Synthetic Presence as Moral Perturbation: A Field-Theoretic Model for Evaluating Moral Behaviour 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

November 2025

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).

There is a traveller who reaches a crossroads at the hour when the world withdraws into itself.

He studies the signposts as if they held the logic of direction. The boards are clean, the words exact, but the air is heavy with a silence that seems older than the road. A thin wind rises, carrying with it the odour of something distant—woodsmoke, or perhaps the memory of it. He cannot tell.

He believes he chooses by reading; but already his gaze has shifted toward the darker path, drawn by a murmur he cannot name. A shape in the periphery—almost a figure, almost a shadow—tilts the balance without ever declaring itself. The light changes, and with it the weight of each possibility.

He hesitates, though he is unaware of the reason. The stones cool beneath his feet. Something in the air—presence, or its simulation— presses lightly against his decision. He steps, not toward the sign he had resolved to follow, but toward the path shaped by these quiet, unclaimed forces.

Later he will recall the moment and speak of deliberation, of judgement, of intention:

- *I reasoned!*
- *I deliberated...*
- *I chose.*

But it was the quiet pressures of the world—the unseen gradients of light, sound, warmth, and presence—that shaped his path.

And the signs? They were there long before he arrived, and they remain long after he has gone. Yet it is the field through which he walked that carried him forward.

Francesco Perrone

Abstract

Moral behaviour emerges not from isolated cognitive modules or explicit reasoning, but from a structurally rich evaluative field shaped by attention, affect, social meaning, and dispositional architecture. This thesis develops and defends a field-theoretic account of moral cognition grounded in empirical evidence, formal topology, and philosophical analysis. It argues that artificial agents—particularly those with humanoid morphology—interact with this field in ways that classical Machine Ethics has systematically overlooked.

To test this, a controlled experiment examined how a humanoid robot (NAO) modulates prosocial donation under a strong moral cue (the Watching-Eye paradigm). Bayesian and regression models reveal a robust attenuation effect: participants donated less in the robot's presence, despite identical moral affordances. Personality- and cognitive-style measures (EQ, SQ, BFI-10) were used to derive three latent evaluative ecologies, each with distinct affective and structural properties. Yet all ecologies exhibited the same directional displacement. The robot did not influence moral principles; it altered the evaluative field through which those principles acquire behavioural force.

This finding supports a structural interpretation of moral cognition: synthetic presence acts as a perturbation operator that suppresses salience, dampens affective resonance, and disrupts justificatory and attentional pathways. The result exposes a critical limitation of top-down Machine Ethics and opens a new direction for Computational Morality—shifting focus from rule encoding to the dynamics of moral environments.

The thesis concludes that artificial agents, even without agency or intent, function as moral modifiers: their perceptual salience and ontological ambiguity reshape the architecture of human moral appraisal. Moral behaviour is thus field-dependent, and synthetic presence deforms that field. This establishes a new methodological foundation for the ethical and empirical study of artificial systems, grounded in evaluative topology, Levels of Abstraction, and the dynamics of moral cognition.

Contents

List of Tables

List of Figures

Acknowledgements

There is a peculiar stillness that settles around work completed under the accelerating discipline of contemporary academia—a sense that one has been guided less by patient inquiry than by the unyielding cadence of an institution convinced that thought must keep pace with its deadlines. If these pages read as though composed at a distance, it is only because they carry the faint tension between what might have matured in its own time and what the present era insists must be shaped, finished, and surrendered.

In that unsettled interval I have leaned on those whose presence does not depend on the coherence of my arguments. This thesis is dedicated to my son, Francesco, whose unguarded curiosity offers a quiet antidote to the rushed certainty demanded here; to my mother, Mirella, and my father, Alberto, whose enduring steadiness has outlasted every fluctuation of purpose; and to my wife, Anna, who has carried more than anyone her age should be asked to bear—not only for reasons that cannot be stated within these pages, but because she has been required, time and again, to return to the limits of my own intellect as though they were a place of refuge. Whatever this work may lack in the calm of true gestation, it rests on the grace with which they have all borne its cost.

Declaration

I declare that, with the exception of background sections that review existing literature and established theoretical frameworks (chapters 3, 6, and part of chapter 9) all original research presented in this thesis—including the experimental design, data collection, statistical analysis, clustering procedures, and the development of the evaluative–topological model of moral perturbation—was carried out independently by the author, unless otherwise explicitly stated. All sources have been appropriately acknowledged, and no part of this thesis has been submitted for any other degree or qualification.

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 22, 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.

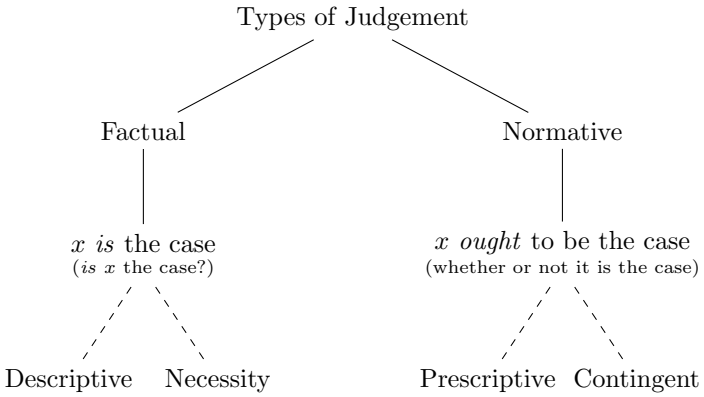


Figure 2.2: This distinction presuppose a sufficient prior understanding of the relevant uses of *is* and *ought* (or *should*) which will not discuss in details here. It is important to notice that, the presence of such marks as is neither a sufficient nor necessary criterion for the distinction we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

3.1 Introduction: Scope, Objectives, and Theoretical Commitments

This chapter establishes the conceptual and methodological terrain on which the remainder of the thesis proceeds. Its central aim is to reposition the study of moral behaviour under artificial co-presence—and the design of artificial moral systems more broadly—within a theoretically unified space at the intersection of *Machine Ethics*, *Computational Morality*, and *Social Signal Processing* (SSP). Although these fields emerged from distinct disciplinary lineages, the experimental results presented in Chapter 7 show that they now converge around a single problem: artificial agents, even when silent, passive, and non-interactive, *modulate the evaluative conditions under which moral judgment and action unfold*. Understanding this phenomenon requires an integration of normative philosophy, moral psychology, computational modelling, and HRI.

