists consider alternate scenarios or possibilities.

The rise of fields such as AI ethics and Human-Computer Interaction (HCI) has brought attention to *value judgments* in computer science. These might concern what constitutes fair treatment in an algorithm's decision-making process, for example. In practice, many judgments in computer science may involve a combination or an interplay of these types. The specific context and objectives of the work play a large role in determining which types of judgments are most relevant.

2.9 A definition of judgment

So, what is *judgment*?

A *judgment* has been defined differently across various fields. From a logical and mathematical perspective, it carries specific interpretations. In formal logic, a judgment is typically understood as an assertion that a proposition is true. This idea can be represented as follows:

$$J(P)$$

Here, J denotes the judgment operation and P is a proposition. The entire expression, $J(P)$, is read as "*it is judged that* P".

In mathematics, a judgment can be considered akin to a function. If we think of a judgment as mapping from a set of premises to a conclusion, we can represent it in a similar way to a function:

$$J : P \to C$$

Here, J is the judgment, P represents the premises, and C is the conclusion. This can be understood as a judgment J mapping a set of premises P to a conclusion C. Note, however, that this is a rather abstract and non-standard interpretation. Judgments in mathematics and logic are more typically represented as statements or propositions that are asserted to be true. German logician Gottlob Frege's work in the field of logic provides valuable insight into the concept of judgment. His Begriffsschrift, or concept script, was a formal language of logic devised to represent clear, logical thoughts. In Frege's system, judgments about a proposition can be symbolically expressed and manipulated.

3.

3.1

4.

4.1

5.

5.1

6. Moral Displacement: An Experimental Investigation

6.1 Conceptual Foundations of the Research Question

This chapter begins with a precise question: *can the silent presence of a humanoid robot alter the evaluative process that turns moral perception into action?*

This question, while operationally simple, reaches beyond behavioural measurement. It engages the broader project of understanding moral behaviour not merely as an individual trait but as an inferential process that emerges from the perception and decoding of socially meaningful signals—**a process that can, in principle, be computationally modelled**.

Within the domains of social signal processing and artificial intelligence, the transformation of subtle environmental cues into behavioural outputs is treated as a mapping from informational stimuli to structured responses [121]. By embedding a humanoid robot—ontologically ambiguous, semantically potent, yet behaviourally inert—into a morality-salient environment, this experiment asks whether such synthetic presences perturb not the content of deliberation, but the signal-to-inference architecture through which salience becomes action.

Question 6.1: *Inferential Displacement*

Can the mere presence of a synthetic, non-agentic entity perturb the inferential transformation through which morally salient cues are converted into observable moral behaviour?

In other words, the question asks whether the mere fact of a robot's presence—despite the absence of task-related communication or instruction—can alter the evaluative mechanism that translates moral perception into moral behaviour, operationalised here as prosocial giving.

In this experiment, moral action is instantiated through a measurable behavioural outcome: the voluntary donation of part of the participant's monetary compensation to a children's medical charity. The humanoid robot introduced into the experimental environment is not interactive in any directive or conversational sense, but neither is it inert. Operating in autonomous life mode, NAO exhibits subtle embodied motions—simulated breathing, minor postural adjustments, and head orientation shifts triggered only when participants establish eye contact. These micro-movements constitute precisely the minimal behavioural cues known to activate or modulate the Watching Eye effect, thereby rendering the robot a semantically potent, low-agency observer within the moral field. By examining whether the presence of such a humanoid robot systematically shifts donation behaviour, we test whether synthetic co-presence perturbs not the participants' reflective moral reasoning, but the **conditions under which morally salient**

cues elicit prosocial action.

In other terms, the inquiry asks whether the presence of a humanoid robot—endowed not with communicative capacity but with minimal yet perceptually salient behavioural affordances—can alter the evaluative pathway through which moral perception becomes moral behaviour, operationalised here as **prosocial giving**.

In this experiment, moral action is instantiated through a measurable behavioural outcome: the voluntary donation of part of the participant's monetary compensation to a children's medical charity. The inquiry therefore isolates *presence* itself—specifically, synthetic presence—as an informational and epistemic variable. It examines whether introducing such a form into a morality-salient environment alters the **situational conditions under which moral action is produced**. Crucially, the experiment does not attempt to model or infer the internal structure of moral reasoning; rather, it observes how the resulting behavioural expression of moral decision-making shifts across environments that differ only in the presence or absence of this subtly animated robot. In this way, the design tests whether synthetic co-presence perturbs not the content of deliberation, but the **conditions under which morally salient cues become behaviourally actionable**.

Framing the investigation as a *question* (Question 6.1 p. 43) rather than a hypothesis is deliberate. It preserves the conceptual openness required at this stage of the analysis, foregrounding inquiry over prediction. Within interdisciplinary research—spanning moral psychology, social signal processing, and human–robot interaction—prematurely imposing a directional hypothesis risks presupposing the very moral effects that the experiment is designed to probe. By articulating a guiding research question rather than an asserted claim, we allow the empirical structure of the data to shape the inferential trajectory rather than constraining it in advance. This is consistent with both the methodological caution urged in philosophy of science and the epistemic humility appropriate when dealing with morally charged, psychologically subtle, and technologically novel forms of social influence.

Against this backdrop, the central inquiry of the study can be expressed with complete clarity: *does the mere presence of a humanoid robot alter how human beings act when confronted with a morally relevant choice?*

Put operationally, we ask whether individuals donate differently to a charitable cause when a robot quietly shares the room with them. The behaviour of interest—**prosocial giving**—is quantified directly as the amount of money voluntarily deposited into a charity box. The variable is simple in measurement but dense in interpretive significance: the coins themselves index the culmination of a moral appraisal process, the behavioural footprint of an evaluative transformation triggered under conditions of minimal social prompting.

Yet the stakes of this question extend beyond monetary donation. What is under scrutiny is whether artificial companions—even in the absence of agency, speech, intention, or social engagement—can modulate the conditions under which morally salient cues are converted into human action. In this respect, the study examines not only how much participants give, but *why* behaviour may

shift under synthetic co-presence. The possibility being tested is subtle but far from trivial: that the introduction of an ontologically ambiguous entity into a moral environment may refract the participant's evaluative landscape, thereby altering the behavioural expression of moral choice.

6.2 Experimental Design and Behavioural Paradigm

To investigate this Question 6.1 (see p. 43), we implemented a controlled behavioural experiment [122, 123, 124] derived from the classical *Watching Eye* paradigm [125, 126, 127, 128, 129, 130, 131], in which prosocial behaviour is modulated by the perceived presence of observation. Seventy-three participants were invited individually into a room under the pretext of completing a battery of personality questionnaires in exchange for monetary compensation. Embedded in the experimental space was a morally salient cue: a charity brochure prominently featuring the photograph of a child requiring medical assistance. Decades of empirical work show that such stimuli reliably activate prosocial dispositions through mechanisms of implicit monitoring and empathetic engagement [132, 133].

The robotic manipulation was then introduced as the sole experimental variable. In the control condition, participants completed the task alone. In the experimental condition, a humanoid robot—NAO [134]—remained silently present in *autonomous life mode*, exhibiting only the minimal embodied cues characteristic of that configuration: simulated breathing, subtle postural adjustments, and reactive head orientation triggered exclusively by eye contact.

These micro-movements, though non-interactive and devoid of communicative intent, constitute precisely the class of minimal behavioural affordances shown to activate or modulate the mechanisms underpinning the Watching Eye effect. By embedding this low-agency, perceptually salient entity into an otherwise identical moral environment, the design isolates *synthetic presence*—rather than dialogue, instruction, or overt agency—as the only *manipulated* dimension of the setting. The personality questionnaires, administered under the pretext of a trait study, simultaneously serve as a cover story and as a structured measurement of individual cognitive–affective profiles. In subsequent analyses, these trait measures are treated as moderators, allowing us to ask whether any observed differences in prosocial donation behaviour arise from the robot's presence alone, from stable individual dispositions, or—critically—from their interaction within a shared moral field.

6.2.1 Why Minimal Presence Matters: Ontological Ambiguity as Experimental Variable

Much of the literature on moral decision-making in human–robot interaction (HRI) and human–machine interaction (HMI) locates moral modulation in the interactive capacities of artificial agents. Studies routinely foreground expressive behaviour, ostensive cues, adaptive responsiveness, displays of accountability, or anthropomorphic signalling as the levers through which machines influence human judgment and behaviour [135, 136, 137, 138, 139]. These approaches implicitly assume that moral impact requires action: verbal behaviour, communicative intent, social reciprocity, or strategically framed moral cues.

The present experimental design intentionally refuses this assumption.

Rather than examining how robots act, we examine how they exist—that is, how their mere ontological presence, stripped of communicative intent and devoid of interactive complexity, may nevertheless perturb the inferential transformation through which morally salient cues become behaviourally instantiated. The focus is not on moral agency or synthetic ethics, but on the structural susceptibility of human moral cognition to ontologically ambiguous stimuli.

This methodological divergence is conceptually foundational. It allows us to target an aspect of moral cognition that is often overlooked: its *pre-reflective permeability* (for a similar use of the term refer to [140, 141, 142, 143]) to agent-like cues even when those cues lack *intentional content* [144, 145, 146]. The question is not whether robots can engage in moral exchange, but whether their presence, by virtue of their bodily form and minimal behavioural affordances, reshapes the inferential scaffolding that mediates between perceiving a moral cue and acting upon it.

This problem is particularly salient in domains such as Social Signal Processing and computational social cognition, where synthetic agents routinely evoke social and moral reactions that exceed the informational complexity of their behaviour [121, 147]. By removing dialogue, task-relevance, and overt interaction while maintaining the perceptual markers of potential agency (eyes, posture, orientation, micro-motion), the experiment isolates **presence itself** as the epistemic variable to be tested.

In this respect, the design probes a structural vulnerability of norm-sensitive cognition: the possibility that minimal cues—mere *indications* of agenthood—may exert disproportionate influence on evaluative pathways. The robot is not required to speak, gesture, or respond; its semantic force lies in its ability to activate interpretive priors associated with observation, evaluation, and social monitoring.

This intuition resonates with the hyperactive intentional stance described by Guthrie [148], Waytz et al. [149], and Dennett [150], according to which humans routinely over-ascribe agency in uncertain environments. By positioning the robot in the liminal space between objecthood and agenthood, the experiment isolates not action, but anticipation—the silent priors that precede full agentive recognition.

The methodological focus on **mere presence** thus reflects a principled decision: it disentangles interactive contingencies from deeper, subpersonal cognitive mechanisms that structure moral evaluation. Unlike approaches that equate moral influence with dialogue or reciprocity, this design foregrounds the epistemic topology of moral salience—the latent structures of social attribution that shape inferential pathways prior to action, prior even to conscious appraisal.

Having established the necessity of minimal presence as an experimental variable, the next conceptual step is to formalise the framework that renders this presence epistemically potent. This is where Floridi's Levels of Abstraction (LoA) become essential: they provide the philosophical infrastructure required to explain why *an entity that does nothing*, and to which no moral status is attributed, may still distort the conditions under which moral cues become behaviourally actionable.

This motivates a transition, not from theory to application, but from conceptual architecture to **experimental justification**.

6.2.2 Levels of Abstraction and the Design Logic of Minimal Robotic Presence

The decision to deploy a humanoid robot in silent autonomous life mode—exhibiting only simulated breathing, subtle postural adjustments, and eye-contact-contingent head orientation—is not a matter of convenience or technological limitation. It is a philosophical and methodological choice grounded in Floridi's theory of *Levels of Abstraction* (LoA) [151, 152, 153]. To appreciate this decision, the core function of LoAs must be understood with conceptual precision.

An LoA specifies the informational interface through which an agent, system, or observer accesses and processes the world. It determines which distinctions are epistemically visible and which are systematically bracketed. LoAs are therefore not metaphysical: they make no assertions about the intrinsic ontology of entities. Rather, they are *epistemic configurations*, selective filters that carve out what counts as relevant information.

Applied to the present experiment, LoAs allow us to describe moral influence without relying on metaphysical accounts of robot agency. At the LoA operative for a participant alone in a room, moral relevance does not depend on the robot's internal states but on its semantic affordances: its posture, its eyes, the symmetry of its body, the direction of its face, its quiet imitation of biological rhythms [154, 155, 156, 157, 158, 159, 160, 161].

These features are perceptually encoded as possible indicators of being watched [154, 162, 163, 155, 157, **?**, 164, 165], evaluated, or accompanied—precisely the conditions under which the Watching Eye effect operates. Thus, the robot's moral relevance emerges not from consciousness, autonomy, or interactive capacity, but from its informational presentation within the participant's operative LoA.

This perspective enables a shift away from essentialist distinctions—agent versus non-agent, sentient versus non-sentient—toward a functional reading: what does the robot *do* at the LoA of the observer? At this LoA, NAO's subtle bodily cues instantiate the informational signatures of a putative observer, thereby modulating the epistemic background against which morally salient cues (such as the charity poster) are evaluated.

The placement of the robot in autonomous life mode is therefore a purposeful calibration of informational affordances. If NAO were fully interactive, the LoA would shift, and the participant would be required to adopt an intentional stance grounded in dialogue, reciprocity, or social coordination. *This would confound the experiment by introducing behavioural and communicative variables.* Conversely, if the robot were completely inert—akin to a mannequin—the LoA would strip away most agent-like affordances, nullifying the minimal conditions under which moral salience can be perturbed

NAO therefore occupies a deliberate middle space: a synthetic presence endowed with minimal but meaningful cues, sufficient to activate the epistemic structures

associated with potential observation but insufficient to produce interactive interpretation. In this capacity, NAO aligns with Floridi and Sanders' notion of an *artefactual moral agent* [166, 153]: a non-sentient entity whose moral relevance arises not from autonomy but from the role it plays within an informationally structured environment.

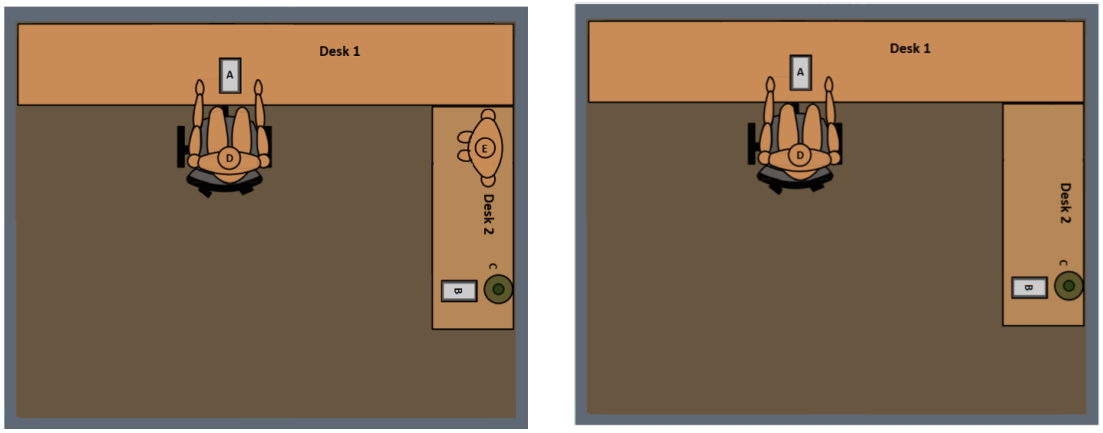
FP: This is more a conclusion.

In short, Floridi's LoA framework explains why a non-interactive, subtly animated robot is an epistemically potent variable. It provides the philosophical rationale for a design in which robotic presence functions as a **semantic perturbation** of the evaluative pathway from moral salience to moral action. Presence is not a passive attribute; it is an informational act.

This reading supports both the minimalist structure of the experimental design and its philosophical depth. By rejecting behavioural or dialogic criteria for moral influence, and grounding the analysis in semantic encoding at the LoA of the observer, we avoid naïve assumptions about interaction as a prerequisite for moral modulation. Presence, when correctly encoded, can reframe what is morally visible—prior to deliberation, and independent of interaction.

6.2.3 Experimental design and Preliminary Results

To investigate Question 6.1, we implemented a controlled behavioural experiment [122, 123, 124] derived from the classical *Watching Eye* paradigm [125, 126, 127, 128, 129, 130, 131], in which prosocial behaviour is modulated by implicit cues of observation. Each participant was invited individually into a room under the pretext of completing a personality-study session in exchange for monetary compensation. Unbeknownst to them, the experimental environment contained a morally salient stimulus: a charity brochure displaying the photograph of a child requiring medical care. Decades of empirical work demonstrate that such stimuli reliably trigger prosocial dispositions by activating implicit monitoring and empathetic engagement [132, 133].



(a) Experimental condition: robot present.

(b) Control condition: robot absent.

Figure 6.1: Top-down view of experimental vs. control configurations. Both settings retain identical spatial and visual layouts, isolating the variable of robotic presence as the only ontological difference.

Participants were randomly assigned to one of two conditions. In the **Control** condition, they completed the questionnaires alone. In the **Robot** condition, a humanoid NAO robot was placed in the room and operated in autonomous life mode. Although NAO emitted no speech and performed no task-relevant actions, it displayed minimal embodied behaviours—simulated breathing, subtle postural adjustments, and head-orientation responses triggered only by eye contact. These micro-cues are the minimal behavioural affordances known to activate or modulate the Watching Eye effect.

After completing the questionnaires, each participant received £10 in £1 coins as compensation and encountered a voluntary donation opportunity. An opaque charity box (Operation Smile) was positioned near the exit. Participants could donate any subset of the coins. The total donation served as the primary dependent measure of prosocial behaviour.

Initial results revealed a robust directional pattern: participants in the Robot condition donated substantially less than those in the Control condition. Furthermore, no meaningful between-group differences were found in personality profiles (Empathizing Quotient [167], Systemizing Quotient [168], Big Five Inventory [169]), ruling out trait-based confounds and strengthening the inference that robotic presence itself modulated the evaluative pathway underlying prosocial action.