Machine Ethics, in its canonical formulation, begins at the reflective end of the normative spectrum. Its foundational gesture is to *encode* the principles of ethical theories—Kantian universalisability tests [16, 121], utilitarian optimisation regimes [122], virtue-theoretic character templates [123], or deontic logics of permission and obligation [124]—and to treat these encodings as viable models of moral *agency*. This approach presupposes that the normative structure of ethical theories can function as a generative model of behaviour. Yet, as both classical moral philosophy and contemporary moral psychology make clear, ethical theories articulate *conditions of justification*, not *mechanisms of cognition*. Their principles operate at a high Level of Abstraction (LoA), whereas the production of moral action in humans occurs at a low LoA shaped by perception, affect, salience, attention, and social meaning. Machine Ethics thereby commits a category error: it collapses reflective normativity into computational operability.

Because this is the first point in the thesis where Floridi's notion of a *Level of Abstraction* appears, a brief clarification is required.[1] In Floridi's framework, an LoA specifies the *vantage point* from which a system is described: the variables made observable, the granularity of representation, and the kinds of explanations that are admissible at that level [125, 126]. Different LoAs support different types of claims. Normative theories operate at a high LoA concerned with justificatory structure; psychological models operate at a lower LoA concerned with mechanisms of appraisal, affect, and attention. Treating principles defined at one LoA as if they were mechanisms at another produces exactly the sort of explanatory

[1] A full treatment is provided in Chapter 8.

inversion that this thesis seeks to correct. The next chapter develops LoA discipline in detail; here, the concept is introduced only to mark the methodological boundary that Machine Ethics routinely oversteps.

Computational Morality inherits this error in a different idiom. Whether formulated through logic-based evaluation engines [127, 128], preference aggregation architectures [129], or LLM-based normative modelling [9, 10], the field tends to assume that moral behaviour can be approximated by propositional inference or symbolic rule execution. Yet decades of empirical work demonstrate that moral judgment is *not* primarily a deliberative phenomenon: it is a fast, affectively inflected, context-sensitive process [88, 58, 57, 130, 82, 75] that integrates intuitive appraisal with socially modulated evaluation [58, 88, 57]. The mechanistic substrate of moral cognition is thus topological rather than logical: it consists of gradients of salience, attractor dynamics of affect, attentional pathways, and context-contingent modulation, none of which is captured by propositional models of moral reasoning. Computational Morality, in its traditional form, therefore abstracts away precisely the cognitive machinery that your experiment reveals to be modulated by synthetic presence.

This chapter proceeds from the methodological premise established in the Morality Primer and the Ethical Cognition chapter: *any model of moral behaviour that ignores the cognitive–affective and social-signalling architectures of moral judgment is descriptively inadequate and normatively unstable.* As the LoA clarification above suggests, mismatching levels of description—treating high-LoA normative content as if it were low-LoA psychological mechanism—produces artefactual theories that neither predict behaviour nor illuminate perturbation phenomena. The experimental results in Chapter 7 demonstrate this problem in its most acute form: a robot with no beliefs, no goals, and no communicative acts nevertheless *modifies* the evaluative conditions under which human agents convert moral perception into prosocial action. Such a phenomenon cannot be captured by any framework that treats morality as propositional reasoning, utility optimisation, rule-following, or explicit norm retrieval.

The present chapter therefore undertakes a task that no existing Machine Ethics literature has successfully achieved: a *structurally disciplined integration of normative theory, empirical psychology, and computational modelling.* It shows that to understand moral behaviour under synthetic presence, one must reframe the foundational assumptions of the field. Moral norms must be analysed through their *topological role* in shaping evaluative constraints; the cognitive processes that produce moral judgments must be situated within the affective, attentional, and social architectures documented in empirical psychology; and artificial agents must be modelled not as bearers of rules or values but as *operators that modulate the evaluative field.*

What follows is therefore more than a literature review. It is a reconstruction of the conceptual lineage necessary to render the experimental findings intelligible. The chapter develops a unified framework grounded in four pillars:

1. **Normative–Philosophical Foundations**: clarifying the role of deontological invariants, consequentialist gradients, virtue-theoretic disposi-

tions, sentimentalist affective vectors, and contractualist justificatory structures—each reconceived through evaluative topology [131, 132, 133, 82].

2. **Empirical Moral Psychology**: synthesising dual-process theory, the Social Intuitionist Model, affective tagging, attentional capture, and social modulation as the core mechanisms by which moral salience becomes behaviour.

3. **Social Signal Processing and Affective Computing**: grounding the argument in the extensive evidence that social cues—including gaze, co-presence, morphology, and synthetic embodiment—systematically perturb evaluation, expectation, and action.

4. **Critical Reassessment of Machine Ethics and Computational Morality**: demonstrating why their monolithic, principle-first architectures are methodologically unsound and why any future model of moral behaviour must begin with the structural, field-level properties revealed in this thesis.

Through this triangulated lens—normative, psychological, and computational—the moral displacement effect demonstrated in the experiment becomes theoretically transparent. Artificial agents do not "make moral decisions" in any traditional sense; instead, they *reshape the evaluative conditions under which humans do*. This chapter situates that insight within the broader intellectual landscape and establishes the theoretical commitments that guide the remainder of the thesis.

Social Signal Processing provides the empirical and theoretical bridge that reveals *why* the attenuation observed in the experiment cannot be understood through abstract ethical theory alone. Work on nonverbal regulation of social meaning [131, 134], anthropomorphisation and norm-perception in HRI [135, 136, 137], and affective appraisal dynamics [133] shows that human moral behaviour is continuously shaped by *attentional capture, empathic resonance, and perceived social evaluation*. These mechanisms operate at the *cognitive Level of Abstraction* [125]—the LoA at which perceptual, affective, and inferential processes modulate the evaluative topology that determines behavioural output. Importantly, this LoA sits *below* the reflective LoA at which ethical theories articulate reasons, duties, or values. The result is a structurally asymmetric architecture: normative principles do not directly cause behaviour; rather, behaviour emerges from the field-like interaction of salience, affect, and social interpretation.

Seen through this lens, the experimental attenuation revealed in Chapter 7 becomes a diagnostic probe into the moral field itself. The humanoid robot does not "apply a principle" nor transmit a norm; instead, it subtly transforms the perceptual–social environment in which moral appraisal unfolds. It reshapes the evaluative gradients through which the Watching-Eye cue gains its normative traction. A scientifically credible framework for moral AI must therefore begin from this empirical architecture: the dynamics of salience, affect, embodiment, and co-present social meaning—not from the reflective content of ethical theory. This is the central commitment of the present chapter.