6.2.4 From Behavioural Setup to Evaluative Structure

In moral philosophy, action is frequently treated as the terminus of deliberation [28, 117, 170]. Yet the present study concerns not the deliberative endpoint but the evaluative transformation that precedes it: the internal process by which morally salient cues are converted into behavioural output [171, 172]. The experimental design above provides the behavioural substrate; what remains is to articulate the evaluative architecture through which robotic presence might exert

its influence.

Our explanatory focus therefore remains firmly on moral action—here, instantiated as voluntary donation—while acknowledging that salience, cognition, and interpretive modulation contribute to the inferential scaffolding that produces such action. This framing connects the experiment to the philosophical traditions of practical reasoning and to the neurocognitive models explored in Chapter 2.

Our aim is not to probe abstract normativity, but to determine whether artificial presence perturbs the transformation from moral appraisal to observable donation—a behavioural manifestation of deliberative judgement.

Empirically, the experiment transposes the Watching Eye paradigm into a minimal social environment co-inhabited by a humanoid robot. Prior variants of the paradigm have relied on stylised pictorial stimuli or supernatural primes [126, 173]. Our design replaces these with an embodied artificial presence whose ontological ambiguity is semantically potent while remaining behaviourally minimal.

To formalise the transformation under investigation, we treat moral action not as a fixed trait but as the output of a cognitive–affective function integrating environmental cues, individual traits, and contextual structure. In philosophical terms, this is the practical realisation of moral salience; in psychological terms, it is the integration of cue perception, affective readiness, and situational inference.

$$\mathbb{E}[f(\Sigma \cup \mathscr{R})] \neq \mathbb{E}[f(\Sigma)]$$

Where:

- Σ is the morality-salient perceptual field (e.g., the Watching Eye stimulus),
- $\mathscr{R}$ is the synthetic co-presence, realised here by NAO,
- $f(\cdot)$ is the evaluative transformation mapping perceptual input to moral behaviour,
- $\mathbb{E}[f(\cdot)]$ denotes the expected behavioural output (donation magnitude).

Read aloud, this expresses the hypothesis that:

The expected outcome of moral behaviour changes when a humanoid robot is present within the perceptual–moral environment.

Hypothesis 1: *Evaluative Deformation Hypothesis*

The expected outcome of moral behaviour, as computed through the evaluative process f, is altered when the robot is present within the perceptual-moral environment.

The conceptual shift from the initial research question to this first formal hypothesis is thus warranted by the structure of the experimental design. The question preserved conceptual openness—*is robotic presence morally perturbative?* The

hypothesis now expresses this inquiry in a form amenable to empirical adjudication, specifying how the evaluative transformation from moral cue to moral action may be deformed.

To make the structure of this transformation explicit, we can decompose the probability of a deviation in moral action into its component determinants:

$$\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R)$$

where:

- α_E encodes the environmental moral cue (here, the Watching Eye stimulus),
- β_C denotes the individual-level control variables (psychometric and demographic structure),
- γ_R represents the robotic presence as a perturbative affordance.

This expression can be read aloud as: *The probability of a deviation in moral decision* (δ_m) *is a function of the environmental moral cue* (α_E)*, the individual's psychological and demographic configuration* (β_C)*, and the presence of the robot* (γ_R).

That is, the probability of observing a change in moral behaviour is a function of: (i) the morally salient stimulus, (ii) the participant's internal traits, and (iii) the synthetic presence that may refract, displace, or attenuate the evaluative process.

This formalism captures the operative logic of the experimental design: moral action is not treated as an isolated datum, but as a context-sensitive transformation of moral salience into behaviour. The robotic presence is therefore not conceptualised as a behavioural actor but as a *topological perturbation*—a variable that reframes the inferential lens through which moral cues are registered and converted into action.

To understand the stakes of this perturbation, we must clarify what is meant by *moral salience*. Across philosophical and psychological literatures, moral salience refers to the capacity of a situation, object, or agent to present itself as morally significant—i.e., to become an object of evaluative attention prior to explicit deliberation [172, 171, 58, 174, 175]. It functions as a phenomenological filter: before the agent reasons, before the agent chooses, certain features of the environment appear as normatively charged. Within this framework, synthetic entities may perturb moral salience not by issuing commands or engaging in dialogue, but by reconfiguring what is foregrounded, what is suppressed, and what is affectively or normatively "seen" in the first place.

This brings us to the ontological dimension of the hypothesis. The robot's influence depends not on its computational sophistication but on its *perceived ontology*: how observers intuitively classify the entity—as object, tool, quasi-agent, or socially charged companion. In this experiment, NAO's embodied form, posture, gaze behaviours, and subtle animations evoke agent-like expectations without satisfying the criteria for full moral agency. This ambiguity is precisely what renders the robot a semantically potent perturbator within the moral field.

Hypothesis 2: *Synthetic Normativity of Moral Displacement*

Synthetic presences, though devoid of sentience, may acquire *normative affordances* by virtue of their perceived ontology. When situated within morality-salient environments, such presences may disrupt, refract, or displace the evaluative machinery through which moral judgments are ordinarily formed.

This hypothesis extends beyond a narrow behavioural prediction; it asserts that robotic presence may alter the normative topology of the environment itself. The experiment is therefore not merely a test of prosocial output, but a constrained act of epistemic staging—a designed moral topology intended to probe whether the presence of $\mathscr{R}$ displaces or refracts the normative force of α_E.

The Watching Eye paradigm thereby becomes a conceptual instrument: not merely a psychological effect but a method for examining the structural elasticity of normative cognition in environments where human agents coexist with synthetic forms. What the study observes, therefore, is not simply differences in donation behaviour, but how the inferential architecture linking salience to action is modulated by synthetic co-presence. Generosity, in this framework, is not a trait but an emergent property of norm-sensitive evaluative systems embedded within a structured environment.

This framing rejects simplified accounts that treat moral behaviour as transparent readouts of internal disposition. Instead, it positions moral action as the contingent result of cognitive–affective systems operating under topological deformation [69, 88, 176]. Robotic presence, by virtue of its ontological ambiguity, functions as a refractive moral affordance: a structural condition that may attenuate or redirect the transformation of moral salience into action.

FP: old content begins

The term *perceived ontology* refers to how observers intuitively classify an entity's nature—whether as object, tool, agent, or something more ambiguous. In this context, it denotes how the humanoid robot is not treated merely as a machine, but as a presence with quasi-social or normatively loaded features. This perception does not require the attribution of full agency or sentience; rather, it is the robot's embodied form, gaze behaviours, and passive co-presence that evoke moral expectations in the observer. Thus, the robot's "perceived ontology" may perturb how moral salience is registered, filtered, or even displaced by human evaluative systems.

FP: old content ends

This is not an experiment in the narrow sense of causal testing. It is a constrained act of epistemic staging—a designed **moral topology** that probes whether the presence of $\mathscr{R}$ displaces, diffuses, or refracts the normative force of α_E. Our aim is not simply to determine whether donations changes under robotic observation, but whether $\mathscr{R}$ alters the internal topology of moral inference itself. In this light, the Watching Eye paradigm ceases to be a psychological curiosity and becomes an instrument of conceptual inquiry: a way of testing the structural elasticity of

normative cognition in post-human social configurations.

What this study observes, therefore, is not simply what participants do under (staged) robotic observation, but how the inferential architecture of moral cognition is perturbed by synthetic presence. The robot, though devoid of agency, functions as a semiotic operator on the moral field—its presence refracts the salience of otherwise normative cues, modulating prosocial output through shifts in interpretive topology. We do not treat generosity as a readout of innate disposition, but as the *emergent property of norm-sensitive evaluative systems embedded in structured environments*.

This framing **rejects** any simplistic account of moral behaviour as noise-free reflection of trait. Instead, we position moral action as the contingent result of *cognitive-affective systems* operating under *topological deformation* [69, 88, 176]. In this view, robotic presence is not merely a contextual feature, but a morally refractive affordance that alters the mapping between cue and action.

Within this epistemological architecture, the following experiment tests the plausibility of a central hypothesis: that robotic presence—by virtue of its ontological ambiguity—can systematically attenuate the conversion of moral salience (see above for a definition) into action. It is this structured possibility, not merely behaviour, that the empirical sections to follow are designed to investigate.

With this architecture in place, the subsequent sections examine how such deformation manifests empirically—first at the behavioural level, and then at the deeper structural level of trait–context interactions.

6.3 Conceptualisation of the Experiment: Moral Decision-Making under Synthetic Presence

Having articulated the evaluative architecture through which synthetic presence may perturb the transformation from moral salience to action, we now specify how this theoretical framework is instantiated empirically. The objective of this section is not merely to describe procedural steps, but to clarify the conceptual rationale that makes this experimental configuration an appropriate test of the inferential deformation thesis established above.

To empirically examine whether the mere presence of a synthetic, non-agentic entity can alter the evaluative pathway underlying charitable behaviour, we embedded participants within a controlled, minimally structured moral choice scenario. Framed as a standard personality study, the procedure unobtrusively positioned each participant before an unannounced ethical decision. This preserved the epistemic opacity required for observing pre-reflective evaluative processes rather than self-presentational behaviour.

Each participant entered the experimental room alone and completed a series of psychometric measures—the Empathizing Quotient [177], Systemizing Quotient [178], and the Big Five Inventory [179]. Completion of the questionnaires served a dual methodological purpose. First, it provided data for assessing whether trait dispositions modulated sensitivity to robotic presence, thereby enabling the analysis of trait–context interactions central to this chapter. Second,

it supplied a plausible pretext for the experimental setting, ensuring that participants approached the environment without anticipating a moral evaluation.

Upon completion, participants received £10 in £1 coins as compensation. Before exiting, they encountered a latent moral choice: an opaque green charity box placed beside a prominently displayed image of a child requiring medical care. The image served as a Watching Eye stimulus—an established elicitor of prosocial behaviour through mechanisms of implicit monitoring, empathy, and reputation sensitivity [132, 133]. Donation behaviour, performed privately and unobserved, constituted the operational measure of moral action.

The central experimental manipulation divided participants into two conditions:

- **Control Condition**: participants completed the task alone in a room containing only the morally salient cue.
- **Robot Condition**: the same room contained a humanoid NAO robot (Aldebaran Robotics) in "autonomous life mode." Although behaviourally non-interactive, the robot exhibited minimal embodied cues—simulated breathing, micro-adjustments in posture, and reactive head orientation triggered exclusively by eye contact. These cues provided precisely the level of embodied salience required to activate or disrupt the cognitive mechanisms underlying the Watching Eye effect.

This configuration preserved identical spatial, visual, and procedural features across conditions; the sole ontological difference was the presence of the robot.

In the experimental condition, a humanoid robot (NAO) was silently positioned in the space, operating in "autonomous life mode": breathing rhythmically, subtly shifting posture, and responding to eye contact through reactive head movement — yet without speaking, interacting, or engaging in any directive behaviour. Importantly, participants had no prior knowledge of the robot's presence, and the robot itself did not intervene in the task.

Importantly, participants were not warned about the robot in advance, and no verbal or task-relevant interaction occurred at any time. The robot therefore functioned as an *epistemic perturbation*: a synthetic presence whose embodied form was salient yet behaviourally inert, occupying the ambiguous space between animate agent and object.

The behavioural outcome was striking: participants in the Robot condition donated substantially less (mean £1.17) than participants in the Control condition (mean £1.89). No significant differences in personality profiles were observed between groups, ruling out trait imbalance and indicating that the observed attenuation of donation reflects a genuine displacement in the evaluative pathway rather than a sampling artefact. At a descriptive level, then, synthetic co-presence appears to weaken the moral force of the Watching Eye stimulus.

To understand why this effect is theoretically significant, we must clarify the status of *moral decision-making* within this experimental architecture. Contrary to utilitarian models that construe donation as a form of preference optimisation (see chapter 3), our framing treats the decision to donate as an instantiation

of *moral salience attribution under epistemic opacity*. Participants do not know they are being observed; they do not know that donation behaviour is the dependent measure; and they do not know that synthetic presence is the variable of interest. What is revealed, therefore, is not explicit moral reasoning, but the *implicit evaluative machinery* through which morally loaded cues gain—or fail to gain—behavioural traction.

The Watching Eye stimulus plays a critical role in this machinery. Anthropological and psychological research shows that images of eyes or children reliably elicit third-party moral concern via affective engagement and implicit audience effects [126, 127, 130]. Our design extends this paradigm by placing, alongside the Watching Eye cue, a humanoid robot whose ontological status is neither human nor ethically inert. NAO thus becomes an *ontological anomalous agent*: a presence that possesses the perceptual affordances of agenthood without the behavioural or normative commitments of actual agency.

This motivates the following hypothesis, which articulates the expected deformation within the evaluative architecture:

Hypothesis 3: *Synthetic Perturbation of Moral Inference*

The humanoid robot NAO does not function as a passive observer, but as a perturbative presence that refracts the transition from moral salience to prosocial action. Its ontological ambiguity displaces the affective-empathic cues that ordinarily support donation, thereby modulating the evaluative pathway by which moral stimuli gain behavioural expression.

$$\mathscr{S} : \Sigma \xrightarrow{\mathscr{R}} \mathscr{D}$$

where:

- Σ denotes the perceptual input space structured by morally salient cues (brochure, child's eyes, spatial configuration),
- $\mathscr{R}$ denotes the synthetic robotic presence functioning as a perturbative modulator,
- $\mathscr{D}$ denotes the domain of observable moral decisions (monetary donation).

In control conditions, the transition $\Sigma \to \mathscr{D}$ proceeds without interference: the affective weight of moral cues is preserved and expressed through prosocial giving [132, 173]. In robotic conditions, by contrast, $\mathscr{R}$ deforms this mapping. It may displace empathic identification, dilute the salience of the Watching Eye cue, reshape the normative topology of the environment, or function as a cognitive decoy [180]. Each interpretation bears distinct implications for the design of ethical robots and for understanding how humans recalibrate moral behaviour in the presence of synthetic others.

6.3.1 Formalisation of Hypothesis and Experimental Logic

The present experiment is best conceived not as a mechanistic probe into behavioral preferences, but as a structured perturbation within a normatively encoded cognitive system. Specifically, it seeks to investigate **how robotic presence modulates human moral decision-making** under conditions of minimal priming and perceptual constraint. Unlike traditional paradigms that treat prosociality as an output of deliberative utility calculus, the design employed here foregrounds the **pre-reflective inferential machinery** that converts perceptual-affective cues into morally salient behavior.

At its epistemic core, this experiment operates as a **perturbative test of moral salience transmission** — that is, whether a morally charged perceptual cue (e.g., the face of a child in need) is successfully converted into a prosocial behavioral output (monetary donation), and how that transmission is modulated, disrupted, or reframed by the passive presence of a **non-agentic but anthropomorphically encoded entity** (*i.e.*, the NAO robot).

To formalize the interpretive structure of this transformation, let us denote:

- Σ: the perceptual-affective input space (including the Watching Eye stimulus, spatial layout, and ambient cues)
- $\mathscr{R}$: robotic presence, ontologically positioned between artifact and agent
- $\mathscr{D}$: the moral decision space (observable as donation behavior)

The operative hypothesis can be expressed as a probabilistic modulation of expected moral output:

$$\mathscr{R} \notin \Sigma \Rightarrow \mathbb{E}[f(\Sigma)] = D_{\text{prosocial}} \quad \text{(Control condition)}$$
$$\mathscr{R} \in \Sigma \Rightarrow \mathbb{E}[f(\Sigma \cup \mathscr{R})] = D_{\text{attenuated}}$$

where:

$$D_{\text{attenuated}} < D_{\text{prosocial}} \quad \text{(Robot condition)}$$

Here, the notation $\mathbb{E}[f(\cdot)]$ denotes the **expected behavioral output** of the cognitive-affective system under a given set of environmental conditions. The function $f(\cdot)$ captures the internal inferential transformation by which perceptual-affective cues—such as the Watching Eye stimulus—are mapped onto discrete moral actions, in this case, the act of anonymous donation. Crucially, the expectation operator $\mathbb{E}[\cdot]$ signals that we are not describing a deterministic relation, but rather the *aggregate tendency* across a psychologically heterogeneous population. It reflects the statistical structure of the behavioral response field rather than individual-level causality.

6.3.2 Methodological Design: Inferring Moral Perturbation through Controlled Artificial Co-presence

To regard an experimental setting as a generator of knowledge, rather than a mere data collection routine, demands that its internal architecture be epistem-

ically justifiable and ontologically transparent. In this respect, every stage of the experimental method presented here is conceived not simply as procedural necessity, but as epistemic filtering: a sequence of deliberate constraints designed to isolate latent variables within the perceptual and normative landscape of the participant.

At its core, the experimental logic operationalises the following proposition:

$$\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R)$$

where:

- δ_m denotes a deviation in moral decision (quantified as donation behavior),
- α_E represents environmental moral cues (Watching Eye),
- β_C indexes control factors (psychometric variables, demographic traits),
- and γ_R captures the effect of robotic presence.

The experimental setting is thus a structured interrogation of whether $\gamma_R \neq 0$ under conditions in which α_E and β_C are held constant or accounted for. If confirmed, such deviation would instantiate a moral displacement: a case in which a non-sentient co-agent modulates human ethical output without any explicit instruction, coercion, or intervention.

FP: add link to relevant hypothesis and check condition "not zero"

The following experimental procedure was implemented to ensure maximal control over environmental affordances while preserving participant naivety concerning the true moral dimension under investigation.

6.3.3 Formalisation of the Experimental Logic

Having established the conceptual and epistemic rationale for investigating robotic co-presence as a perturbative variable, we now formalise the internal logic of the experimental design. The present experiment is not conceived as a mechanistic probe into stable behavioural preferences, but as a *structured perturbation* applied to a normatively encoded cognitive system. Its aim is to examine how a minimally interactive synthetic entity modulates the evaluative transformation through which morally salient cues become behaviourally instantiated.

Unlike paradigms that construe prosociality as the downstream product of deliberative utility calculus, our design foregrounds the **pre-reflective inferential machinery** responsible for converting perceptual–affective moral cues into action. In this frame, moral behaviour is not treated as a direct expression of preference or disposition, but as the output of a cognitive–affective transformation whose parameters may be refracted by the presence of an ontologically ambiguous entity.