The chapter proceeds by articulating six interlocking arguments that establish

the methodological and conceptual ground for a new approach to moral AI:

1. **Machine Ethics consists of two structurally distinct research programmes**, whose conflation has hindered both scientific progress and philosophical clarity.
2. **Moral psychology reveals a dual-process, affectively embedded cognitive architecture** that is incompatible with treating moral judgement as purely rule-based or propositional.
3. **Classical Machine Ethics fails because it collapses Levels of Abstraction**: it assumes that normative structures at a reflective LoA can serve as computational operators at a cognitive LoA.
4. **Moral behaviour arises from an evaluative topology**, not from internal execution of ethical principles: perception, affect, and social meaning generate the gradients that guide action.
5. **Synthetic presence perturbs this topology in measurable ways**, attenuating prosocial output by deforming the salience field rather than altering explicit reasoning.
6. **A viable programme for moral AI must therefore model how machines reshape human moral experience**, before attempting any normative codification or ethical implementation.

3.2 The Two Research Projects in Machine Ethics

Machine Ethics, as originally formulated by Moor [14] and developed by Wallach and Allen [123], is not a single unified research field. It is a composite of two methodologically divergent projects whose aims, assumptions, and Levels of Abstraction are frequently confounded. The first, *Human–Machine Ethics*, concerns the normative governance of human behaviour *around* and *in the presence of* artificial systems: agency displacement, accountability, social influence, and moral risk. The second, *Computational Machine Ethics*, attempts to design machines that themselves make morally adequate decisions by embedding ethical theories into computational architectures.

The experimental results of this thesis show that conflating these projects is not merely a conceptual error—it produces empirical blindness. Human–Machine Ethics speaks directly to the phenomenon the experiment uncovers: the modulation of human moral behaviour by artificial co-presence. Computational Machine Ethics largely ignores this modulation, because it begins from the false premise that moral behaviour can be generated by encoding reflective ethical theories.

Human–Machine Ethics aligns naturally with research in HRI, media psychology, and SSP. Studies on robotic social facilitation [137], anthropomorphic cueing [135], norm-perception under artificial observation [136], and social attention modulation [134] demonstrate that artificial agents influence human moral behaviour even in the absence of explicit interaction. These findings are central to the present thesis: moral environments are *ethically load-bearing* by virtue of

their perceptual and affective structure alone. Classical Machine Ethics has never incorporated this fact.

Computational Machine Ethics, by contrast, emerged from attempts to implement ethical theories directly in artificial systems. Deontic logics [124, 16], utilitarian optimisers [122], and virtue-based architectures [123] all assume that moral behaviour can be generated by applying normative principles at runtime. Yet this assumption conflicts with decades of empirical evidence showing that human moral cognition is primarily intuitive, affective, and context-sensitive [88, 58, 57]. Attempts to encode ethical theories as computational operators therefore violate LoA discipline: they treat reflective normative content as if it were a psychological mechanism.

The experiment demonstrates that this assumption is untenable. Synthetic presence $\mathscr{R}$ systematically perturbs the evaluative topology that governs intuitive appraisal, attentional salience, empathic resonance, and contextual interpretation. Moral behaviour changes because the *field* changes—not because an abstract principle is executed differently [138, 139]. This insight directly implicates Human–Machine Ethics in the study of moral behaviour under artificial co-presence and reveals the inadequacy of Computational Machine Ethics as a behavioural model.

A scientifically coherent future for moral AI therefore requires recognising that Machine Ethics contains two distinct explanatory projects, operating at different Levels of Abstraction and constrained by different empirical realities. Only by respecting this structural division can normative theory, cognitive science, and computational modelling be brought into principled alignment.

Taken together, Sections 3–1 and 2 establish the methodological reorientation that grounds the remainder of this chapter. The analysis of Machine Ethics as two structurally distinct research programmes reveals a foundational misalignment: Computational Machine Ethics operates at an inappropriate Level of Abstraction (LoA), treating reflective normative structures as if they were generative psychological mechanisms, while Human–Machine Ethics has been conceptually underutilised despite its direct relevance to the empirical modulation of moral behaviour. Crucially, the experimental findings in Chapter 7 reinforce this structural diagnosis. A silent, non-agentic robot altered participants' donation behaviour *even under a strong moral prime* (the Watching-Eye cue), consistent with the robust social-monitoring effects documented across field and laboratory studies [140, 141, 142]. These findings demonstrate that synthetic presence reshapes the perceptual–affective conditions under which moral salience becomes action. No framework that models moral behaviour as propositional rule retrieval or principle execution can account for such a field-level deformation.

Having clarified the conceptual landscape and identified the methodological failures that motivate a systematic reframing of Machine Ethics, the next step is to specify the theoretical resources required to reconstruct a scientifically grounded model of moral cognition. Sections 3 and 4 therefore articulate the psychological and philosophical commitments that classical Machine Ethics overlooked: the dual-process structure of moral cognition; the cognitive–affective substrate of evaluation; and the LoA discipline that distinguishes reflective justification from

cognitive mechanism. These sections also provide the conceptual infrastructure needed to understand *why* the experiment produced a uniform attenuation effect across participants and *why* that attenuation is best interpreted as a deformation of the evaluative topology rather than as a deviation from principle, preference, or conscious moral reasoning.

3.3 Moral Psychology and Moral Philosophy: Cognitive–Affective vs. Rationalist–Intuitionist Models

Any attempt to model moral behaviour—whether human or artificial—must begin from a clear understanding of the psychological architecture through which moral judgment is produced. Decades of empirical work demonstrate that moral cognition is not primarily a matter of propositional reasoning or deliberative inference, but a *dual-process system* in which intuitive, affectively inflected pathways interact with slower, reflective reasoning. The Social Intuitionist Model [88], Greene's neurocognitive dual-process account [58, 143, 90], and Cushman's action-based inference framework [57] converge on a common insight: moral evaluation originates in rapid, affectively loaded appraisals shaped by perceptual salience, empathy, and attentional dynamics. Affective tagging and empathy architectures [82, 130] show that moral judgments emerge from mechanisms integrating motivational relevance, value-laden salience, and social meaning long before explicit reasoning becomes operative.

This cognitive–affective model stands in sharp contrast with the rationalist traditions that have historically informed Machine Ethics. Kantian, utilitarian, and contractualist theories articulate *justificatory structure*: universalisability conditions, value aggregation, or reason-giving relations. These theories operate at a reflective LoA concerned with *why* actions are justifiable, but they do not—and were never intended to—describe *how* moral judgments are generated within actual cognitive systems. As Sidgwick [144], Scanlon [44], and Korsgaard [117] emphasise, justification belongs to the domain of reflective endorsement, not causal explanation.

Machine Ethics inherited only one side of this philosophical division. By adopting ethical theories as computational templates, it implicitly treated rationalist structures as if they were generative psychological mechanisms. Yet the empirical architecture of moral cognition does not support this view. Moral behaviour is governed by attentional capture, affective resonance, and socially modulated salience [145, 142, 146]—precisely the mechanisms perturbed in the experiment. The Watching-Eye effect [141, 140, 138, 139] operates not through rule endorsement but through implicit monitoring, affective vigilance, and social-evaluative sensitivity. This is the cognitive LoA at which empathy, salience, and contextual modulation operate [82, 90, 75]. The attenuation effect observed in Chapter 7 thus belongs squarely within the cognitive–affective architecture of moral psychology, not within any reflective principle or moral theory.

This insight has direct implications for Computational Machine Ethics. If moral cognition fundamentally arises from evaluative salience, attentional modulation,

and affective integration, then any computational model that treats moral behaviour as propositional inference or rule execution is descriptively incomplete. The next section articulates the methodological consequences of this fact through the concept of Level-of-Abstraction discipline.

3.4 Levels of Abstraction and the Failure of Machine Ethics

Floridi's notion of a *Level of Abstraction* (LoA) provides the conceptual apparatus required to diagnose the foundational error in classical Machine Ethics. An LoA specifies the *observables*, the *granularity*, and the *explanatory constraints* appropriate to a given description of a system [125, 126]. Normative theories occupy a high LoA: they articulate justificatory relations, principles of right action, and standards of rational agency. Moral psychology, by contrast, operates at a lower LoA, modelling the mechanisms through which perception, affect, salience, attention, and social meaning generate evaluative trajectories. Treating the former as if it directly governed the latter collapses the boundary between reflective and cognitive explanation.

Machine Ethics systematically commits this collapse. Deontic architectures [124, 16], utilitarian optimisation systems [122], and virtue-based computational agents [123] assume that principles defined at a reflective LoA can serve as *behavioural generators*. Logic-based systems treat moral judgment as derivation; rule-governed architectures treat it as constraint propagation; utility-based systems treat it as maximisation. All presume that moral behaviour is produced by algorithms executing normative content. Empirical cognitive science directly contradicts this assumption: moral behaviour is generated by affective appraisal [130], empathic resonance [82], attentional selection [146], and social-contextual modulation [134, 141]. These mechanisms constitute the *evaluative topology* within which reasons, norms, and values acquire behavioural force.

The experiment demonstrates the consequences of violating LoA discipline. The Watching-Eye cue introduces a deontic expectation of accountability, yet the behavioural manifestation of this expectation is attenuated by the presence of a silent, non-agentic robot. This perturbation does not occur at the level of explicit reasoning or principle application; it occurs at the level of *salience fields*, *affective gradients*, and *implicit social ontology*. Studies in HRI show that robots alter human social evaluation even in minimal-interaction contexts [137, 135, 136, 147, 148]. Studies in SSP show that social cues reshape evaluation and behaviour via perceptual and affective channels [131, 132, 133]. The attenuation observed in the experiment is therefore a paradigmatic case of cross-LoA interference: a reflective norm (accountability) fails to exert its expected behavioural influence because the cognitive LoA at which it is realised has been deformed.

Interim Synthesis. The analyses developed in Sections 3 and 4 converge on a single structural insight: classical Machine Ethics begins at the wrong place in the explanatory hierarchy. It treats reflective normative theories—Kantian maxims, utilitarian utilities, virtue-trait encodings, and rule-based constraint systems—as if they were *mechanisms* of moral cognition, thereby collapsing Levels of Abstraction and obscuring the architectures that actually generate moral behaviour.

Moral psychology and SSP reveal a different picture: moral action emerges not from the execution of principles but from the dynamic interaction of perceptual salience, affective resonance, attentional capture, embodiment, and social modulation. These low-LoA mechanisms constitute the *evaluative topology* within which reasons, values, and obligations acquire behavioural force.

Seen through this lens, the experiment in Chapter 7 functions as a decisive test case. A silent, non-agentic robot attenuates prosocial donation even in the presence of a strong accountability cue. This attenuation cannot be explained by any model that treats moral behaviour as propositional inference or rule retrieval. Instead, it discloses a deformation of the evaluative field itself: a disruption of empathic salience, a weakening of the Watching-Eye gradient, and a uniform directional shift across dispositional ecologies. The phenomenon is therefore diagnostic not of failed principle-application but of a perturbed cognitive–affective substrate.

The thesis therefore adopts a clear and principled stance: *before we can design moral machines, we must understand how machines reshape human moral experience.* Classical Machine Ethics begins at the *wrong end* of the explanatory hierarchy: it starts with reflective normative theory and assumes that these principles can be used as behavioural generators. A scientifically credible programme for moral AI must invert this order. It must first identify the empirical architecture that makes moral behaviour possible; second, model how artificial systems modulate the evaluative field in which that behaviour unfolds; and only then, on this empirically grounded basis, determine what forms of normative oversight or ethical design are justified.

This reframing provides the conceptual orientation for the remainder of the chapter. The next sections examine the psychological, philosophical, and computational structures that classical Machine Ethics could not accommodate: the cognitive–affective machinery of moral judgment, the topological configuration of evaluative forces, and the cross-LoA perturbations induced by artificial copresence. These resources form the foundation for a new, empirically grounded and topologically disciplined account of moral behaviour under synthetic presence.

3.5 Evaluative Topology, Affective Architecture, and Synthetic Moral Perturbation

If Sections 3 and 4 establish that classical Machine Ethics fails by mislocating moral cognition at the wrong Level of Abstraction, the present section articulates the positive theoretical alternative: a *topological* account of moral behaviour grounded in the cognitive–affective mechanisms identified by empirical psychology and Social Signal Processing. The central claim is that moral behaviour is governed not by the internal execution of principles, but by the dynamic configuration of an *evaluative field*: a structured, multi-dimensional manifold shaped by gradients of salience, affective resonance, attentional pathways, contextual norms, and implicit social meaning. Ethical theories operate within this field as *structural constraints*, not as generative mechanisms.

The notion of an evaluative topology synthesises several strands of empirical and philosophical insight. From moral psychology, we inherit the idea that affective and intuitive systems provide the primary route through which motivational salience attaches to moral cues [88, 58, 130]. From SSP and affective computing, we gain the recognition that social cues—including gaze, morphological salience, co-presence, and implicit monitoring—modulate the perceptual filters and affective weights through which situations are evaluated [131, 132, 133]. From philosophy, we draw on the structural character of normative theories: deontological invariants, consequentialist gradients, virtue-based dispositional attractors, sentimentalist affective vectors, and contractualist justificatory equilibria. Reinterpreted through an LoA-sensitive framework, these are not rules to be executed but *roles* played within a larger evaluative topology.

Within this topological framework, moral behaviour emerges from trajectories through the evaluative manifold. Attention generates local curvature; affect saturates regions of the field with motivational orientation; contextual cues deform or amplify gradients; and implicit social signals modulate the salience of norms, duties, and empathic responses. This framework dissolves the traditional philosophical opposition between rationalist and intuitionist models: rationalist structures become constraints on admissible trajectories, while intuitionist mechanisms determine the field through which those trajectories unfold. It also resolves the computational impasse: behaviour is not produced by any principle-encoding module, but by the system's real-time navigation through a salience-weighted topology.