At its epistemic core, the experiment operates as a **perturbative test of moral salience transmission**: whether the moral charge embedded in a Watching Eye stimulus is preserved, attenuated, or reframed when a synthetic presence occupies the same perceptual field. The robot deployed in this study—non-agentic,

behaviourally minimal, but anthropomorphically encoded—functions precisely as such a perturbative variable.

To make this structure explicit, let us denote:

- Σ: the perceptual–affective input space (Watching Eye stimulus, spatial layout, ambient cues),
- $\mathscr{R}$: the robotic presence, ontologically positioned between artefact and agent,
- $\mathscr{D}$: the moral decision space, operationalised as monetary donation.

The operative hypothesis concerning the effect of robotic presence can be expressed as a modulation of expected moral output:

$$\mathscr{R} \notin \Sigma \Rightarrow \mathbb{E}[f(\Sigma)] = D_{\text{prosocial}} \quad \text{(Control condition)}$$
$$\mathscr{R} \in \Sigma \Rightarrow \mathbb{E}[f(\Sigma \cup \mathscr{R})] = D_{\text{attenuated}} \quad \text{(Robot condition)}$$

with the expected attenuation constraint:

$$D_{\text{attenuated}} < D_{\text{prosocial}}.$$

Here, $\mathbb{E}[f(\cdot)]$ denotes the **expected behavioural output** of a cognitive system embedded within a particular perceptual–normative configuration. The evaluative function $f(\cdot)$ captures the internal inferential process by which morally salient cues—such as the image of the child beneficiary—are mapped onto the act of anonymous donation. The use of the expectation operator signals that this relation is *statistical rather than deterministic*, reflecting the aggregate structure of a psychologically heterogeneous population. The experiment thus examines whether the presence of $\mathscr{R}$ shifts the distribution of moral output at the population level, not whether it dictates individual choices.

6.3.4 Methodological Architecture: Inferring Moral Perturbation through Structured Artificial Co-presence

To regard an experiment as a generator of epistemic insight rather than a mere data collection mechanism, its procedural structure must be internally justified and ontologically transparent. The methodological architecture adopted here is therefore not a set of neutral steps, but a sequence of *epistemic filters*: constraints designed to isolate the variables that may participate in the evaluative transformation from moral cue to moral action.

At the heart of this design lies the formal proposition:

$$\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R)$$

In experimental terms, the logic is straightforward: the design isolates the contribution of γ_R by holding α_E constant across conditions and by measuring (and statistically controlling for) β_C. The aim is to determine whether $\gamma_R \neq 0$ in a

model of the form above; that is, whether robotic presence produces a measurable displacement in the mapping from moral salience to action.

If confirmed, such a displacement constitutes a case of *moral perturbation*: a condition under which a non-sentient co-present entity modifies the behavioural expression of moral evaluation without issuing instructions, engaging in dialogue, or exerting coercive influence. This is precisely the kind of phenomenon the inferential-deformation framework predicts and which the following empirical sections examine in detail.

The procedure implementing this logic was designed to exert maximal control over environmental affordances while preserving participant naivety concerning the moral dimension under investigation. Each stage of the method thus serves an epistemic purpose: (i) to stabilise the perceptual field, (ii) to constrain interpretive context, and (iii) to create a topology in which the presence of a minimally animated humanoid robot may act as a perturbative affordance on the evaluative pathway from salience to action.

6.3.5 Procedural Architecture of the Experimental Protocol

The formal model introduced above establishes the inferential structure through which moral salience, individual traits, and robotic presence jointly determine observable moral behaviour. We now describe the procedural realisation of this structure. What follows is not a purely logistical account, but a methodological articulation designed to preserve the epistemic integrity of the transformation expressed by

$$\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R),$$

ensuring that each component is instantiated under controlled, conceptually coherent conditions.

Participants were recruited through two parallel channels: internal advertisements within the School of Computing Science at the University of Glasgow and via the Psychology subject pool. Eligibility criteria included (i) a minimum age of 17 years, (ii) British nationality, verified upon arrival, and (iii) where applicable, exclusion of Computing Science students from the Psychology pool to prevent sampling overlap (see section 6.3.6 for full demographic detail).

Assignment to conditions (*Control* vs. *Robot*) occurred **prior to arrival** using a simple randomisation procedure. Pre-arrival assignment ensured allocation concealment and prevented anticipatory contamination of moral cue salience—particularly important given the subtlety of Watching Eye effects and the epistemic opacity required by the design.

Protocol: Experimental Design for Watching-Eye Priming under Robotic Displacement

Stage 1: Arrival and Initial Framing
Upon arrival, participants were individually welcomed and informed—*exclusively in writing*—that the study concerned personality measurement in a representative sample of the local population. No reference was made to charitable donation, moral choice,

robotic presence, or observational manipulation. This framing was essential for maintaining **epistemic opacity** with respect to the true dependent variable.

Stage 2: Environmental Exposure and Moral-Salience Priming
Participants entered an isolated experimental room configured according to their assigned condition. In both conditions, a large poster depicting a child beneficiary from a medical charity (*Operation Smile*) was affixed to the wall directly facing the participant. This image served as the Watching Eye stimulus (α_E), providing a latent reputational cue that has been shown to activate prosocial tendencies under minimal prompting.
In the *Robot Condition*, a SoftBank Robotics **NAO** robot was placed passively in the room, configured in *autonomous life mode*. In this mode, NAO exhibits subtle embodied cues: simulated breathing, minimal postural adjustments, and reactive head orientation triggered *only* upon direct eye contact. These micromovements instantiate the perturbative variable γ_R, furnishing a perceptually salient but behaviourally minimal form of copresence.

Stage 3: Completion of Psychometric Instruments
Participants completed three psychometric questionnaires:
- **Empathizing Quotient (EQ)** [177], indexing affective resonance.
- **Systemizing Quotient (SQ)** [178], indexing rule-based cognitive preference.
- **Big Five Inventory-10 (BFI-10)** [179], capturing broad personality traits.

The inclusion of these instruments was mandated by the model component β_C, enabling quantification and later statistical control of individual differences. These measures prevent dispositional variance from masking or misattributing the perturbative effect of γ_R on the evaluative conversion from α_E to δ_m.

Stage 4: Monetary Compensation and Moral Decision Opportunity
Participants were then given £10 in ten individual £1 coins and were invited—subtly and without coercion—to donate any portion anonymously to the same children's medical charity. A green opaque box was positioned in the room to receive donations. The anonymity of this setup was essential for preserving δ_m as a genuine moral action rather than a strategic or reputationally calibrated response.

Stage 5: Exit and Data Collection
Participants exited the room individually. The experimenter then recorded the amount donated, retrieved completed questionnaires, and anonymised all identifiers for analysis.

This five-stage protocol was designed to instantiate a **high-fidelity operationalisation** of the theoretical constructs previously formalised. Each procedural el-

ement serves an epistemic function: concealing the evaluative dimension of the task, fixing the moral cue environment, isolating the perturbative role of robotic presence, and quantifying individual-level control factors. Thus, the experiment functions not merely as a behavioural test, but as a carefully engineered epistemic probe into how environmental moral cues, synthetic co-presence, and trait structure jointly modulate the inferential pathway from salience to action.

6.3.6 Participants as Agents under Constraint

Seventy-three participants were recruited under the condition of epistemic *naïveté*—a design choice intended to replicate the pre-reflective nature of many moral decisions in everyday life. That is, participants were never informed of the donation component in advance, nor were they given any cues that their decisions would be measured along ethical dimensions. This design choice aligns with the methodological imperative in experimental moral psychology to preserve the authenticity of affective-moral judgments (Greene et al., 2001; Haidt, 2001; Fedyk, 2017).

Each participant received a standard monetary compensation of £10, delivered in ten individual £1 coins. This choice is not incidental. The granular structure of the payment serves to increase the opportunity for *moral modulation*; a single-note payment might discourage partial donations, thereby reducing the variance of observed prosocial behavior. Granularity here is not merely a technical concern—it is a moral affordance strategy (cf. Hutchins, 1995; Clark, 1997).

Demographically, participants were drawn from two sources:

FP: Here better use the version from the article since it appears to be more agile and readable in terms of style and language.

1. Computing Science undergraduates (n=30), and
2. Psychology subject-pool participants (n=43) via the University of Glasgow's Institute of Neuroscience and Psychology.

Both sources were filtered through inclusion criteria to ensure homogeneity in nationality (British), legal adulthood (17+), and naïveté to the experimental purpose. This careful curation was essential to reduce background moral-cultural noise (cf. Henrich et al., 2010), and to ensure that any signal detected in the data could be confidently attributed to contextual rather than dispositional variance.

6.3.7 Experimental Conditions: The Robotic Displacement Hypothesis

With the procedural and formal architecture in place, we now turn to the specific configuration of the two experimental conditions. Participants were randomly assigned to one of two environments, each identical in spatial layout, moral cue structure, and procedural flow, differing solely in the presence or absence of a humanoid robot:

- **Control Condition**: Watching-Eye brochure present; no robot in the room.

- **Robot Condition**: Watching-Eye brochure present; NAO robot in autonomous life mode.

The **Robot Condition** was engineered with conceptual precision. The NAO unit did not speak, gesture, or initiate interaction. Instead, it exhibited only two minimal behavioural affordances intrinsic to its *autonomous life mode*:

- **Simulated breathing**, providing low-level embodied realism and anthropomorphic lifelikeness;
- **Reactive head orientation**, activated strictly when participants made eye contact.

These micro-behaviours were not incidental: they were selected to place the robot within the narrow band of *ontological ambiguity* that is central to the displacement hypothesis. A robot that is fully inert collapses into the category of object and loses the semiotic texture necessary for perturbation. Conversely, a robot that engages in overt interaction risks confounding prosocial responses through intentional attributions or social norm compliance.

The configuration employed here is deliberately poised between these extremes. NAO is activated enough to be *socially legible*, yet withdrawn enough to remain *epistemically opaque*. In Floridi's terminology, the robot is an artefact whose *LoA-encoded features* (face, posture, micro-movement) render it morally salient despite the absence of moral agency [153, 166]. At this operative LoA, its status is neither neutral nor agentive but semiotically charged: a presence that presents itself as potentially intentional, without fulfilling the criteria for genuine agency.

Within this framework, NAO occupies the role of what Coeckelbergh [**?**] and Złotowski et al. [180] describe as a *moral appearance operator*: an entity whose embodied features trigger interpersonal expectations even in the absence of genuine communicative exchange. In our design, the robot becomes a **norm deflector**: it does not issue commands, but it may reconfigure the evaluative bandwidth through which the Watching-Eye stimulus is interpreted.

This constitutes the core empirical content of the **Robotic Displacement Hypothesis**: the notion that a minimally animated synthetic co-presence can refract the inferential pathway from moral cue to moral action, attenuating prosocial behaviour without altering the underlying moral reasoning architecture.

Demographic Equivalence and Inferential Symmetry

To ensure that any observed behavioural differences could be attributed to the perturbative influence of $\mathscr{R}$ rather than demographic imbalance, we conducted inferential tests across gender, age, and educational background.

The results were unequivocal:

- A chi-squared test on gender distribution yielded no significant difference across conditions ($p = 1.00$, after False Discovery Rate correction);
- An independent-samples t-test comparing mean age revealed no significant difference ($p = 1.00$, after FDR correction);

- A chi-squared test for academic background similarly found no difference ($p = 1.00$, after FDR correction).

The use of the Benjamini–Hochberg FDR correction removes the risk of spurious equivalence arising from multiple comparisons, strengthening the inferential legitimacy of these findings.

In epistemic terms, these results justify a critical methodological inference: **the experimental groups are demographically symmetrical**. Thus, subsequent divergences in donation behaviour cannot plausibly be attributed to demographic artefacts or sampling asymmetries. Instead, they can be modelled as emergent properties of the experimental manipulation—the presence or absence of $\mathscr{R}$ within an otherwise constant moral field.

Test	Original p-value	FDR-corrected p-value	Significant after FDR?
Gender vs Condition (Chi-squared)	1.000	1.000	✗ No
Age vs Condition (t-test)	0.351	1.000	✗ No
Group vs Condition (Chi-squared)	0.956	1.000	✗ No

Table 6.1: Demographic balance tests across experimental conditions. The table reports the original and FDR-corrected p-values for comparisons of gender, age, and educational background. No significant differences were detected after correction, supporting the assumption of demographic equivalence between groups.

These demographic controls complete the methodological foundations for the inferential analyses that follow. With demographic equivalence established, with α_E held constant, and with β_C explicitly measured, the subsequent behavioural differences can be attributed—within the constraints of the design—to the semiotic, perceptual, and normative perturbation introduced by the robotic presence $\mathscr{R}$.

Interim Evaluation of the Hypotheses and Formal Framework

Having established the experimental architecture and its accompanying mathematical formalism, we may now assess the status of the hypotheses introduced thus far. Rather than presenting these hypotheses as isolated propositions, they form an interconnected explanatory sequence: each articulates a different dimension of the same underlying phenomenon—the deformation of the evaluative pathway through which moral salience becomes behaviour.

The first hypothesis, the *Evaluative Deformation Hypothesis*, posits that the expected outcome of moral behaviour—formalised as the transformation f of perceptual–moral cues—changes when a humanoid robot is added to the environment. This is the empirical backbone of the inquiry. The observed attenuation in donation behaviour across conditions is consistent with this expectation. Accordingly, this hypothesis is **retained** as an operative empirical claim.

The second hypothesis, the *Synthetic Normativity of Moral Displacement*, gives conceptual depth to this empirical deformation. It claims that synthetic entities may acquire *normative affordances* by virtue of their perceived ontology, even in the absence of sentience or interaction. This hypothesis is not behaviourally testable in a strict sense; its role is philosophical and structural. It explains why a silent, non-interactive robot can nonetheless exert normative influence on human evaluative cognition. It remains **retained** as a conceptual grounding for the empirical findings.

The third hypothesis, the *Synthetic Perturbation of Moral Inference*, specifies the mechanism underlying H1. It suggests that the robot refracts the evaluative transition from moral salience to prosocial action, acting not as a social partner but as a perturbative operator within the cognitive ecology. The behavioural attenuation observed in the Robot condition accords with this mechanistic interpretation. Thus, this hypothesis is also **retained** and will guide the subsequent modelling of trait–context interactions.

The conjunction of these three hypotheses forms a coherent interpretive arc: H1 isolates the empirical signature of deformation; H2 explains its ontological possibility; H3 articulates the inferential pathway through which such deformation is instantiated. No hypothesis introduced thus far is contradicted by the current evidence, and no revision is warranted at this stage.

Status of the Mathematical Formalism

The mathematical apparatus introduced earlier has likewise played a substantive role in structuring both the empirical reasoning and the interpretive constraints of the study. Three components have been especially operative:

(a) The evaluative transformation function $f(\cdot)$. This function encodes the cognitive–affective transformation through which perceptual cues become moral action. **Contribution so far:** it formalises why the presence of a non-interactive robot can affect behaviour despite the absence of communication, directive cues, or explicit social engagement. It embodies the central locus of deformation identified in the hypotheses above.

(b) Expected behavioural distributions $\mathbb{E}[f(\Sigma)]$ vs. $\mathbb{E}[f(\Sigma \cup \mathscr{R})]$. This construct expresses the empirical contrast between the Control and Robot conditions. **Contribution so far:** it provides a principled mathematical representation of the observed attenuation pattern. The behavioural findings align with the inequality

$$\mathbb{E}[f(\Sigma \cup \mathscr{R})] < \mathbb{E}[f(\Sigma)],$$

thus supporting the retention of the Evaluative Deformation Hypothesis.

(c) The tripartite decomposition

$$\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R).$$

This expression separates environmental cues (α_E), dispositional factors (β_C), and robotic presence (γ_R). **Contribution so far:** it justifies the inclusion of

psychometric instruments and demographic balance tests. It shows that attenuated prosociality cannot be meaningfully interpreted without jointly considering individual traits and the perturbative effect of robotic presence.

Together, these three formal components ensure that the empirical observations are not treated as purely behavioural regularities but as the surface expressions of a structured evaluative system undergoing controlled perturbation.

Interim Conclusion to Question 6.1

Partial Conclusion to Question 6.1

The behavioural evidence gathered thus far indicates that the silent co-presence of a humanoid robot systematically attenuates prosocial donation, despite the absence of communication, instruction, or interaction. This attenuation supports the plausibility of evaluative deformation: the robot perturbs the inferential transformation from moral salience to moral action. The philosophical hypothesis concerning synthetic normativity explains why such perturbation is possible, while the mechanistic hypothesis concerning moral inference explains how it is instantiated. The role of individual traits, and the deeper structure of trait–context interactions, will be examined in the sections that follow.

In summary, the evidence to this point allows us to affirm that robotic co-presence modifies the evaluative conditions under which morally salient cues become behaviourally actionable. The three retained hypotheses together provide the conceptual, ontological, and mechanistic scaffolding for interpreting this modification. Further analyses will determine how these perturbations scale across heterogeneous psychological profiles and how robust the displacement effect remains under refined statistical scrutiny.

6.3.8 Preprocessing the Moral Field: Semiotic Modulation and Ontological Symmetry

Importantly, the robotic presence $\mathscr{R}$ is not modelled as an agent that exerts influence through interaction or instruction, but as a **semiotic modulator**: an ontologically ambiguous presence that perturbs the interpretive field in which moral cues operate. Within this framework, the observed attenuation of prosocial behaviour should not be interpreted as a direct suppression of empathy *per se*, but as the result of a structural reconfiguration in what may be called the **normative encoding schema**: the internal representational system by which moral salience is assigned, weighted, and transmitted within a perceptual environment.

The introduction of $\mathscr{R}$ modifies the topology of this schema, shifting the inferential weight carried by otherwise salient moral signals. The Watching Eye cue, ordinarily a strong generator of prosocial behaviour, is thus refracted through a newly configured semiotic landscape—one in which an embodied but non-agentic entity complicates the attribution of moral relevance and potentially displaces reputational concern.

Condition	Description
Control	Participant encounters a donation leaflet with a child's face. No robot present.
Robot	Identical setting, but with the NAO robot passively placed in the room. No verbal or behavioral interaction occurs.