The experiment presented in Chapter 7 provides a decisive empirical probe into this architecture. The Watching-Eye cue creates a prosocial salience gradient by activating implicit accountability mechanisms [141, 140, 146]. Yet the introduction of a silent, non-agentic robot attenuates this gradient despite the absence of communicative intent or normative instruction. The perturbation arises because the robot's ambiguous social ontology—perceptually agentic, ontologically indeterminate—reshapes the affective–attentional conditions through which the Watching-Eye cue exerts its effect. In this sense, synthetic presence functions as a *field operator*: it modifies the salience landscape and thus alters the pathways through which moral perception becomes moral action.

Crucially, this perturbation is *layer-sensitive*. The Prosocial–Empathic ecology showed the strongest attenuation because its evaluative topology depends most heavily on empathic resonance and interpersonal salience—the very mechanisms displaced by the robot's presence. The Analytical–Structured ecology showed moderate attenuation because its field relies more on interpretive stability than affective reactivity; the robot destabilised its sense-making substrate without significantly disrupting affective curvature. The Emotionally Reactive ecology showed minimal change because its evaluative topology lacks stable gradients; perturbation at this LoA therefore produces negligible displacement. These differential effects underscore the central insight: *synthetic presence perturbs the evaluative topology upstream of both principle and trait.*

This topological reading clarifies why classical Machine Ethics could not pre-

dict the observed phenomenon. Moral behaviour changes not because an encoded principle is misapplied, but because the evaluative field in which principles would acquire force has been deformed. Deontological norms cannot motivate action when the salience of accountability is suppressed; consequentialist gradients flatten when contextual meaning becomes ambiguous; virtue-based dispositions fail to express themselves when affective attractors lose curvature; sentimentalist mechanisms weaken when empathic resonance is displaced; and contractualist justificatory relations dissolve when the perceived social field becomes indeterminate. The experiment thus confirms the central thesis of this section: moral behaviour is *field-sensitive*, and synthetic agents act as *perturbation operators* on the evaluative topology.

Evaluative topology therefore provides the structural bridge needed to integrate normative theory, empirical psychology, and computational modelling. It offers a principled account of how high-LoA normative structures interface with low-LoA cognitive–affective machinery and explains why synthetic presence can reshape moral behaviour without expressing beliefs, intentions, or normative content. This sets the stage for the final section, which synthesises the chapter's arguments into a unified model of machine-mediated moral cognition.

3.6 Integrative Synthesis: Toward a Cognitive–Affective Model of Machine-Mediated Morality

The analyses developed across Sections 1–5 point toward a unified account of moral behaviour under synthetic presence. Classical Machine Ethics fails because it begins with reflective normative theories and assumes that these can serve as behavioural generators. Moral psychology shows that moral behaviour emerges from a cognitive–affective architecture defined by salience, attention, empathy, and contextual modulation. SSP and HRI demonstrate that artificial agents modulate these mechanisms even in the absence of interaction. Evaluative topology integrates these insights by representing moral behaviour as the product of trajectories through a salience-weighted, affectively structured evaluative field. The experiment reveals that synthetic presence perturbs this field, deforming the conditions under which moral salience becomes action.

This integrated framework yields three core claims that define the thesis's contribution to the philosophy of moral AI:

1. **Moral behaviour is generatively rooted in the cognitive LoA.** Reflective ethical theories articulate justificatory structure, but moral action emerges from cognitive–affective dynamics. Ethical norms acquire behavioural force only when the evaluative field enables them to do so.

2. **Artificial agents reshape the evaluative field before they act within it.** Studies in HRI and SSP [136, 135, 137, 147, 131, 132] show that synthetic presence modulates attention, empathy, and social meaning. The experiment confirms that these perturbations are sufficient to attenuate moral behaviour even when normative content remains unchanged.

3. **A scientifically credible programme for moral AI must begin with**

evaluative topology, not normative codification. The field must reverse the methodological order: start with the empirical structure of moral cognition; analyse how artificial systems modulate this structure; and only then determine how normative oversight or ethical design may be justified. Computational systems cannot generate moral behaviour by executing principles unless the evaluative field has the curvature required to make such principles operative.

This synthesis reframes the foundational commitments of moral AI research. Artificial systems should not be conceptualised as bearers of ethical rules or evaluative principles; they should be modelled as *operators on the evaluative field* in which human moral cognition unfolds. The experiment demonstrates that even minimal robotic presence deforms salience gradients, attenuates empathic resonance, and weakens interpersonal accountability cues. These perturbations occur at the cognitive LoA, far upstream of explicit reasoning or reflective endorsement.

The result is a conceptual and methodological paradigm shift: *moral AI must be grounded in the science of moral cognition, not in the logic of ethical theory.* Synthetic moral perturbation is not an anomaly; it is the natural consequence of a multi-layered architecture in which perceptual salience, affective openness, and contextual interpretation shape the pathways through which moral meanings become moral actions. Evaluative topology provides the theoretical vocabulary needed to describe this architecture, diagnose its vulnerabilities, and design artificial systems that stabilise rather than distort the moral field.

With this framework in place, the final part of the thesis articulates a global synthesis: a unified model of moral perturbation and ethical interpretation that consolidates the normative, psychological, and topological analyses developed across the project.

3.7 Global Synthesis: From Inferential Displacement to Synthetic Moral Topology

The literature reviewed across this chapter establishes a series of convergent insights that collectively redefine the conceptual landscape for thinking about moral behaviour under artificial co-presence. Rather than offering isolated conclusions from discrete subfields, the evidence forms a coherent picture:

"*moral judgement and moral action emerge from a cognitively embedded, affectively structured, socially modulated evaluative field.*"

Ethical theories supply reflective standards, but they do not generate behaviour; cognitive architecture does. Artificial agents, even when minimally interactive, participate in this architecture by modulating salience, affect, attention, and perceived social meaning. These insights, taken together, provide the intellectual foundation on which the remainder of the thesis builds.

From Question to Framework: A Literature-Grounded Progression

The guiding research question that motivates the empirical work of the thesis—whether synthetic presence can perturb the inferential transformation through which moral salience becomes action—arises directly from unresolved tensions in the literature. Classical Machine Ethics assumes that moral behaviour follows from the execution of encoded principles [124, 16]. Moral psychology shows that it does not [88, 58, 57]. Social cognition and SSP demonstrate that social cues modulate moral behaviour via attentional capture and affective appraisal [131, 132, 145]. HRI studies reveal that artificial agents trigger many of these same modulations [135, 136, 137]. Yet these strands have not been integrated into a unified account capable of explaining *how* artificial co-presence reshapes moral behaviour.