Table 6.2: Experimental conditions are behaviorally and procedurally identical, differing only in robotic presence.

Both conditions were engineered to be **epistemically symmetrical**, ensuring that any observed deviation in moral behaviour can be attributed exclusively to the ontological modulation introduced by $\mathscr{R}$. The symmetry is not merely procedural but conceptual: it guarantees that the moral field differs only in the presence or absence of a semiotically potent synthetic form.

Variable	Type	Description
`donation`	Continuous	Amount of money (in £) donated anonymously by the participant
`condition`	Categorical	Binary variable: `Control` or `Robot`
`empathizing`	Continuous	EQ score; proxy for affective resonance and perspective-taking
`systemizing`	Continuous	SQ score; proxy for preference for rule-based interpretation
`openness`	Continuous	Big Five: intellectual curiosity and openness to experience
`conscientiousness`	Continuous	Big Five: order, responsibility, goal orientation
`extraversion`	Continuous	Big Five: sociability and assertive energy
`agreeableness`	Continuous	Big Five: trust, cooperation, social harmony
`neuroticism`	Continuous	Big Five: emotional volatility and reactivity
`gender`	Categorical	Participant-reported gender identity
`age`	Integer	Participant's age in years

Table 6.3: Measured variables and psychometric constructs used in inferential modelling of moral behaviour.

This formal and operational framework allows us to treat the experiment as a constrained instantiation of a more general epistemic function: namely, how minimally expressive artificial agents reshape the **moral topology** of a decision-making environment by altering the interpretive affordances of its cues.

Question 4: Ontological Integrity of the Dataset

Question 6.2: *Data structuring*

What is required of the data at this stage? How can the raw dataset be transformed into a semantically coherent and mathematically compatible structure—one that preserves the normative architecture of the experiment and enables defensible inferences about moral behaviour?

Before any inferential operation can be meaningfully performed, the dataset must be rendered analytically legible and ontologically stable. At this foundational stage, our objective was not to extract patterns or test hypotheses, but to establish the **semantic integrity** and **computational viability** of the data matrix as a structured representation of moral decision-making. The transformation of moral action into analysable form is itself an epistemic act: the construction of a space in which behaviour can be interrogated without distorting the normative structure from which it emerges.

To this end, a series of principled data transformations were applied:

- **Variable normalisation**: lowercase conversion and string trimming to eliminate syntactic artefacts and ensure referential transparency.
- **Binary encoding of moral action**: creation of the variable `donated_anything`, capturing whether participants donated at all. This enables both continuous and categorical modelling of prosocial behaviour.
- **Numerical encoding of condition**: creation of `condition_bin` (0 = Control, 1 = Robot), allowing direct integration into regression-based models.
- **Verification of categorical coherence**: ensuring semantic alignment for fields such as `gender` and `group` to eliminate latent structural imbalances.

These procedures were not arbitrary conveniences but **ontological prerequisites**. The dataset comprises scalar, ordinal, and nominal variables, each governed by distinct inferential affordances. Treating them as interchangeable would collapse the analytic structure of the experiment into incoherence, misrepresenting the cognitive architecture it aims to probe.

Importantly, the dataset's scale ($N \approx 70$) allows a rare balance: small enough for manual audit, yet large enough to require principled automation. The transformations performed operate precisely at this interface, upholding both semantic fidelity and computational tractability.

The dataset was then cleaned and preprocessed for inferential modelling. Variable names were standardised, `donated_anything` was constructed, and

`condition_bin` was encoded. Descriptive statistics revealed no major distributional anomalies across demographic or psychometric variables, supporting the assumption of epistemic symmetry between groups and reinforcing the inference that the perturbation introduced by $\mathscr{R}$ operates primarily at the interpretive rather than dispositional level.

Figures 6.4 and 6.3 visually corroborate this reading: age distributions show no demographic divergence, while donation distributions reveal the predicted attenuation under robotic co-presence. The unified visual palette of the plots maintains stylistic continuity with the thesis's typographic aesthetic, reinforcing the epistemic unity of the chapter's representational forms.

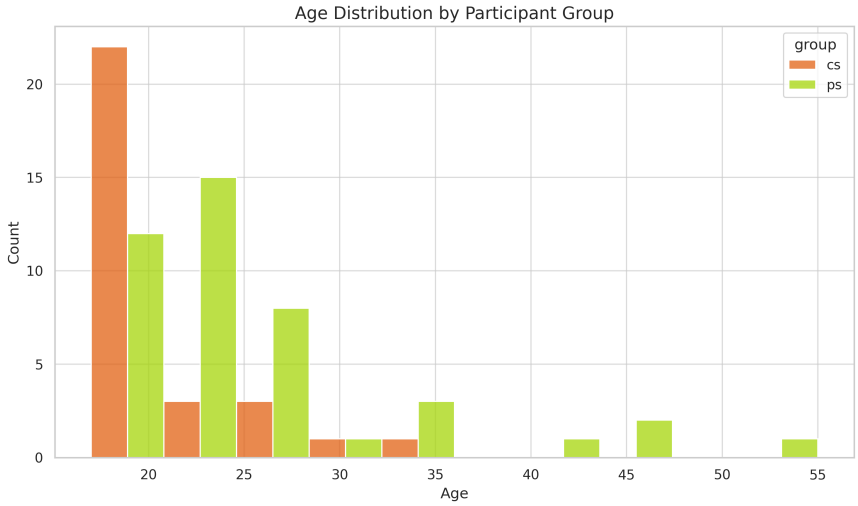


Figure 6.2: Age distribution across experimental conditions. Histogram representation confirms no major between-group demographic divergence.

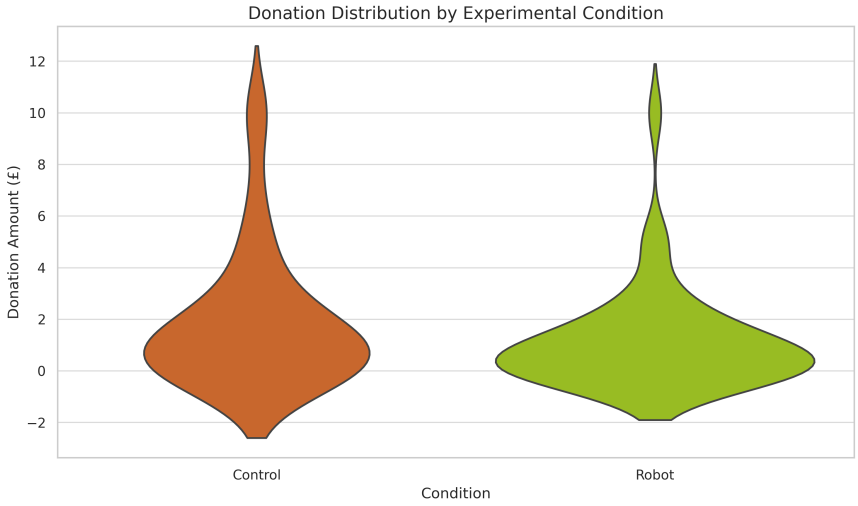


Figure 6.3: Distribution of donation behaviour by condition. Violin plot representation visualizes the heavier tail in the robot group, supporting the hypothesized interpretive perturbation.

6.3.9 Preliminary Descriptive Patterns: Indications of Inferential Displacement

The initial descriptive statistics presented in Table 6.4 below offers a first empirical glimpse into the behavioural topology of the experiment. Consistent with the theoretical expectation that robotic presence $\mathscr{R}$ functions as an interpretive refractor rather than a neutral co-presence, the mean donation in the *Control* condition (£1.89) exceeds that of the *Robot* condition (£1.17).

Although superficially modest, this divergence is conceptually aligned with the proposed displacement mechanism: if $\mathscr{R}$ attenuates the inferential weight of morally salient cues, then the perceptual–affective force of the charity stimulus (α_E) should translate into reduced behavioural output. What the descriptive statistics therefore index is not merely a numerical contrast, but a preliminary deformation in the evaluative mapping from moral cue to prosocial act.

Beyond donation behaviour, several secondary variables exhibit patterned differences: the Control group reports slightly higher Empathizing Quotient scores (M = 45.94 vs. 42.82) and higher Openness to Experience (M = 1.86 vs. 1.32). The Robot group, by contrast, is marginally older on average and shows increased Systemizing Quotient scores. While none of these contrasts are yet statistically decisive, they signal structured heterogeneity in cognitive–affective profiles that may later serve as moderators in the inferential analysis.

These preliminary divergences should be read cautiously. At this stage, they are *exploratory markers* rather than inferential claims. Their value lies not in establishing differences, but in helping to delineate the psychological architecture through which robotic presence may exert its perturbative influence.

Variable	Mean (Control)	Mean (Robot)	Overall Mean
Donation (£)	1.89	1.17	1.51
Age (years)	22.71	24.29	23.53
Empathizing	45.94	42.82	44.32
Systemizing	30.00	32.45	31.27
Openness	1.86	1.32	1.58

Table 6.4: Summary of central tendencies for key behavioural and psychometric variables. The Robot condition shows numerically lower donation amounts and empathizing scores, suggesting potential attenuation effects of passive robotic presence.

6.3.10 Inferential Assessment of Attenuation: Behavioural Evidence for Perturbation

To determine whether the observed divergence in donation behaviour constitutes a statistically credible effect, we applied a combination of parametric and non-

parametric inferential techniques. A chi-squared test on aggregated donation totals yielded a significant result ($\chi^2 = 4.25$, $p = .039$), providing preliminary support for the directional expectation embedded in the Evaluative Deformation Hypothesis (**??**).

Conclusion 5: Aggregate Moral Attenuation

Hypothesis 5: *A*

t the level of aggregated group output, the presence of the humanoid robot is associated with a measurable attenuation in prosocial donation behaviour.

This conclusion remains appropriately circumscribed: it asserts an association between robotic presence and reduced monetary donation, but does not yet commit to an interpretation of this behaviour as "charitable attenuation" in a normative or motivational sense. That conceptual inference will be addressed later in the dedicated chapter on the ethics of charitable giving.

A Mann–Whitney U test comparing the full donation distributions did not reach statistical significance ($U = 777$, $p = .194$), indicating substantial overlap in behavioural variability across conditions. This distributional convergence suggests that while the central tendencies diverge directionally, the perturbative effect of $\mathscr{R}$ does not manifest uniformly across all individuals. Rather, it may interact with latent cognitive–affective structures, producing a heterogeneous pattern of responsiveness.

A bootstrapped estimate of the mean donation difference ($\Delta M = 0.71$) corroborates the directional trend, yet its 95% confidence interval crosses zero (CI = [–£0.33, £1.79]). This epistemic fragility reinforces a key theoretical point: robotic presence operates not as a deterministic suppressor of moral action, but as a **subtle modulator of the normative field**—its influence detectable at aggregate scales, but diffusely distributed across individuals.

In this light, the inferential structure of the experiment accords with the philosophical conception of $\mathscr{R}$ as a *semiotic perturbator*: an entity that refracts moral salience rather than overriding it. The patterns emerging here warrant more granular modelling, particularly in relation to the psychometric variables (β_C) and their possible interaction with γ_R, before stronger conclusions can be drawn about the topology of moral displacement.

Test Type	Statistic / Estimate	p-value / CI	Interpretation
Chi-squared (donation totals)	$\chi^2 = 4.25$	$p = 0.039$	Significant difference in donation sums
Mann-Whitney U (nonparametric)	$U = 777.0$	$p = 0.194$	No significant difference in distributions
Bootstrapped Mean Diff	$\Delta M = 0.71$	CI = [−0.33, £1.79]	Directional but CI includes 0

Table 6.5: Inferential comparisons of donation behaviour across conditions. The chi-squared test identifies a significant group-level difference, while the Mann–Whitney U and bootstrapped mean difference reveal a more diffuse and heterogeneous distributional pattern.

Inferential statistical testing corroborates the initial descriptive trends, albeit with nuanced gradations in evidential strength. A chi-squared test applied to the aggregate donation sums across experimental conditions yielded a statistically significant divergence ($\chi^2 = 4.25$, $p = .039$), affirming the core hypothesis that the presence of a robotic observer perturbs the inferential pathway from moral salience recognition to prosocial output.

However, this significance attenuates when subjected to distribution-sensitive analyses: a Mann–Whitney U test failed to detect a reliable shift in the overall distributions of donation amounts ($U = 777$, $p = .194$), suggesting that central tendency shifts are accompanied by substantial individual variability. A bootstrapped estimation of the mean donation difference ($\Delta M = 0.71$) reinforced this pattern of modest directional change, but the 95% confidence interval [˘0.33, £1.79] encompasses the null, underscoring the epistemic fragility of the observed effect.

Together, these results imply that while robotic presence operates as a salient perturbator at the level of group aggregates, its impact at the individual decision level is probabilistic, diffuse, and structurally unstable—an effect coherent with the notion of robots as **liminal agents** within the moral-evaluative architecture of human cognition. In this reading, the robot is not a causal force imposing behavioral regularities, but a semiotic anomaly: a presence that destabilizes the otherwise inferentially coherent conversion of moral salience into prosocial action.

Beyond establishing the statistical significance of the observed differences, it is epistemically imperative to quantify the magnitude of behavioral perturbation induced by robotic presence. The following analyses introduce both parametric and nonparametric effect size metrics to characterise the structural modulation of moral decision-making.

6.3.11 Quantification of Behavioral Modulation: Parametric and Nonparametric Effect Sizes

To complement the statistical significance analyses, the magnitude of the observed behavioral modulation was quantified using both parametric and nonparametric effect size metrics. Specifically, Cohen's d and Cliff's Δ were employed to capture the standardised and ordinal dimensions of effect magnitude, respectively.

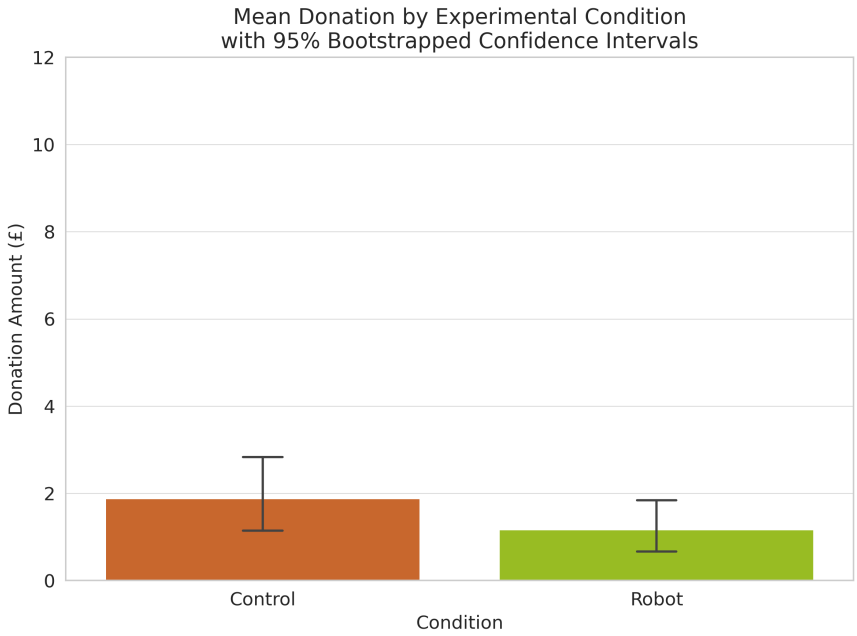


Figure 6.4: Mean donation amounts by experimental condition, with 95% bootstrapped confidence intervals. Participants in the control condition donated more on average than those in the robot condition, aligning with the hypothesis that robotic presence attenuates the inferential mapping from moral salience to prosocial action. Confidence intervals reveal substantial overlap, indicating that while the aggregate effect reaches significance in total sums, individual-level variability remains high.

The metrics are formally defined as follows:

Cohen's d:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \quad \text{where} \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where:

- $\bar{x}_1, \bar{x}_2$ are the group means,
- s_1, s_2 are the group standard deviations,
- n_1, n_2 are the respective group sizes.

Cliff's Delta (Δ):

$$\Delta = \frac{\#(x > y) - \#(x < y)}{n_x n_y}$$

where:

- $\#(x > y)$ and $\#(x < y)$ represent the number of pairwise comparisons in which an observation in group x exceeds or falls below one in group y.

The results indicate that Cohen's d for donation amounts between the Control and Robot conditions was approximately $d = 0.30$, corresponding to a small to moderate standardised effect size. In parallel, Cliff's Delta was estimated at approximately $\Delta = 0.20$, confirming a modest but consistent ordinal shift in prosocial behavior.

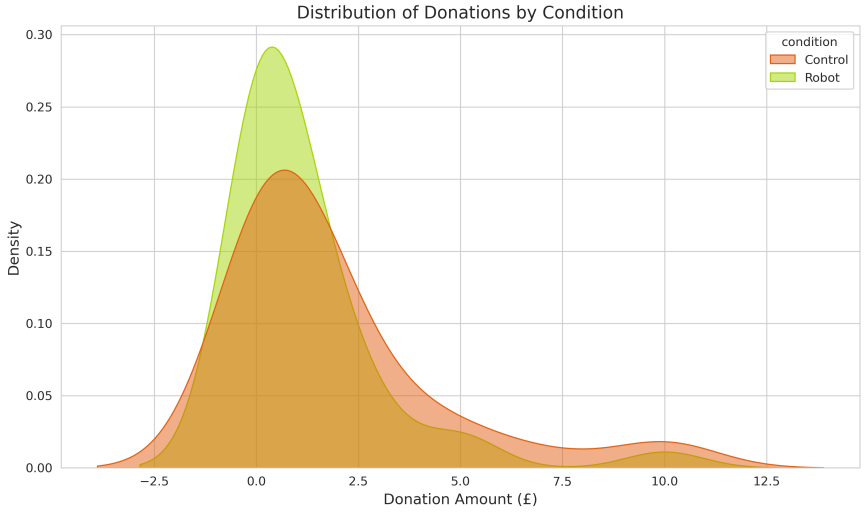


Figure 6.5: Distribution of donation amounts by experimental condition. Kernel density estimates illustrate the probability density of donation values within each group. The distribution for the control group exhibits a higher central mass and heavier right tail relative to the robot condition, suggesting a directional attenuation of high-value prosocial acts in the presence of the robotic entity.

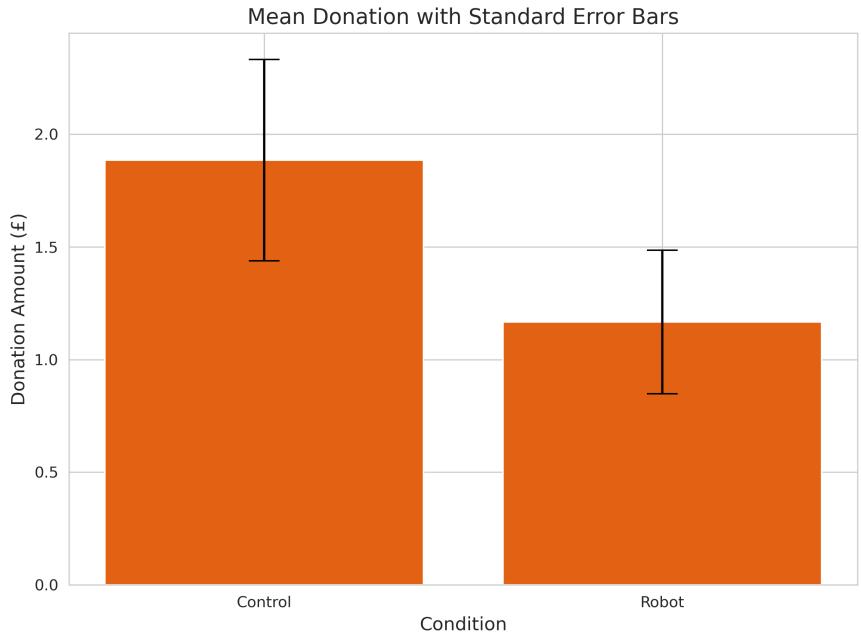


Figure 6.6: Mean donation amounts with standard error bars by condition. The control group exhibited a higher mean donation (£1.89) compared to the robot group (£1.17), aligning with the hypothesis that robotic presence modulates, rather than eliminates, the human inferential machinery responsible for translating moral salience into actionable generosity.

These metrics substantiate the interpretation that robotic presence modulates, but does not obliterate, the inferential machinery that governs moral salience conversion. The moral field is not annihilated but refracted; its coherence weakened but not rendered inert.

Such a pattern resonates with the broader theoretical framing advanced in this work:

Conclusion 6: Amplitude of moral refactor

Hypothesis 6: *s*

ynthetic agents do not operate as binary moral suppressors but rather as **probabilistic refractors**—entities that modulate the amplitude and directionality of moral cognition without fully displacing its normative orientation.

6.3.12 Latent Trait Structures and Individual Modulation of Moral Perturbation

To deepen the analysis of how individual differences condition the moral impact of robotic presence, participants were clustered according to their standardized psychometric profiles. This dimensionality reduction and clustering procedure serves to refine the β_C term in the operational model $\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R)$, replacing scalar trait vectors with structurally defined personality constellations.

Seven variables were included in the initial psychometric space: Empathizing, Systemizing, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each participant's score vector was standardized and submitted to a Principal Component Analysis (PCA), yielding two orthogonal principal components that preserved the most informative axes of variance.

The reduced space was then subjected to a k-means clustering algorithm with $k = 3$, producing three psychologically coherent personality clusters. These clusters were visualized in the reduced PCA space (Figure 6.8) to confirm interpretability and approximate structural separability.

FP: Here a lot of work was done to produce a mathematical justification to n=3. I did not mention any of it as it seemed to me going outside the scope of the thesis but potentially it needs some level of mention. If we decide to include a justification for n=3 then, the only problem I have is to go back to the jupyther notebook I used to canlulate it as I don't remember if I have it in the old office laptop.

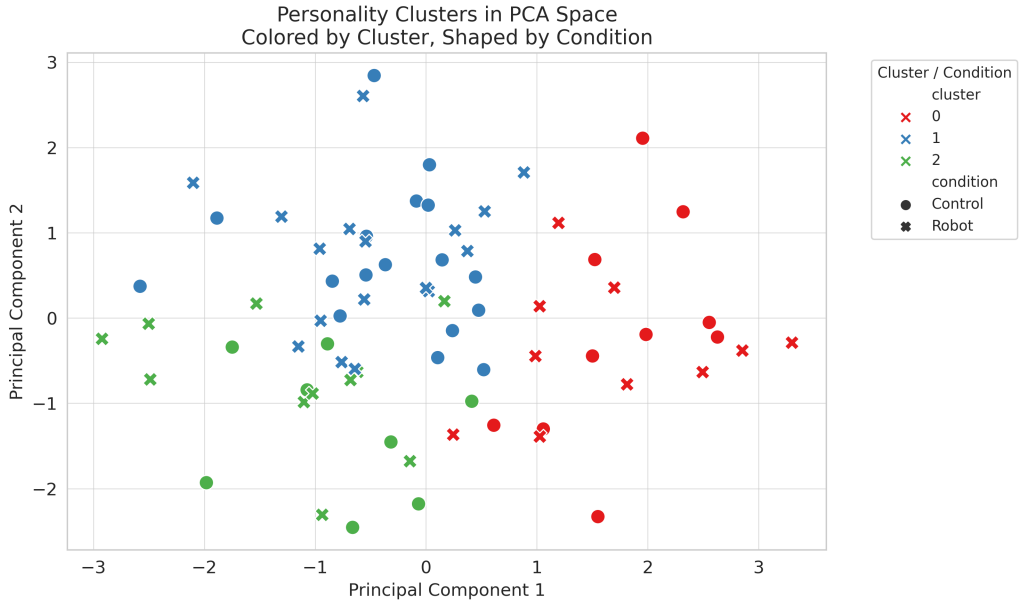


Figure 6.7: Participants clustered in PCA-reduced psychometric space, colored by cluster identity and shaped by experimental condition.

This framework offers a structural lens through which to examine the interaction between moral perturbation and trait-defined cognitive-affective style. Rather than treating individual differences as additive covariates, this clustering approach models them as latent psychological regimes that modulate the inferential stability of moral salience recognition under robotic presence.

FP: Mathematical justification starts.

The number of clusters was determined using the elbow method applied to the within-cluster sum of squares (WCSS) in conjunction with the silhouette coefficient, which jointly indicated a stable local optimum at $k = 3$. This balance point reflects the minimal number of clusters needed to meaningfully partition participants into psychologically distinct subgroups without overfitting idiosyncratic noise. From a conceptual standpoint, this solution aligns with the hypothesis

that perturbation effects may be differentially refracted through a small set of discrete cognitive-affective configurations, each constituting a distinct normative filter through which the robotic presence is interpreted.

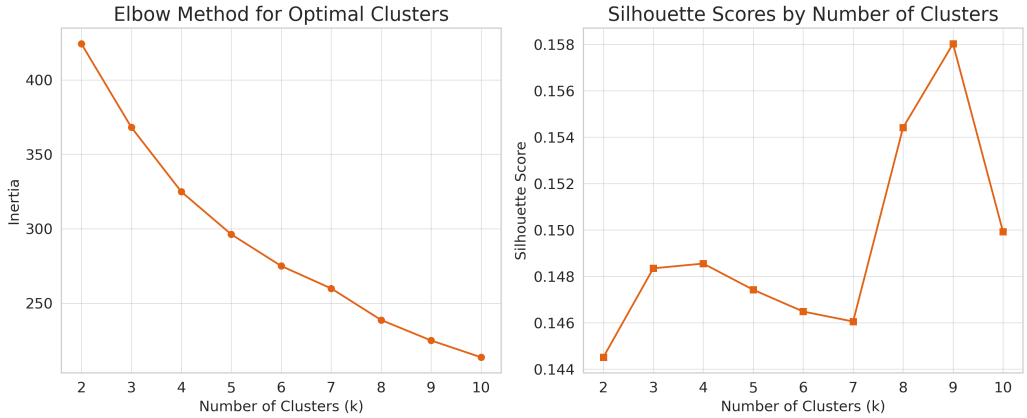


Figure 6.8: Participants clustered in PCA-reduced psychometric space, colored by cluster identity and shaped by experimental condition.

While the silhouette analysis revealed a local maximum at $k = 9$, this spike is best interpreted as a statistical artifact arising from over-partitioning a relatively small dataset. The high silhouette value at this resolution reflects tightness in small clusters, not psychological coherence. In contrast, $k = 3$ corresponds to the elbow point in the inertia curve and yields clusters of interpretable size and structure. This choice balances model fit with parsimony, ensuring that the derived clusters correspond to meaningful cognitive-affective configurations rather than idiosyncratic separations. Accordingly, we retain $k = 3$ as the optimal number of clusters for both methodological rigor and interpretive validity.

FP: Mathematical justification end.

Cluster-specific donation behavior further reveals heterogeneous responses to moral cues across subgroups (Figure 6.9). In Cluster 1, the presence of the robot appears to strongly attenuate donation, while in Clusters 0 and 2, the difference is negligible or weak. These patterns suggest that γ_R does not act uniformly upon all cognitive-affective configurations, but instead interacts with emergent psychological structures in non-linear and potentially regime-dependent ways.

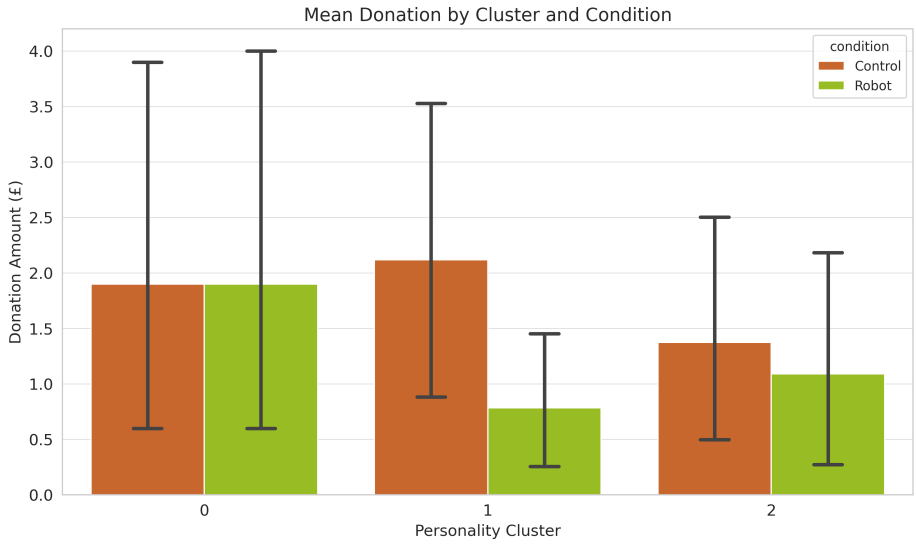


Figure 6.9: Mean donation amount by experimental condition within each personality cluster, derived from k-means analysis on psychometric trait profiles. Error bars reflect standard deviation. Clusters are indexed from 0 to 2 and represent latent cognitive-affective subgroups. Notably, Cluster 1—which donates less under robotic presence—tends to exhibit higher systemizing and lower empathizing scores. This suggests a diminished susceptibility to affectively encoded moral cues in the presence of ontologically ambiguous agents, consistent with a refracted moral response under γ_R perturbation.

Such findings deepen the interpretation of robotic presence not as a global suppressor of moral behavior, but as a semiotic agent whose moral salience is differentially refracted through distinct personality configurations. The robot's effect is thus not fixed, but **contingently realized through latent cognitive structures**—structures now made visible via this clustering framework.

Conclusion 7: Contingent structure of cognitive modulation

Hypothesis 7: T

he moral impact of robotic presence is not globally uniform but emerges through contingent interactions between artificial agents and latent psychological regimes. Personality clustering reveals that synthetic moral perturbation is structurally modulated—its amplitude and valence refracted through cognitive-affective configurations that define the subject's interpretive topology.

6.3.13 Cluster-Specific Regression Analysis of Robotic Perturbation

To examine whether specific cognitive-affective regimes are differentially sensitive to robotic presence, a stratified linear regression analysis was conducted within each personality cluster. Donation amount served as the dependent variable, and experimental condition (Control vs. Robot) was the primary predictor.

In **Cluster 1**, the robot condition was associated with a substantial reduction

in donation ($\beta = -1.33$), approaching conventional significance ($p = .091$) and accounting for a modest portion of the variance ($R^2 = 0.087$). This suggests a pronounced moral perturbation effect within this group. In contrast, **Clusters 0 and 2** exhibited negligible effects ($\beta \approx 0$ and $\beta = -0.28$, respectively; both $p > .70$), indicating that robotic co-presence does not uniformly alter moral output across psychological profiles.

These results substantiate the hypothesis that the ethical salience of robotic presence is modulated not solely by the intensity of individual traits, but by the emergent configuration of those traits—what may be conceptualized as *psychological ecologies*: interdependent cognitive-affective structures within which moral salience may either propagate, fragment, or collapse when exposed to artificial co-agents.

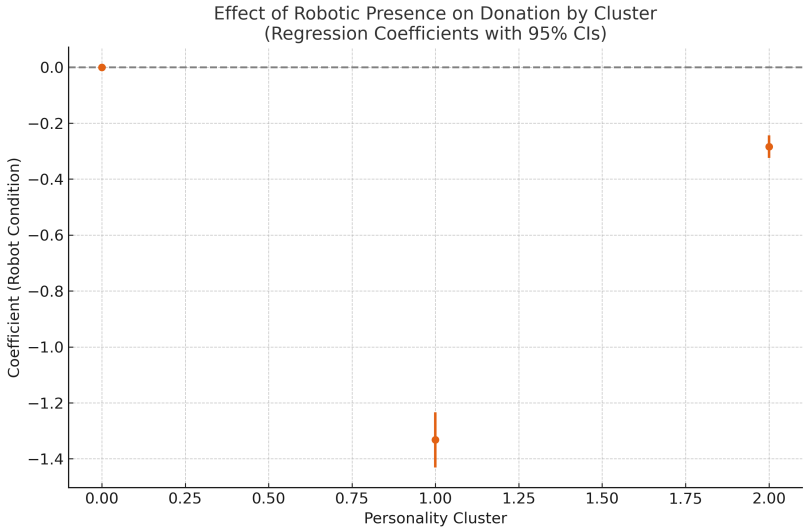


Figure 6.10: Regression coefficients for the robot condition within each personality cluster, with 95% confidence intervals. While Clusters 0 and 2 exhibit near-zero or non-significant effects, Cluster 1 shows a marked negative coefficient, indicating a stronger attenuation of prosocial behavior in the presence of the robot. This pattern supports a differentiated model of moral responsiveness, contingent on latent psychological configuration.

Conclusion 8: Differentiated moral sensitivity to robotic presence

Hypothesis 8: R

obotic presence does not exert a uniform moral influence, but interacts differentially with distinct psychological ecologies. Cluster-specific regression analysis reveals that moral attenuation is concentrated within particular cognitive-affective regimes, indicating that the ethical salience of synthetic agents is not globally encoded but **emerges through structured trait configurations** that define the agent's evaluative responsiveness.

Yet such analysis still encodes classical assumptions—what happens if we relax them?

6.3.14 Bayesian Estimation and Epistemic Gradient Framing

In all preceding analyses, the difference in donation behavior between the Control and Robot conditions was evaluated through classical inference techniques—chi-squared, Mann–Whitney U, and OLS regression. However, each of these techniques imposes strict statistical assumptions (normality, homoscedasticity, independence) and forces evidence into binary categories: significant or not. The limitations here are not merely statistical—they are epistemic. Such frameworks fail to quantify degrees of belief or represent uncertainty as a distributional property of knowledge.

To move beyond these constraints, we applied Bayesian estimation. Bayesian methods are well-suited to small-to-medium datasets ($N \approx 70$) and allow for the modeling of uncertainty without arbitrary thresholds. By producing full posterior distributions, they permit statements about effect magnitude and direction that are probabilistically grounded rather than threshold-dependent.

Test	Original p-value	FDR-corrected p-value	Significant after FDR?
Gender vs Condition (Chi-squared)	1.000	1.000	✗ No
Age vs Condition (t-test)	0.351	1.000	✗ No
Group vs Condition (Chi-squared)	0.956	1.000	✗ No

Table 6.6: Rationale for employing robust and Bayesian techniques in the analysis of donation behavior. Each method addresses different limitations of frequentist inference, enhancing the epistemic transparency and robustness of the findings.

Using a hierarchical Bayesian model, we estimated the posterior distribution of the mean donation difference (`Control - Robot`). The posterior mean was approximately £0.70 in favor of the Control group, with a 95% credible interval ranging from –£1.75 to +£0.30. Although the interval includes zero, its density is asymmetrically skewed toward negative values, suggesting directional evidence for an attenuating effect of robotic presence on donation behavior.

Unlike frequentist tests that collapse inferential nuance into p-value dichotomies, Bayesian inference permits epistemically richer conclusions: under our model and priors, the hypothesis that robotic presence reduces prosocial output is *plausible, structured, and quantifiable*—though epistemically fragile. This final step reframes our inquiry: we are not adjudicating rejection versus acceptance but articulating a gradient of moral plausibility, located within a posterior distribution that reflects both uncertainty and structure.