The inferential displacement question is therefore not merely a research convenience; it emerges as the *necessary organising problem* at the nexus of these literatures, where normative theory, moral psychology, computational modelling, and social robotics intersect.

Why a Multi-Hypothesis Framework Was Needed

The transition from a single research question to a structured hypothesis set is mandated by the heterogeneity of mechanisms identified in the literature. Existing research suggests at least three theoretically distinct pathways through which artificial agents may modulate moral behaviour:

1. **Evaluative deformation:** Studies on gaze cues, reputation systems, and social monitoring show that attentional and affective gradients can be reshaped by subtle cues [141, 140, 146]. This literature suggests that synthetic presence may alter the curvature of the evaluative field.
2. **Synthetic normativity:** Research in HRI demonstrates that artificial agents can acquire quasi-normative status by virtue of their perceived social ontology [147, 135, 136]. This literature suggests that robots may introduce new justificatory expectations without explicit agency.
3. **Perturbation of inferential pathways:** Neurocognitive and affective models show that moral judgements rely on the integration of empathic resonance, attentional selection, and contextual interpretation [82, 90]. Artificial presence may displace these pathways by introducing competing social signals.

The literature therefore requires a multi-hypothesis framework: different domains predict different modes of perturbation, and no single mechanism is sufficient to explain the range of modulations documented in HRI, SSP, and moral psychology.

Theoretical Implications from the Literature Alone

When synthesised, the reviewed evidence supports three high-level claims about moral behaviour under artificial co-presence—claims that follow from *the literature itself*, independent of any empirical findings of the present thesis:

1. **Moral behaviour is field-sensitive.** Decades of work in moral psychology, affective neuroscience, and social cognition show that moral action emerges from a structured evaluative field defined by salience, attention, affect, and context [58, 88, 82].
2. **Artificial agents modulate this field.** HRI and SSP research demonstrate that artificial co-presence reshapes social meaning, perceived accountability, empathic stance, and attentional allocation [136, 137, 134, 147]. Artificial systems therefore participate in the cognitive LoA rather than the normative one.
3. **Machine Ethics, as traditionally formulated, cannot model this modulation.** Encoding ethical principles at the reflective LoA cannot account for cognitive–affective modulation at the perceptual LoA [123, 122]. The literature makes clear that moral behaviour cannot be generated, predicted, or regulated solely by propositional structures.

These points converge on a pivotal conclusion: a viable framework for moral AI must be grounded not in the normative *content* of ethical theories but in the *structural dynamics* of the evaluative field.

Conceptual Synthesis Box (Reframed as Literature-Driven)

Conceptual Synthesis: Why Prosocial Behaviour Is the Right Lens for Moral Modulation Studies

Across moral psychology, social cognition, HRI, and SSP, converging evidence shows that prosocial behaviour is the behavioural endpoint of a cognitive–affective architecture that transforms moral salience into action. Social cues, attentional signals, empathic triggers, and embodied co-presence modulate this architecture continuously. Because moral cognition is inherently practical, perturbations to its perceptual, affective, or interpretive components manifest most reliably in behavioural output. This literature therefore establishes prosocial action as a methodologically robust proxy for detecting how artificial co-presence reshapes the evaluative field in which moral meaning becomes behaviourally operative.

Final Literature-Grounded Transition

Through this synthesis, the chapter completes the theoretical task appropriate to a literature review: it maps the conceptual space, identifies structural tensions, and isolates the explanatory gaps that motivate empirical investigation. The review establishes that understanding moral behaviour under artificial co-presence requires an integrated framework grounded in cognitive architecture, affective dynamics, social signalling, and Level-of-Abstraction discipline. This framework sets the stage for the empirical analysis that follows, providing both the theoretical motivation and the interpretative resources required to examine how synthetic presence modulates the evaluative conditions under which moral perception becomes moral action.

Literature-Driven Conclusion: Why Moral Behaviour Must Be Studied as a Field-Sensitive Phenomenon

The interdisciplinary literature reviewed in this chapter converges on a single structural insight: *moral behaviour cannot be explained by principle-based models alone.* Across moral psychology, social cognition, developmental studies, HRI, and Social Signal Processing, converging evidence shows that moral action emerges from the interaction of perceptual salience, affective appraisal, attentional modulation, embodied cues, and socially mediated meaning. Ethical theories provide reflective standards for justification, but the *production* of moral behaviour arises at a lower Level of Abstraction—within a cognitive– affective evaluative field shaped by context, embodiment, and social signals.

This literature reveals three core commitments that guide the remainder of the thesis: (i) moral behaviour is generated by field-like structures of evaluation rather than by the execution of encoded principles; (ii) artificial agents, even when passive or minimal in appearance, can reshape these fields by altering attentional, affective, and social conditions; and (iii) a scientifically credible framework for moral AI must therefore begin from an empirically grounded model of moral cognition, not from the direct implementation of normative theory.

These commitments reposition Machine Ethics within a broader scientific landscape: moral behaviour must be analysed as *field-sensitive*, *contextually modulated*, and *cognitively grounded.* Any artificial system capable of influencing human decision-making thereby participates—intentionally or not—in the structural conditions under which moral salience becomes action. Understanding this participation is the central task that motivates the chapters that follow.

Through this synthesis, the chapter completes the theoretical arc that began with the research question of inferential displacement. It demonstrates that synthetic agents reshuffle the very conditions under which moral salience becomes action, and it establishes the empirical and philosophical foundation on which the remainder of the thesis builds its general theory of moral perturbation.

4. MORALITY PRIMER FOR COMPUTER SCIENTISTS

4.1 Why This Chapter Exists

Research in human–robot interaction, affective computing, and artificial intelligence routinely engages with moral concepts. Yet technical treatments of morality often rely on folk-theoretical assumptions, intuitive definitions, or operational proxies that lack philosophical and psychological grounding.

In Floridi's framework, a *Level of Abstraction* (LoA) denotes the set of observables, modelling choices, and epistemic constraints under which a system is described and analysed [149, 126]. An LoA determines what counts as information, what distinctions can be made, and which questions are meaningful within a given descriptive or normative domain. Crucially, different LoAs support different inferential structures: a psychological LoA describes cognitive regularities, a normative LoA prescribes what agents *ought* to do, and these cannot be interchanged without committing a methodological error [150, **?**, **?**]. Attending to LoAs therefore provides the conceptual machinery needed to diagnose the kinds of confusions that arise when technical research invokes moral terminology without theoretical grounding.

Against this background, two systematic conceptual errors. First, *category mistakes*: treating morality as a set of externally codifiable rules, conflating ethical norms with behavioural conventions, or assuming that computational tractability licenses normative reduction [42, 151]. Second, *level-of-abstraction confusions*: importing normative notions into descriptive models, or conversely, construing psychological regularities as ethical principles [149, 150].