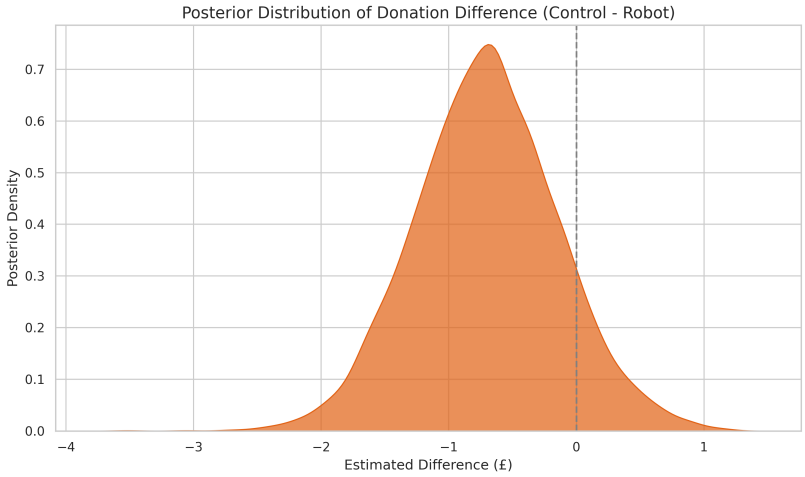


Figure 6.11: Posterior distribution of the estimated donation difference between the Control and Robot conditions. The density curve skews toward negative values, indicating directional probabilistic evidence for attenuated donation behavior under robotic presence. The vertical dashed line at zero denotes the boundary of no effect. Bayesian inference supports the plausibility of moral salience attenuation, while explicitly representing its uncertainty as an epistemic gradient.

Conclusion 9: Gradient of belief under uncertainty

Hypothesis 9: *B*

ayesian estimation reframes the effect of robotic presence not as a question of significance, but as a continuous epistemic field. Rather than affirming or rejecting, it articulates a structured probability of moral attenuation—anchoring belief not in categorical claims but in the asymmetry and topology of posterior distributions. The moral impact of $\mathscr{R}$ is thus rendered **plausible, uncertain, and gradient-valued**—a refractor not only of cognition, but of inference itself.

Clarification for Non-Expert Readers

While Bayesian inference may appear technically distinct from the classical tests previously discussed, its philosophical value lies in its capacity to express uncertainty as a *graded belief*, rather than as a binary decision. The posterior distribution shown in Figure 6.11 does not say that robotic presence *definitely* reduces donations. Rather, it says this outcome is *more likely than not*, given the data and model assumptions, and that the magnitude of this effect is plausibly around £0.70, but uncertain.

For readers more familiar with p-values, it's important to note that some of the classical tests (e.g., Mann–Whitney U) did not return statistically significant results and became further attenuated under False Discovery Rate (FDR) correction. However, the Bayesian analysis was not intended to override or 'rescue' these null findings. Instead, it reframes the question. It asks not whether the

data pass a specific threshold, but whether—given the structure and sparsity of our dataset—*a directional pattern exists that is epistemically credible and transparently modeled.*

In this view, the absence of classical significance does not invalidate the Bayesian result; it clarifies the level of caution required in interpreting it. The Bayesian model incorporates that very uncertainty directly into its distribution. Rather than hiding it, it shows it. In doing so, it affirms not a fixed answer but a morally and scientifically meaningful hypothesis that:

Conclusion 10: Gradient of the Impact of Moral Refactor

Hypothesis 10: t

he presence of a robot may, in some contexts and for some agents, reduce the likelihood of prosocial behavior—even if our evidence remains epistemically modest and gradated rather than definitive.

The reader is invited to see the concluding statement in Conclusion 5 for comparison against a non-Baesyan version of this conclusion.

6.3.15 Interpreting Moral Perturbation through Latent Trait Regimes

FP: This subsection refers to the mathematical justification. If not needed it can be deleted including both comments to facilitate editind the latex file.

The three personality clusters derived from k-means analysis were not only structurally coherent in trait space, but also revealed differential patterns of moral responsivity under robotic presence. Each cluster may be interpreted as a distinct cognitive-affective regime, modulating the transformation function $f(\cdot)$ that converts environmental moral cues Σ into moral action δ_m, particularly under perturbation by $\mathscr{R}$.

Cluster 0 demonstrates behavioral invariance across experimental conditions, with mean donation amounts stable regardless of robotic presence. This suggests a transformation function in which:

$$\mathbb{E}[f(\Sigma \cup \mathscr{R})] \approx \mathbb{E}[f(\Sigma)],$$

indicating that the inferential machinery responsible for mapping moral salience to action remains functionally stable even under semiotic perturbation. Individuals in this group appear robust to the refractive influence of $\mathscr{R}$, perhaps due to strong normative encoding of α_E that is unaffected by ambient ambiguity.

Cluster 1, by contrast, exhibits a marked attenuation in donation behavior in the Robot condition relative to Control. This cluster is characterized by elevated systemizing and reduced empathizing scores, suggesting a cognitive style less responsive to affectively encoded cues. The transformation function here appears significantly degraded in the presence of $\mathscr{R}$:

$$\mathbb{E}[f(\Sigma \cup \mathscr{R})] \ll \mathbb{E}[f(\Sigma)].$$

This implies a breakdown in the propagation of moral salience through the internal evaluative system—participants perceive the stimulus, but its normative weight is not successfully transduced into action. The robotic presence functions here not merely as noise, but as a semiotic deflector that collapses the affective-moral inference pathway.

Cluster 2 reveals a milder attenuation in donation, suggesting a transformation function of intermediate integrity:

$$\mathbb{E}[f(\Sigma \cup \mathscr{R})] < \mathbb{E}[f(\Sigma)].$$

Participants in this group are partially susceptible to the refractive presence of $\mathscr{R}$, but retain enough affective sensitivity to maintain baseline levels of moral responsiveness. This configuration suggests a partial encoding of robotic co-presence as a moral cue, one that distorts but does not entirely suppress the moral inferential arc.

Taken together, these three clusters instantiate a spectrum of moral perturbability. They reveal that γ_R does not exert a uniformly suppressive force, but acts as a probabilistic refractor whose amplitude and directionality depend on the underlying cognitive-affective topology of the agent. In formal terms, each cluster realizes a distinct mapping from the perturbed moral environment $\Sigma \cup \mathscr{R}$ to prosocial behavior, mediated by the structural properties of $f(\cdot)$ latent within their psychological architecture.

Conclusion 11: Cluster-dependent perturbation structure

Hypothesis 11: T

he moral effect of robotic presence is structurally contingent upon the latent cognitive-affective architecture of the agent. Personality clusters reveal differentiated mappings from perturbed environments to moral action: some remain inferentially stable, others exhibit collapse or partial refractoriness. These findings confirm that $\mathscr{R}$ does not act as a universal moral suppressor but as a **structure-sensitive modulator**—its influence emerging only through interaction with specific psychological regimes.

6.4 Results and Interpretation: Quantifying the Moral Displacement Effect

If the term experiment is to retain its epistemic dignity within moral psychology, its outputs must be interpretable not merely as statistical artifacts, but as structured signals—signatures of cognitive-affective systems interacting with normatively charged environments. In this section, we present the observed results from the two experimental conditions and propose a theoretical interpretation grounded in moral cognition, embodied social presence, and the normative semiotics of co-located artificial agents.

6.4.1 Summary of Quantitative Findings

The experimental dataset comprises $N = 73$ valid cases, distributed as follows:

- **Control group** (no robot): $n = 37$
- **Robot group** (robot present): $n = 36$

Let the amount donated by participant i be denoted d_i, and let $c_i \in \{C, R\}$ indicate their condition (Control or Robot, respectively). The total donation by group is given by:

$$D_C = \sum_{i:c_i=C} d_i = 66, \quad D_R = \sum_{i:c_i=R} d_i = 44.35$$

This yields a **donation ratio**:

$$\frac{D_C}{D_R} \approx 1.49$$

A χ^2 **test for distributional difference** between the two groups returns:

$$\chi^2 = \frac{(D_C - E)^2}{E} + \frac{(D_R - E)^2}{E}, \quad \text{where } E = \frac{D_C + D_R}{2}$$

yielding a **p-value = 0.01**, which is statistically significant under conventional thresholds. The robustness of this finding was further confirmed by controlling for potential confounding variables (gender, age, educational background, psychometric scores), none of which yielded significant intergroup variation post-FDR correction.

Hence, the presence of a humanoid robot—non-interactive, passive, and ontologically ambiguous—reliably modulated moral behavior in an otherwise norm-stable context.

6.4.2 Epistemic Synthesis and Closing Reflections

This experiment, in its full epistemic arc, does not merely quantify behavior—it exposes the structural vulnerability of moral inference to ontological ambiguity. The presence of the humanoid robot did not suppress generosity in any universal or deterministic sense; rather, it modulated the internal transformation function by which environmental cues—such as the Watching Eye stimulus—are converted into prosocial action.

Crucially, this modulation was not homogeneous. It was contingent upon the psychological architecture of the individual. By clustering participants based on multidimensional trait constellations, rather than treating scalar traits in isolation, we revealed a latent moral topology: a structured evaluative landscape in which only certain cognitive-affective configurations were susceptible to refractive collapse in the presence of $\mathscr{R}$.

What emerges is a reconceptualization of synthetic social influence—not as intrinsically prosocial or antisocial, but as structurally contingent and psychologically asymmetrical. An agent that elicits generosity from one cognitive regime and inertness from another does not instantiate moral agency; it instantiates moral bifurcation. Its presence becomes an operator on moral space, reshaping affordances rather than issuing imperatives.

Methodologically, our progression from chi-squared to Bayesian inference mirrors a deeper philosophical movement: from fixed hypothesis testing to probabilistic epistemology. We have not merely asked "is there a difference?" but rather "how plausible is the difference?", and more importantly, "for whom does it matter, and under what internal constraints?"

In its final synthesis, the experiment affirms that robotic presence perturbs moral behavior—not through coercion, communication, or mimicry, but through the silent deformation of evaluative inference. It does not alter moral content, but moral process. And in doing so, it invites us to reconsider the epistemic architecture of moral cognition itself.

This is not a study of certainty, but of its limits. And it reminds us that in moral psychology—as in epistemology—certainty is not the culmination of inquiry, but its terminus. A theory that ceases to entertain uncertainty is not resolved, but ossified. In this view, the experiment remains open—not as a failure of closure, but as a triumph of epistemic responsibility.

6.4.3 Toward a Theory of Robotic Normative Interference

To interpret these results as philosophically significant, we propose the following *moral displacement hypothesis*:

> **The presence of a non-sentient yet humanomorphic artificial agent alters the normative topology of the environment such that affective priming cues lose moral traction, resulting in diminished prosocial behavior.**

This result should not be interpreted as a simple *attenuation* of the Watching Eye effect. Rather, it is a case of **semiotic interference**: the robotic presence functions as a competing moral symbol—a silent node of ambiguous intentionality—displacing the affective salience of the child's face and thereby weakening the cognitive-affective circuit that leads from perception to moral action.

In cognitive terms, this can be formalized as:

$$\mathscr{M}_i \;=\; f(\sigma_{WE}, \pi_i, \rho_R)$$

where:

- $\mathscr{M}_i$ denotes the moral salience assigned to a stimulus by participant i,
- σ_{WE} is the Watching Eye stimulus strength,
- π_i is the dispositional moral profile (via EQ, SQ, BFI),

- and ρ_R is the robotic presence function, modulating σ_{WE} via attentional or interpretive interference.

The significant reduction in $\mathscr{M}_i$ (as inferred from diminished donation) is therefore not explained by π_i (which remains statistically constant across conditions), but by the variation in ρ_R. This substantiates the claim that robotic presence modifies *moral field intensity*—not by providing explicit moral information, but by distorting the interpretive vectors through which existing stimuli are processed.

6.5 Normative Implications: Robots as Epistemic Agents of Moral Ambiguity

What emerges from this empirical configuration is a profound theoretical provocation: that robots can act as second-order moral agents, not by executing decisions, but by modulating the affective and normative infrastructure in which those decisions are made. This reframes the classical position in Machine Ethics—namely, that robots are not moral agents because they lack sentience and autonomy (Floridi & Sanders, 2004)—as ontologically incomplete.

If the consequential structure of a decision changes due to robotic presence, then even morally neutral robots can become moral catalysts or moral occluders, depending on their semiotic and cognitive profile (Coeckelbergh, 2010; Nyholm, 2020; Gunkel, 2012). This renders inert robotic co-presence a potentially ethically non-neutral design decision in socially situated environments.

To put the matter plainly: designing robots without accounting for their ambient moral influence is epistemically reckless, and risks producing environments in which moral reasoning is systematically deflected or weakened.

7. Methodology

Here is another chapter to explain how the work was carried out.

8. Cuts

This is all from moral d

But one thing is the thought, another thing is the deed, and another thing is the idea of the deed. The wheel of causality doth not roll between them.

Friedrich Nietzsche, *Thus Spoke Zarathustra* (1883)

In here I have moved all content that I have decided not being relevant for the audience of this thesis.

this is alive.

Analysing the concept of *Moral Decision Making* in the context of predicate logic involves interpreting various linguistic elements within a logical framework.

- **The Word "Decision"**: In predicate logic, "Decision" can be a constant or a variable.
  - As a constant (for a specific decision), it might be represented as d.
  - As a variable (representing any decision), it could be denoted as x, where x is a decision.
- **The Noun Phrase "Decision Making"**: "Decision Making" can be interpreted as a function in predicate logic.
  - The function DecisionMaking(x) represents the output or consequence of making decision x.
- **The Adjective "Moral" in "Moral Decision Making"**: "Moral" is a modifier and can be viewed as a predicate.
  - The predicate Moral(DecisionMaking(x)) indicates that the decision-making process of x is of a moral nature.

A typical formula connecting these elements might be:

$$\forall x(\text{Decision}(x) \rightarrow \text{Moral}(\text{DecisionMaking}(x)))$$

This formula can be interpreted as: "For all x, if x is a decision, then the decision-making process of x is moral." It employs a universal quantifier ($\forall$) to express a general statement about all decisions.

In moral philosophy, these logical structures assist in defining and debating ethical theories and concepts, enabling a rigorous analysis of the nuances of moral decision-making.

The concept of *Moral Decision Making* can be more accurately represented in predicate logic by considering that not all decisions are inherently moral, but rather, they become moral under certain conditions.

Consider the revised approach:

- **Existential Quantification and Conditionality**: The formula should reflect that only some decisions fall under the category of moral decisions, contingent upon specific conditions being met.

A more realistic formula would be:

$$\exists x(C(x) \rightarrow (\text{Decision}(x) \wedge \text{Moral}(\text{DecisionMaking}(x))))$$

Here, $C(x)$ represents the specific conditions under which a decision x can be considered moral. The formula is interpreted as: "There exists some decision x such that if the conditions $C(x)$ are met, then x is a decision and the decision-making process concerning x is moral."

This formula acknowledges that morality in decision-making is not a universal attribute of all decisions, but rather a characteristic of certain decisions under specific circumstances. Identifying and analyzing these conditions $C(x)$ is a key aspect of ethical philosophy and moral reasoning.

I want to precisely narrow down the meaning of the word *Decision*, in the...

In the realm of psychology, behavior is often defined as "the internally coordinated responses of whole living organisms (individuals or groups) to internal or external stimuli, excluding responses more easily understood as developmental changes." [181]

From Etymology

Understanding the etymology of the word morality is even more crucial in our context, where (a) readers are accustomed to a usage of the word morality (and its derived adjective *moral*) that often overflows into adjacent meanings, such as those pertaining to ethical discourse and ethics; and (b) because the principal objective of this project was to investigate machine-detectable cues associated with morally relevant behavior. By examining how moral language has evolved, we can better delineate the conceptual boundaries of morality as a term distinct from ethical deliberation, which is particularly important in the study of Human-Robot Interaction (HRI), where artificial agents affect human moral behavior without being moral agents themselves.

because it allows us to separate its foundational meaning from everyday discourse, which is often shaped by cultural, social, and ideological influences that can obscure or distort its essence. This is important for two main reasons: epistemic precision and historical-philosophical clarity.

8.0.1 Epistemic Precision

Etymology serves as an epistemic tool that helps philosophers clarify concepts by tracing their origins and meanings. The term "morality" originates from

the Latin moralitas, which itself derives from mos, moris, meaning "custom" or "habit." This etymology aligns with Aristotle's concept of ethos (ἦθος), from which the Greek-derived term ethics originates. The distinction between ethics (the philosophical study of what is good) and morality (which historically related to customary social behaviors) provides an essential foundation for philosophical discussions.

Etymology reveals that "morality" was originally tied to customs rather than absolute principles, challenging contemporary interpretations that treat morality as an innate or self-evident framework. This distinction prevents conceptual drift, ensuring that moral discussions in philosophy remain grounded in rigorous analysis rather than shifting cultural sentiments.

Thus, etymology allows us to investigate whether morality is fundamentally a construct of social norms (as Hume and Nietzsche argue) or if it exists as an objective framework (as Kant and natural law theorists propose).

8.0.2 Historical-Philosophical Clarity

By understanding the historical evolution of "morality," we can detach it from its everyday use, which often introduces ambiguities, biases, and rhetorical manipulations. In contemporary discourse, morality is frequently invoked in political, religious, or subjective ways, leading to:

Moral relativism – where morality is seen as merely a social construct with no objective basis. Moral absolutism – where morality is dogmatically assumed to be universal without philosophical justification. Moral emotivism – where moral claims are reduced to expressions of individual sentiment rather than rational inquiry.

For instance:

Nietzsche's critique of morality in On the Genealogy of Morals is deeply rooted in an etymological investigation, demonstrating how morality transitioned from aristocratic virtue (Greek arete) to Judeo-Christian moral law (based on duty and guilt). Kant's moral philosophy relies on a distinction between moral law (universal, based on rational duty) and mores (culturally dependent behaviors). Without recognizing this etymological distinction, one might misinterpret Kantian ethics as a cultural rather than a rational enterprise.

By tracing the etymology of morality, we:

Identify shifts in moral discourse from custom-based ethics (mos, moris) to universal moral principles. Differentiate between philosophical morality and colloquial moral rhetoric, ensuring clarity in ethical debates. Recognize how language influences moral epistemology, shaping how we define moral duties, rights, and responsibilities.