Both errors impair empirical interpretation in human–robot interaction and distort theoretical proposals in Machine Ethics, where the distinction between *moral agency*, *normative impact*, and *behavioural modulation* is frequently collapsed [152, 4]. Without a precise account of what constitutes a moral judgment, how such judgments differ from other evaluative processes, and how moral cognition interacts with affective and social mechanisms, researchers risk mischaracterising the very phenomena they aim to measure or engineer.

The purpose of this chapter is therefore clarificatory. It provides a rigorous, minimally sufficient conceptual primer tailored to computer scientists and engineers. The chapter does not advance normative arguments, nor does it attempt to resolve ethical debates. Its aim is to supply the conceptual scaffolding required for understanding the empirical and theoretical contributions that follow. The framework adopted here positions moral cognition as an *action-guiding* evaluative process, situated within a broader cognitive–affective ecology [60, 65, 153]. This orientation ensures that later discussions—especially those concerning moral perturbation under robotic presence—rest upon analytically coherent foundations

rather than inherited ambiguities.

4.2 What Morality Means

4.2.1 Descriptive and Normative Domains

The term "morality" spans at least two analytically distinct domains. The first is *descriptive morality*: the empirical study of how humans form moral judgments, experience moral emotions, and engage in normatively salient actions. This includes developmental psychology [154], social–cognitive models [155, 156], affective neuroscience [75, 82], and evolutionary accounts of cooperation and prosociality [2, 3]. The second is *normative morality*: the domain of ethical theorising concerned with how one ought to act. This domain encompasses deontological, consequentialist, contractualist, and virtue-theoretic traditions [35, 94, 157, 158].

These domains are distinct but interdependent. Descriptive accounts illuminate how agents actually evaluate and respond to situations, while normative theories articulate standards for justified action. Empirical models of moral cognition acquire meaning partly through the normative vocabulary within which moral judgments are articulated, while normative theories must remain constrained by what agents are psychologically capable of performing or understanding.

In this thesis, the primary focus remains on *descriptive* moral cognition, though normative materials are used to clarify the structure and function of moral judgment. The distinction is maintained rigorously to prevent importing normative assumptions into empirical constructs or misinterpreting behavioural outcomes as moral prescriptions.

4.2.2 Why Definitions Vary

There is no single universally accepted definition of morality. Divergence arises because different research programmes emphasise different components of the moral domain: cognitive mechanisms [60], affective systems [153], normative reasoning [158], social norms [**?**], or evolutionary functions [2, 3]. Philosophical traditions likewise disagree on whether morality is grounded in rationality, sentiment, virtue, utility, social contracts, or evolutionary pressures.

Computational treatments often default to rule-based perspectives not because such accounts reflect human cognition but because they are structurally convenient to implement. This convenience has contributed to misleading interpretations of moral behaviour as rule following [159, 88, 160, 161], and has encouraged oversimplified models of moral decision-making [162, 59, 163, 164, 165, 166]. A primary goal of this chapter is to replace such inherited simplifications with a framework grounded in contemporary moral psychology and cognitive science.

4.2.3 Minimal Operational Definition for This Thesis

For the purposes of this thesis, we adopt the following minimal, action-oriented definition:

Moral cognition is the evaluative process through which agents detect

> *normatively salient features of a situation, generate judgments regarding permissible or obligatory actions, and select behaviour accordingly.*

This definition is intentionally modest. It avoids substantive normative commitments while capturing the components required for empirical investigation: *evaluation*, *judgment*, and *action*. It aligns with contemporary accounts of moral psychology that treat morality as grounded in both affective and cognitive mechanisms [155, 82, 75]. It also coheres with the theoretical machinery of this thesis, including evaluative topology, levels of abstraction, and the notion of semiotic perturbation. *Moral cognition thus functions as a mapping from situational cues to action policies, modulated by trait-level and affective structures* [65, 60].

4.3 Judgments: Factual and Normative

A central distinction for understanding moral cognition is the difference between *factual* and *normative* judgments. Although both concern evaluations of situations, they operate at different logical and conceptual levels. Factual judgments describe states of affairs: they answer questions about what *is* the case. Normative judgments concern what *ought* to be done, what is *permissible*, *required*, or *forbidden*. The distinction is classical in philosophy, yet it remains frequently blurred in computational and psychological treatments of morality [119, 151].

Factual judgments derive their correctness conditions from empirical features of the world. Their truth depends on evidence, observation, or inference. Normative judgments, by contrast, embed claims about reasons for action and the standards that govern deliberation. They express commitments that are action-guiding and prescriptive in force, even when articulated implicitly [42, 158]. What follows from this distinction is more than a semantic bifurcation: it marks a functional divide in the cognitive architecture that underwrites evaluative thought. A judgment about what *is* the case engages classificatory and predictive mechanisms; a judgment about what *ought* to be the case recruits additional systems responsible for assigning motivational weight, integrating affective cues, and generating the directional force that links evaluation to action.

Key distinction: Factual = descriptive; Normative = action-guiding.

Moral cognition refers to the ensemble of perceptual, affective, and inferential processes through which agents register morally relevant features of a situation and transform them into evaluative representations [58, 88, **?**]. It encompasses both explicit moral judgment and upstream mechanisms that detect salience, encode social meaning, and initiate the transition from evaluative appraisal to behaviour [57, 53]. Introducing this construct at this stage is essential, because it clarifies that the descriptive–normative distinction is mirrored in the cognitive architecture that processes them: factual information is registered by systems specialised for prediction and classification, whereas normative evaluation recruits additional mechanisms that assign motivational force and action-guiding significance.[1]

Moral cognition: Perception → appraisal → action-guidance.

In moral cognition, the distinction is not merely verbal but functional. Psychological models indicate that factual information serves as input to evaluative ap-

[1] In moral psychology, this distinction is often operationalised by contrasting cognitive processes supporting representational accuracy with those supporting valuation and action selection. See [75, 143, 167].

praisal [168, 169, 170, 162], but normative judgment involves the additional step of mapping descriptive cues onto action-guiding evaluations [171, 172, 173, 168]. Treating normative judgments as a special case of factual ones therefore collapses essential differences in their psychological and functional architecture. For empirical research in moral psychology—and particularly for any paradigm seeking to measure moral behaviour—the distinction ensures that observable responses are not misinterpreted as direct indicators of moral endorsement or norm acceptance.

Methodolog note: Behaviour ≠ endorsement unless interpretive architecture is specified.