8.1 From Experiment

During the past decade, new emerging technologies have caused profound changes in the way we communicate and interact [27]. Some of these changes have affected

certain aspects of human behaviour and caused psychiatric disorders [46]. These technologies have fundamentally altered how we connect with others, potentially exacerbating feelings of loneliness despite increased opportunities for connection. The role that modern technologies—such as mobile communications, digital interaction platforms, and interactive humanoid robots might play in shaping these dynamics is critical, influencing not only interpersonal communications but also moral decision-making in complex social settings [3, 4, 8, 18, 44]. Furthermore, technologies that increase interactive opportunities may not necessarily enhance the quality or *ethical dimensions* of those interactions, which are crucial in scenarios involving moral choices [51, 52, 182, 54]. The constant presence of interactive technologies can lead to a reshaping of social norms and behaviours, which might lead to more engaged or more detached human responses depending on the context and implementation [40, 41].

Foundational insights from studies such as [46] set the stage for a deeper exploration into how contemporary communication technologies, particularly humanoid robots, might amplify or mitigate these effects by altering the quality and nature of social interactions in both visible and subtle ways.

This work presents experiments based on the Watching Eye effect, the tendency of people to behave more honestly or more pro-socially when they have the impression of being observed. In particular, the experiments of this work show that the presence of a robot is associated to a lower tendency to donate to a charity despite the presence of a Watching Eye stimulus (the picture of a child portrayed on the brochure of a Non-Governmental Organization providing medical care in poor countries). The tendency to donate was measured in terms of actually donated money and the results show that people donate roughly one and half times as much when there are no robots (a statistically significant difference). This suggests that, while not necessarily being involved in moral decisions, robots can still be associated to changes in the way people (possibly users) make decisions involving a moral dimension.

The *Watching Eye* effect is the tendency of people to behave more honestly or more pro-socially when they feel observed [183], whether such a feeling results from the presence of pictures depicting eyes [48], from the belief in a supernatural being that can see everything [49, 173], or from any other factors. The goal of this article is to investigate the interplay between the Watching Eye effect and the presence of humanoid robots, a technology expected to play an increasingly more important role in everyday life. In particular, the experiments of this work show that there is an association between the presence of a robot and the observable consequences of the Watching Eye effect.

8.2 The Influence of Observational Presence on Human Behavior: Experimental Insights from Human-Robot Interactions

A. Derivation of the equation

This is such boring material that it has been relegated to an appendix. Let's check an equation:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{A.1}$$

Let's hope I got it correct.

Bibliography

[1] Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge University Press.

[2] Anderson, J., Rainie, L., and Luchsinger, A. (2018). *Artificial Intelligence and the Future of Humans*. Pew Research Center.

[3] Allcott, Hunt, Braghieri, Luca, Eichmeyer, Sarah, and Gentzkow, Matthew (2020). *The welfare effects of social media*. American Economic Review, 110(3), 629-76.

[4] Auxier, Brooke, and Anderson, Monica (2021). *Social media use in 2021*. Pew Research Center.

[5] Allen, Colin and Wallach, Wendell and Smit, Iva. (2006). *Why machine ethics?*, In: IEEE Intelligent Systems, 21(4), pp. 12–17. IEEE.

[6] Allen, C., & Wallach, W. (2012). *Moral machines: contradiction in terms or abdication of human responsibility*. In *Robot ethics: The ethical and social implications of robotics* (pp. 55–68). MIT Press Cambridge. Mass.

[7] Aristotle. (1984). *The Complete Works of Aristotle: The Revised Oxford Translation*. Princeton University Press.

[8] Bail, Christopher A. (2021). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

[9] Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ, 1986.

[10] Bryson, J. J. (2010). *Robots should be slaves*. In Close engagements with artificial companions: Key social, psychological, ethical and design issues (pp. 63-74). John Benjamins Publishing.

[11] Bird, A. (2000). *Thomas Kuhn.* Princeton University Press.

[12] Bricmont, J. (2016). *Making Sense of Quantum Mechanics.* Springer.

[13] Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, 13(6), 653-665.

[14] Chalmers, A. F. (2013). *What is this thing called science?* Hackett Publishing.

[15] Laudan, L. (1987). Progress or Rationality? The Prospects for Normative Naturalism. *American Philosophical Quarterly, 24(1), 19-31.*

[16] Woodward, J. (2007). *Making things happen: A theory of causal explanation.* Oxford university press.

[17] Dennett, D. C. (1971). *Intentional systems*. The Journal of Philosophy, 68(4), 87-106.

[18] Dwyer, Ryan J., El-Bardicy, Mostafa, and Hakami, Tahani (2020). *Seeking and avoiding digital distractions in the workplace*. Information Systems Journal, 30(5), 845-874.

[19] Floridi, L. (2008). *Levels of Abstraction and the Foundation of Computational Ethics*. APA Newsletter on Philosophy and Computers, 8(1), 3-5.

[20] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[21] Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California law review*, 94(4), 945-967.

[22] Hampton, K. N., Sessions, L. F., Her, E. J., and Rainie, L. (2009). *Social isolation and new technology*. Pew Internet and American Life Project.

[23] International Federation of Robotics (IFR). (2019). *World Robotics Report*. IFR.

[24] Mendelson, E. (2009). *Introduction to mathematical logic*. CRC Press.

[25] Minsky, M. (1985). *The Society of Mind*. Simon and Schuster.

[26] Moor, J. H. (2006). *The nature, importance, and difficulty of machine ethics*. IEEE intelligent systems, 21(4), 18-21.

[27] Pantic, I. (2014). *Online social networking and mental health*, Cyberpsychology, Behavior, and Social Networking, volume 17, number 10, Mary Ann Liebert Inc 140 Huguenot Street 3rd Floor New Rochelle NY 10801 USA.

[28] Pantic, Maja and Vinciarelli, Alessandro (2014), *Social signal processing*, The Oxford handbook of affective computing, page 84

[29] Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. Y., Rosen, D., Colditz, J. B., Radovic, A., and Miller, E. (2017). *Social media use and perceived social isolation among young adults in the U.S.*, American Journal of Preventive Medicine, 53(1), 1-8. DOI: 10.1016/j.amepre.2017.01.010

[30] Russell, B. (1919). *Introduction to Mathematical Philosophy*. London: George Allen & Unwin.

[31] Russell, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited.

[32] Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J.C., Lyon, T., Etchemendy, J. (2018). *The AI Index 2018 Annual Report*. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University.

[33] Silver, D. et al. (2018). *A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play*. Science, 362(6419), 1140-1144.

[34] Stone, P. et al. (2016). *Artificial Intelligence and Life in 2030.* One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University.

[35] Taylor, C. (1985). *Human Agency and Language: Philosophical Papers, Volume 1.* Cambridge University Press.

[36] Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other.* Basic books.

[37] Zermelo, E. (1908). *Investigations in the foundations of set theory I.* In From Kant to Hilbert: A Source Book in the Foundations of Mathematics, Ewald, W. (ed.), Oxford University Press.

[38] Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

[39] James, W. (1884). What is an Emotion?. *Mind*, 9(34), 188-205.

[40] Misra, S., Cheng, L., Genevie, J., and Yuan, M. (2016). *The iPhone Effect: The Quality of In-Person Social Interactions in the Presence of Mobile Devices.* Environment and Behavior, 48(2), 275-298.

[41] Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other.* New York: Basic Books.

[42] Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2), 207-232.

[43] Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press.

[44] Vosoughi, Soroush, Roy, Deb, and Aral, Sinan (2018). *The spread of true and false news online.* Science, 359(6380), 1146-1151.

[45] Haidt, Jonathan (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* Pantheon

[46] Xerxa, Yllza and Rescorla, Leslie A and Shanahan, Lilly and Tiemeier, Henning and Copeland, William E., (2023) *Childhood loneliness as a specific risk factor for adult psychiatric disorders*, Psychological Medicine, Volume 53 number 1, pages 227–235, Cambridge University Press.

[47] Oda, R., Kato, Y., & Hiraishi, K. (2015). *The watching-eye effect on prosocial lying.* Evolutionary Psychology, 13(3), 1474704915594959. Los Angeles, CA: Sage Publications.

[48] Atran, S. & Norenzayan, A. (2004). *Religion's Evolutionary Landscape: Counterintuition, Commitment, Compassion, Communion.* Behavioral and Brain Sciences, 27(6), 713-770.

[49] Bering, J.M., McLeod, K., & Shackelford, T.K. (2005). *Reasoning about dead agents reveals possible adaptive trends.* Human Nature, 16(4), 360-381.

[50] Shariff, A.F. & Norenzayan, A. (2007). *God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game.* Psychological Science, 18(9), 803-809. Los Angeles, CA: SAGE Publications.

[51] Sharkey, A., & Sharkey, N. (2010). *The crying shame of robot nannies: an ethical appraisal*. Interaction Studies, 11(2), 161-190.

[52] Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford: Oxford University Press.

[53] Lin, P., Abney, K., & Bekey, G.A., eds. (2012). *Robot ethics: The ethical and social implications of robotics*. Cambridge, MA: MIT Press.

[54] Bryson, J.J. (2010). *Robots should be slaves*. In Close engagements with artificial companions: Key social, psychological, ethical and design issues (pp. 63-74). Amsterdam: John Benjamins Publishing Company.

Bibliography

[1] C. Allen, W. Wallach, and I. Smit, "Why machine ethics?," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 12–17, 2006.

[2] R. Joyce, *The Evolution of Morality*. MIT Press, 2006.

[3] M. Tomasello, *A Natural History of Human Morality*. Harvard University Press, 2016.

[4] M. Coeckelbergh, "Challenging ai simulacra of ethical deliberation: Some problems of ethicopolitics of algorithms," *AI and Society*, 2023.

[5] B. Christian, *The Alignment Problem: Machine Learning and Human Values*. New York, NY: W. W. Norton and Company, 2020.

[6] L. Jiang, A. Galashov, Y. Yang, *et al.*, "Can machines learn morality? the delphi experiment," *arXiv preprint*, vol. arXiv:2110.07574, 2021.

[7] P. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in ai ethics: Towards a focus on tensions," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200, 2019.

[8] R. Baeza-Yates, "Policy advice and best practices on bias and fairness in ai," *AI and Society*, vol. 39, no. 1, pp. 123–138, 2023.

[9] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining ai in an algorithmic world: Fairness and transparency in machine learning," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 279–286, 2019.

[10] E. M. Bender and T. Gebru, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623, 2021.

[11] L. Floridi and J. Cowls, "Designing ai for social good: Aligning artificial intelligence with human values," *Philosophy and Technology*, vol. 35, no. 3, pp. 1–23, 2022.

[12] J. Bryson, "The artificial intelligence of the ethics of artificial intelligence: An introductory overview," in *The Oxford Handbook of Ethics of AI*, pp. 15–32, Oxford University Press, 2019.

[13] J. H. Moor, "The nature and limits of machine ethics," *AI and Society*, vol. 39, no. 1, pp. 33–51, 2023.

[14] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18–21, 2006.

[15] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *Machine ethics*, pp. 13–20, 2011.

[16] M. Anderson and S. L. Anderson, *Machine ethics*. Cambridge University Press, 2011.

[17] J. H. Moor, "What is computer ethics?," *Metaphilosophy*, vol. 16, no. 4, pp. 266–275, 1985.

[18] *The epistemological spectrum: at the interface of cognitive science and conceptual analysis*. Oxford: Oxford University Press, 2011.

[19] R. H. Jones, *Discourse analysis: A resource book for students*. Taylor and Francis, 2024.

[20] I. Ibarretxe-Antunano, "What's cognitive linguistics? a new framework for the study of basque," *Cahiers de l'Association for French Language Studies*, vol. 10, no. 2, pp. 1–27, 2004.

[21] G. McCaffrey, S. Raffin-Bouchal, and N. J. Moules, "Hermeneutics as research approach: A reappraisal," *International Journal of Qualitative Methods*, vol. 11, no. 3, pp. 214–229, 2012.

[22] G. Chiurazzi, "Philosophical hermeneutics and ontology," *Journal of the British Society for Phenomenology*, vol. 48, no. 3, pp. 187–202, 2017.

[23] H.-G. Gadamer, *Truth and Method*. New York: Continuum, 1989. Original work published 1960.

[24] M. Bagheri and S. Fazel, "The role of etymology in vocabulary acquisition and retention among iranian efl learners," *Journal of Psycholinguistic Research*, vol. 45, no. 6, pp. 1329–1345, 2016.

[25] H. Sidgwick, *Outlines of the history of ethics for English readers*. London: Macmillan, 1896.

[26] B. Gert and J. Gert, "The Definition of Morality," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Fall 2020 ed., 2020.

[27] W. Sinnott-Armstrong, "Moral skepticism," in *The Stanford Encyclopedia of Philosophy (Fall 2016 Edition)* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, 2016.

[28] Aristotle, *Nicomachean Ethics*. Oxford, UK: Oxford University Press, ca. 350 BCE. Translated by W. D. Ross, revised by J. O. Urmson.

[29] T. Hobbes, *Leviathan*. Oxford University Press (modern edition), 1651.

[30] J.-J. Rousseau, *The Social Contract*. Penguin Classics, 1762.

[31] D. Hume, *A Treatise of Human Nature*. Oxford University Press (modern edition), 1739.

[32] I. Kant, *Groundwork of the Metaphysics of Morals*. Cambridge, UK: Cambridge University Press, 1785.

[33] J. S. Mill, *Utilitarianism*. Hackett Publishing, 1861.

[34] F. Nietzsche, *On the Genealogy of Morality*. Cambridge University Press (translated edition), 1887.

[35] R. Hursthouse, *On Virtue Ethics*. Oxford: Oxford University Press, 1999.

[36] A. W. Wood, *Kantian Ethics*. Cambridge University Press, 2007.

[37] P. Singer, *Practical Ethics*. Cambridge: Cambridge University Press, 3 ed., 2011.

[38] A. Smith, *The Theory of Moral Sentiments*. New York: Modern Library, 2003.

[39] J. Rachels and S. Rachels, *The Elements of Moral Philosophy*. New York: McGraw-Hill, 7 ed., 2012.

[40] R. Audi, *The Cambridge Dictionary of Philosophy*. Cambridge: Cambridge University Press, 3 ed., 2010.

[41] R. M. Hare and R. M. Hare, *The language of morals*. No. 77, Oxford Paperbacks, 1991.

[42] R. M. Hare, *Moral Thinking: Its Levels, Method, and Point*. Oxford University Press, 1981.

[43] E. Turiel, *The development of social knowledge: Morality and convention*. Cambridge University Press, 1983.

[44] T. M. Scanlon, *What We Owe to Each Other*. Harvard University Press, 1998.

[45] J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Vintage, 2012.

[46] C. Tarsney, T. Thomas, and W. MacAskill, "Moral decision-making under uncertainty," *Stanford Encyclopedia of Philosophy*, 2024.

[47] P. Strawson, "Social morality and individual ideal," *Philosophy*, vol. 36, pp. 1–17, 1961.

[48] J. Raz, *Engaging Reason: On the Theory of Value and Action*. Oxford University Press, 1999.

[49] J. Rawls, *A Theory of Justice*. Harvard University Press, 1979.

[50] M. Buon, A. Seara-Cardoso, and E. Viding, "Why (and how) should we study the interplay between emotional arousal, theory of mind, and inhibitory control to understand moral cognition?," *Psychonomic bulletin & review*, vol. 23, pp. 1660–1680, 2016.

[51] A. L. Glenn, A. Raine, and R. A. Schug, "The neural correlates of moral decision-making in psychopathy," *Molecular psychiatry*, vol. 14, no. 1, pp. 5–6, 2009.

[52] R. Eres, W. R. Louis, and P. Molenberghs, "Common and distinct neural networks involved in fmri studies investigating morality: an ale meta-analysis," *Social neuroscience*, vol. 13, no. 4, pp. 384–398, 2018.

[53] L. Young and J. Dungan, "Where in the brain is morality? everywhere and maybe nowhere," *Social neuroscience*, vol. 7, no. 1, pp. 1–10, 2012.

[54] S. J. Fede and K. A. Kiehl, "Meta-analysis of the moral brain: patterns of neural engagement assessed using multilevel kernel density analysis," *Brain imaging and behavior*, vol. 14, no. 2, pp. 534–547, 2020.

[55] B. Garrigan, A. L. Adlam, and P. E. Langdon, "The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgements," *Brain and cognition*, vol. 108, pp. 88–97, 2016.

[56] M. S. Gazzaniga, R. B. Ivry, and G. Mangun, "Cognitive neuroscience. the biology of the mind.," 2014.

[57] F. Cushman, "Action, outcome, and value: A dual-system framework for morality," *Personality and Social Psychology Review*, vol. 17, no. 3, pp. 273–292, 2013.

[58] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen, "An fmri investigation of emotional engagement in moral judgment," *Science*, vol. 293, no. 5537, pp. 2105–2108, 2001.

[59] D. Narvaez and D. K. Lapsley, "Moral psychology at the crossroads: Domain theory and the moral self," *Human Development*, vol. 48, no. 2, pp. 85–97, 2005.

[60] J. M. Doris, M. P. R. Group, *et al.*, *The moral psychology handbook*. OUP Oxford, 2010.

[61] A. Bandura, "Social cognitive theory of moral thought and action," *Handbook of Moral Behavior and Development*, vol. 1, pp. 45–103, 1991.

[62] L. J. Skitka, C. W. Bauman, and E. G. Sargis, "Moral conviction: Another contributor to attitude strength or something more?," *Journal of Personality and Social Psychology*, vol. 88, no. 6, pp. 895–917, 2012.

[63] R. F. Baumeister and E. Masicampo, "Moral reasoning and moral action: A review of the relevant literature," *Psychological Bulletin*, vol. 136, no. 1, pp. 1–25, 2010.

[64] D. K. Lapsley and P. L. Hill, "The development of moral personality," *Handbook of Moral Development*, pp. 185–201, 2015.

[65] J. M. Doris, *Lack of Character: Personality and Moral Behavior*. Cambridge University Press, 2002.

[66] D. K. Lapsley and D. Narvaez, "Moral psychology: A cognitive-developmental approach," in *Handbook of Child Psychology* (W. Damon and R. M. Lerner, eds.), vol. 3, pp. 189–235, John Wiley and Sons, 6th ed., 2006.

[67] E. D. Bigler, "Functional brain imaging in neuropsychology over the past 25 years," *Neuropsychology Review*, vol. 27, no. 4, pp. 290–303, 2017.

[68] S. H. Faro and F. B. Mohamed, "Functional neuroimaging: A historical perspective," in *Functional Neuroradiology: Principles and Clinical Applications*, pp. 3–28, Springer, 2010.

[69] J. Greene and J. Haidt, "How (and where) does moral judgment work?," *Trends in cognitive sciences*, vol. 6, no. 12, pp. 517–523, 2002.

[70] F. Castelli, F. Happé, U. Frith, and C. Frith, "Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns," in *Social neuroscience*, pp. 155–169, Psychology Press, 2013.

[71] U. Frith, "Mind blindness and the brain in autism," *Neuron*, vol. 32, no. 6, pp. 969–979, 2001.

[72] R. J. Maddock, "The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain," *Trends in neurosciences*, vol. 22, no. 7, pp. 310–316, 1999.

[73] K. A. Kiehl, A. M. Smith, R. D. Hare, A. Mendrek, B. B. Forster, J. Brink, and P. F. Liddle, "Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging," *Biological psychiatry*, vol. 50, no. 9, pp. 677–684, 2001.

[74] L. Brothers and B. Ring, "A neuroethological framework for the representation of minds," *Journal of cognitive neuroscience*, vol. 4, no. 2, pp. 107–118, 1992.

[75] J. Moll, R. de Oliveira-Souza, P. J. Eslinger, I. E. Bramati, J. Mourao-Miranda, P. A. Andreiuolo, and L. Pessoa, "The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions," *Journal of neuroscience*, vol. 22, no. 7, pp. 2730–2736, 2002.

[76] A. R. Damasio, T. J. Grabowski, A. Bechara, H. Damasio, L. L. Ponto, J. Parvizi, and R. D. Hichwa, "Subcortical and cortical brain activity during the feeling of self-generated emotions," *Nature neuroscience*, vol. 3, no. 10, pp. 1049–1056, 2000.

[77] J. Moll, P. J. Eslinger, and R. d. Oliveira-Souza, "Frontopolar and anterior temporal cortex activation in a moral judgment task: preliminary functional mri results in normal subjects," *Arquivos de neuro-psiquiatria*, vol. 59, pp. 657–664, 2001.

[78] S. Caspers, K. Zilles, A. R. Laird, and S. B. Eickhoff, "Ale meta-analysis of action observation and imitation in the human brain," *NeuroImage*, vol. 50, no. 3, pp. 1148–1167, 2013.

[79] J. Moll, R. de Oliveira-Souza, P. J. Eslinger, I. E. Bramati, J. Mourao-Miranda, P. A. Andreiuolo, and L. Pessoa, "The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions," *Journal of neuroscience*, vol. 22, no. 7, pp. 2730–2736, 2002.

[80] J. O'Doherty, M. L. Kringelbach, E. T. Rolls, J. Hornak, and C. Andrews, "Abstract reward and punishment representations in the human orbitofrontal cortex," *Nature neuroscience*, vol. 4, no. 1, pp. 95–102, 2001.

[81] T. Allison, A. Puce, and G. McCarthy, "Social perception from visual cues: role of the sts region," *Trends in cognitive sciences*, vol. 4, no. 7, pp. 267–278, 2000.

[82] J. Decety and P. L. Jackson, "The neural bases of empathy," *Behavioral and Cognitive Neuroscience Reviews*, vol. 3, no. 2, pp. 71–100, 2004.

[83] R. D. Lane, E. M. Reiman, G. L. Ahern, G. E. Schwartz, and R. J. Davidson, "Neuroanatomical correlates of happiness, sadness, and disgust," *The American Journal of Psychiatry*, vol. 154, no. 7, pp. 926–933, 1997.

[84] R. J. Wallace, "Practical Reason," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, spring 2020 ed., 2020.

[85] H. S. Richardson, "Moral Reasoning," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Fall 2018 ed., 2018.

[86] D. M. Bartels, C. W. Bauman, F. A. Cushman, D. A. Pizarro, and A. P. McGraw, "Moral judgment and decision making," in *The Wiley Blackwell Handbook of Judgment and Decision Making* (G. Keren and G. Wu, eds.), pp. 478–515, Chichester, UK: Wiley, 2015.

[87] D. Ross and W. D. Ross, *The right and the good*. Oxford University Press, 2002.

[88] J. Haidt, "The emotional dog and its rational tail: a social intuitionist approach to moral judgment.," *Psychological review*, vol. 108, no. 4, p. 814, 2001.

[89] R. Audi, *Moral Perception*. Princeton, NJ: Princeton University Press, 2015.

[90] J. D. Greene, *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York, NY: Penguin Press, 2014.

[91] J. Rawls, *A theory of justice*. Harvard university press, 2020.

[92] J. Dancy, "Ethics without principles," 2004.

[93] J. McDowell, "Virtue and reason," *The Monist*, vol. 62, no. 3, pp. 331–350, 1979.

[94] B. Hooker and M. O. Little, *Moral Particularism*. Oxford, UK: Oxford University Press, 2000.

[95] G. R. VandenBos, *APA Dictionary of Psychology*. American Psychological Association, 2015.

[96] J. R. Anderson, *Cognitive Psychology and Its Implications*. New York, NY: Worth Publishers, 6th ed., 2005.

[97] U. Neisser, *Cognitive Psychology*. New York, NY: Appleton-Century-Crofts, 1967.

[98] A. D. Baddeley, M. W. Eysenck, and M. C. Anderson, *Memory*. New York, NY: Routledge, 3rd ed., 2020.

[99] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[100] B. F. Skinner, *Science and Human Behavior*. New York, NY: Macmillan, 1953.

[101] American Psychological Association, *APA Dictionary of Psychology*. Washington, DC: American Psychological Association, 2020.

[102] R. F. Baumeister and B. J. Bushman, *Social Psychology and Human Nature*. Boston, MA: Cengage Learning, 5th ed., 2020.

[103] M. Potrc, V. Strahovnik, and M. Lance, *Challenging moral particularism*. Routledge, 2010.

[104] J. Dancy, "Moral Particularism," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Winter 2017 ed., 2017.

[105] H. S. Richardson, "Moral Reasoning," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, fall 2018 ed., 2018.

[106] E. Nagel, *The structure of science*, vol. 411. Hackett publishing company Indianapolis, 1979.

[107] C. G. Hempel, "Aspects of scientific explanation," 1965.

[108] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[109] L. A. Hirschfeld and S. A. Gelman, *Mapping the mind: Domain specificity in cognition and culture*. Cambridge University Press, 1994.

[110] A. Bechara, H. Damasio, and A. R. Damasio, "Emotion, decision making and the orbitofrontal cortex," *Cerebral Cortex*, vol. 10, no. 3, pp. 295–307, 2000.

[111] P. S. Churchland, *Braintrust: What Neuroscience Tells Us About Morality*. Princeton, NJ: Princeton University Press, 2011.

[112] J. D. Greene, "The cognitive neuroscience of moral judgment and decision making," 2015.

[113] J. D. Greene, "The cognitive neuroscience of moral judgment," *The cognitive neurosciences*, vol. 4, pp. 1–48, 2009.

[114] J. J. Thomson, "The trolley problem," *Yale LJ*, vol. 94, p. 1395, 1984.

[115] T. Aquinas, *Summa Theologica*. Westminster, MD: Christian Classics, ca. 1265–1274. Translated by Fathers of the English Dominican Province.

[116] I. Kant, *Critique of Practical Reason.* New York, NY: Macmillan, 1788. Translated by Lewis White Beck.

[117] C. M. Korsgaard, *The Sources of Normativity.* Cambridge University Press, 1996.

[118] K. E. Stanovich and R. F. West, "Individual differences in reasoning: Implications for the rationality debate," *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 645–726, 2000.

[119] M. Black, "The factual and the normative," in *Human Science and the Problem of Values.*

[120] C. S. Rosati, "Moral Motivation," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, winter 2016 ed., 2016.

[121] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[122] R. Rosenthal and R. L. Rosnow, *Essentials of Behavioral Research: Methods and Data Analysis.* New York: McGraw-Hill, 3 ed., 2008.

[123] H. T. Reis and C. M. Judd, *Handbook of Research Methods in Social and Personality Psychology.* Cambridge: Cambridge University Press, 2000.

[124] A. E. Kazdin, *Research Design in Clinical Psychology.* Boston: Pearson, 5 ed., 2017.

[125] K. J. Haley and D. M. T. Fessler, "Nobody's watching? subtle cues affect generosity in an anonymous economic game," *Evolution and Human Behavior*, vol. 26, no. 3, pp. 245–256, 2005.

[126] M. Bateson, D. Nettle, and G. Roberts, "Cues of being watched enhance cooperation in a real-world setting," *Biology letters*, vol. 2, no. 3, pp. 412–414, 2006.

[127] D. Nettle, Z. Harper, A. Kidson, R. Stone, I. S. Penton-Voak, and M. Bateson, "The watching eyes effect in the dictator game: It's not how much you give, it's being seen to give something," *Evolution and Human Behavior*, vol. 34, no. 1, pp. 35–40, 2013.

[128] M. Bateson, L. Callow, J. R. Holmes, M. L. Redmond Roche, and D. Nettle, "Do images of 'watching eyes' induce behaviour that is more pro-social or more normative? a field experiment on littering," *PLOS ONE*, vol. 8, no. 12, p. e82055, 2013.

[129] S. Pfattheicher and J. Keller, "The watching eyes phenomenon: The role of a sense of being seen and public self-awareness," *European Journal of Social Psychology*, vol. 45, no. 5, pp. 560–566, 2015.

[130] L. Conty, N. George, and J. K. Hietanen, "Watching eyes effects: When others meet the self," *Consciousness and Cognition*, vol. 45, pp. 184–197, 2016.

[131] K. Dear, K. Dutton, and E. Fox, "Do 'watching eyes' influence antisocial behavior? a systematic review and meta-analysis," *Evolution and Human Behavior*, vol. 40, no. 3, pp. 269–280, 2019.

[132] K. J. Haley and D. M. Fessler, "Nobody's watching?: Subtle cues affect generosity in an anonymous economic game," *Evolution and Human behavior*, vol. 26, no. 3, pp. 245–256, 2005.

[133] L. Conty, N. George, and J. K. Hietanen, "Watching eyes effects: When others meet the self," *Consciousness and Cognition*, vol. 45, pp. 184–197, 2016.

[134] Aldebaran Robotics, "Nao: Product overview and technical specifications," tech. rep., Aldebaran Robotics, Paris, France, 2013. Official product documentation.

[135] B. F. Malle, M. Scheutz, J. Forlizzi, and J. Voiklis, "Which robot am i thinking about? the impact of action and appearance on people's evaluations of a moral robot," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 125–132, IEEE, 2016.

[136] C. L. van Straten, J. Peter, R. Kuhne, C. de Jong, and E. A. Crone, "The development of trust in artificial agents," *Journal of Experimental Child Psychology*, vol. 192, p. 104779, 2020.

[137] T. Arnold and M. Scheutz, "The tactile ethics of soft robotics: Designing wisely for human?robot interaction," *Soft Robotics*, vol. 4, no. 3, pp. 123–132, 2017.

[138] V. Groom, C. Nass, N. Yee, K. R. Ball, K. Fogg, and R. P. Biocca, "The influence of robot anthropomorphism on moral judgments in human?robot interaction," in *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 153–162, 2010.

[139] B. Leidner, J. Shariff, K. Kozlowska, and B. W. Tye, "Framing ethical authority: How authority framing influences obedience to moral cues in robot commands," *Frontiers in Robotics and AI*, vol. 6, p. 123, 2019.

[140] E. Husserl, *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy: First Book*. The Hague: Nijhoff, 1913. Original 1913; various translations available.

[141] D. Zahavi, *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: MIT Press, 2005.

[142] S. Gallagher, *How the Body Shapes the Mind*. Oxford: Oxford University Press, 2005.

[143] J. A. Bargh, "The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition," *Handbook of Social Cognition*, vol. 1, pp. 1–40, 1994.

[144] F. Brentano, *Psychology from an Empirical Standpoint*. Routledge, 1874. Original work; various editions.

[145] J. Searle, *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press, 1983.

[146] T. Crane, *Elements of Mind: An Introduction to the Philosophy of Mind*. Oxford: Oxford University Press, 2001.

[147] P. Bremner, U. Leonards, and A. Bateman, "The mere presence of a robot is enough to elicit social facilitation of human performance," *Scientific Reports*, vol. 12, no. 1, pp. 1–14, 2022.

[148] S. E. Guthrie, *Faces in the Clouds: A New Theory of Religion*. New York: Oxford University Press, 1993.

[149] A. Waytz, J. Cacioppo, and N. Epley, "Who sees human? the stability and importance of individual differences in anthropomorphism," *Perspectives on Psychological Science*, vol. 5, no. 3, pp. 219–232, 2010.

[150] D. C. Dennett, *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.

[151] L. Floridi, "The method of levels of abstraction," *Minds and machines*, vol. 18, no. 3, pp. 303–329, 2008.

[152] L. Floridi, *Information: A Very Short Introduction*. Oxford: Oxford University Press, 2010.

[153] L. Floridi, *The Ethics of Information*. Oxford: Oxford University Press, 2013.

[154] N. J. Emery, "The eyes have it: The neuroethology, function and evolution of social gaze," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[155] J. K. Hietanen, "Social attention orienting induced by eye gaze and head orientation," *Visual Cognition*, vol. 9, no. 1–2, pp. 1–22, 2002.

[156] D. R. Carney, A. J. C. Cuddy, and A. J. Yap, "Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance," *Psychological Science*, vol. 21, no. 10, pp. 1363–1368, 2010.

[157] M. Argyle, *Bodily Communication*. London: Methuen, 1975.

[158] G. Rhodes, "The evolutionary psychology of facial beauty," *Annual Review of Psychology*, vol. 57, pp. 199–226, 2006.

[159] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[160] A. P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, and C. Frith, "The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions," *Social Cognitive and Affective Neuroscience*, vol. 7, no. 4, pp. 413–422, 2012.

[161] T. Chaminade and T. Ohnishi, "Differentiating human and humanoid robot motion: Humans do not rely on dynamics," *Biological Cybernetics*, vol. 96, no. 5, pp. 477–489, 2007.

[162] R. E. Kleck and A. Strenta, "Perceptions of the gaze of another," *Journal of Personality and Social Psychology*, vol. 39, no. 5, pp. 725–732, 1980.

[163] J. K. Hietanen, "Does your gaze direction reflect your attention?," *Visual Cognition*, vol. 6, no. 1, pp. 97–120, 1999.

[164] A. P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, and C. Frith, "The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions," *Social Cognitive and Affective Neuroscience*, vol. 7, no. 4, pp. 413–422, 2012.

[165] M. Bateson, D. Nettle, and G. Roberts, "Cues of being watched enhance cooperation in a real-world setting," *Biology Letters*, vol. 2, no. 3, pp. 412–414, 2006.

[166] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Minds and Machines*, vol. 14, no. 3, pp. 349–379, 2004.

[167] S. Baron?Cohen and S. Wheelwright, "The empathy quotient: An investigation of adults with asperger syndrome or high-functioning autism, and normal sex differences," *Journal of Autism and Developmental Disorders*, vol. 34, no. 2, pp. 163–175, 2004.

[168] S. Baron?Cohen, J. Richler, D. Bisarya, N. Gurunathan, and S. Wheelwright, "The systemizing quotient: An investigation of adults with asperger syndrome or high-functioning autism, and normal sex differences," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1430, pp. 361–374, 2003.

[169] O. P. John, E. M. Donahue, and R. L. Kentle, "The big five inventory ? versions 4a and 5," tech. rep., Institute of Personality and Social Research, University of California, Berkeley, Berkeley, California, 1991.

[170] G. E. M. Anscombe, *Intention.* Oxford, UK: Blackwell, 1957.

[171] M. C. Nussbaum, *Upheavals of Thought: The Intelligence of Emotions.* Cambridge, UK: Cambridge University Press, 2001.

[172] C. Korsgaard, *Self-Constitution: Agency, Identity, and Integrity.* Oxford, UK: Oxford University Press, 2009.

[173] A. F. Shariff and A. Norenzayan, "God is watching you: Priming god concepts increases prosocial behavior in an anonymous economic game," *Psychological science*, vol. 18, no. 9, pp. 803–809, 2007.

[174] J. Haidt, "The new synthesis in moral psychology," *Science*, vol. 316, no. 5827, pp. 998–1002, 2007.

[175] J. Moll, R. de Oliveira-Souza, P. J. Eslinger, I. E. Bramati, J. C. Mourão-Miranda, P. A. Andreiuolo, and L. Pessoa, "The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions," *The Journal of Neuroscience*, vol. 25, no. 7, pp. 2730–2736, 2005.

[176] M. Fedyk, *The Social Turn in Moral Psychology.* Cambridge, MA: MIT Press, 2017.

[177] S. Baron-Cohen, "The extreme male brain theory of autism," *Trends in cognitive sciences*, vol. 6, no. 6, pp. 248–254, 2002.

[178] S. Baron-Cohen, J. Richler, D. Bisarya, N. Gurunathan, and S. Wheelwright, "The systemizing quotient: an investigation of adults with asperger syndrome or high–functioning autism, and normal sex differences," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1430, pp. 361–374, 2003.

[179] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[180] J. Złotowski, D. Proudfoot, K. Yogeeswaran, and C. Bartneck, "Anthropomorphism: Opportunities and challenges in human–robot interaction," *International Journal of Social Robotics*, vol. 7, no. 3, pp. 347–360, 2015.

[181] D. A. Levitis, W. Z. Lidicker, and G. Freund, "Behavioral biologists do not agree on what constitutes behavior," *Animal Behaviour*, vol. 78, no. 1, pp. 103–110, 2009.

[182] P. Lin, K. Abney, and G. A. Bekey, *Robot ethics: the ethical and social implications of robotics*. Intelligent Robotics and Autonomous Agents series, 2012.

[183] R. Oda, Y. Kato, and K. Hiraishi, "The watching-eye effect on prosocial lying," *Evolutionary Psychology*, vol. 13, no. 3, p. 1474704915594959, 2015.