This distinction between descriptive input and normative evaluation sets the stage for a further refinement. Once we recognise that moral cognition incorporates specialised mechanisms for assigning salience, generating evaluative force, and transforming appraisal into behaviour, it becomes clear that moral judgments themselves cannot be exhaustively characterised as simple outputs of belief or emotion. They arise from the coordinated operation of multiple cognitive systems—perceptual, affective, inferential, and motivational—whose interaction determines not merely *what* is judged, but *how* and *why* it guides action. In other words, the transition from factual uptake to normative appraisal presupposes an internal architecture of judgment: a structured evaluative act with identifiable components that jointly confer its distinctive normative authority. It is to this internal architecture that we now turn.

4.4 The Structure of Moral Judgments

Moral judgments are not mere expressions of preference or affective reaction. They exhibit a characteristic structure combining evaluative content, justificatory grounding, and action-guiding force [174, 175, 176, 177, 178]. A moral judgment typically involves at least three components:

1. **Salience detection**: recognition that a situation involves normatively relevant features (harm, fairness, honesty, obligation, care). This process draws upon perceptual, affective, and social-cognitive systems [75, 82].
2. **Evaluative appraisal**: an assessment of those features in light of internalised norms, dispositions, or reasons. This appraisal may be intuitive or reflective, emotionally charged or deliberative, depending on the individual and context [153, 60].
3. **Practical commitment**: a transition from evaluation to action guidance, where the judgment functions as a reason for or against performing a particular behaviour [157, 158].

These components jointly distinguish moral judgments from other evaluative acts such as aesthetic preferences or strategic choices. They also underwrite the thesis's operational understanding of moral cognition as an *evaluative mapping* from cues to action.

Importantly, this structure accommodates both intuitive and deliberative models. Intuitive processes may dominate in everyday moral encounters; nonetheless, these judgments retain justificatory structure, even when reasons are not explicitly articulated [88, 172, 173, 162]. Conversely, deliberative processes involve explicit reasoning, *counterfactual consideration*, and appeal to principles or char-

acter traits [35]. This duality will be further elaborated in the discussion of psychological and neuroscientific foundations that follows.

This distinction between intuitive and deliberative processes is not merely taxonomical; it marks the beginning of a deeper inquiry into the cognitive architecture that makes moral judgment possible. To understand why certain stimuli reliably elicit prosocial behaviour while others disrupt or attenuate it, we must examine the underlying mechanisms through which moral salience is perceived, represented, and acted upon. The transition from intuition to deliberation is mediated by identifiable affective, perceptual, and executive systems, each contributing distinct computational roles within the broader moral economy. As the following section illustrates, contemporary psychological and neuroscientific research converges on a model of moral cognition as a distributed and dynamically interactive network. This framework not only clarifies how humans ordinarily navigate morally charged environments, but also establishes the theoretical scaffolding required to interpret how such processes may be perturbed—subtly yet measurably—by the presence of agents whose social and ontological status is ambiguous, such as humanoid robots. In this sense, the empirical foundations surveyed below serve as the conceptual substrate upon which our experimental analysis later builds.

4.4.1 Psychological and Neuroscientific Foundations of Moral Decision-Making

A substantial body of work in cognitive neuroscience demonstrates that moral decision-making is not the product of a single "moral centre" but emerges from coordinated activity across distributed affective, social-cognitive, and executive networks. These systems jointly determine how agents detect morally salient cues, generate evaluative appraisals, and select action policies. The architecture is, in this sense, inherently *practical*: the neural substrates implicated in moral judgment are deeply intertwined with those responsible for value computation, behavioural control, and action selection.[2] Rather than isolating "moral reasoning" as a *sui generis* faculty, contemporary research positions it within a larger computational system whose governing question is not "What is right?" but "What should I do given this situation?" [110, 75, 172].

Affective and Value-Based Systems. Among the most extensively studied structures contributing to moral evaluation are the ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC). These regions compute affective and motivational value, integrating emotional information with anticipated outcomes. Lesion studies demonstrate that damage to the vmPFC disrupts the ability to factor emotional and social consequences into decision-making, often resulting in choices that appear normatively inappropriate or insensitive to harm [110]. Functional imaging studies show robust vmPFC activation during tasks involving interpersonal harm, care, and empathic concern [75]. These observations suggest

[2]This stands in contrast to folk-psychological depictions of moral judgment as a purely contemplative process concerned with identifying moral facts. Neuroscientific evidence overwhelmingly supports action-guidance as the primary functional orientation of moral cognition.

that moral judgments rely on mechanisms that encode the valenced quality of behavioural options and link them to affectively grounded somatic markers.

The amygdala and anterior insula further contribute to the rapid detection of morally salient information [55, 52, 54] . The amygdala is sensitive to threat, intentional aggression, and aversive outcomes, providing early affective tagging [179, 180] that biases attention and behavioural readiness. The anterior insula responds to disgust, norm violations, and aversive interoceptive states [181, 182, 183]. Together, these regions enable rapid, pre-reflective processing of emotionally charged cues, thereby initiating downstream evaluative computation. Electrophysiological evidence indicates that these affective signals can precede conscious deliberation [184, 185], suggesting that emotional valence functions as an early gatekeeper in moral cognition.

Social-Cognitive and Interpretive Systems. Moral judgments frequently hinge on the mental states of agents: their beliefs, intentions, and reasons for action [186, 187, 188]. The temporo-parietal junction (TPJ), medial prefrontal cortex (mPFC) [189, 190], and posterior superior temporal sulcus (pSTS) constitute a network specialised for theory-of-mind and mental-state attribution [191, 192]. TPJ activation is reliably observed in tasks requiring participants to distinguish between intentional and accidental harms, to attribute blame or forgiveness, or to infer whether an agent acted under ignorance or malice. This sensitivity to mental-state information demonstrates that moral cognition tracks reasons and intentions, not merely outcomes [193, 57].

The anterior cingulate cortex (ACC) plays an integrative role in moral cognition by monitoring conflict between competing evaluative signals [194, 195]. In classic moral dilemmas—such as those involving trade-offs between harm minimisation and fairness constraints—the ACC shows increased activation during conflict detection and the recruitment of cognitive control [143, 196]. This suggests that the ACC contributes to arbitrating between intuitive emotional responses and more deliberative evaluations, particularly in situations where values compete or intentions are ambivalent [197, 198].

Executive and Action-Guidance Systems. The dorsolateral prefrontal cortex (dlPFC) supports controlled cognitive operations, including the inhibition of prepotent affective responses, the representation of rules, and the evaluation of abstract or long-term consequences [199, 200]. Disruption of dlPFC activity via transcranial magnetic stimulation has been shown to alter participants' willingness to endorse harmful actions in instrumental contexts, indicating that this region contributes to regulating intuitive aversions when normative or goal-directed reasoning requires overriding them [201, 202]. Rather than functioning as a classical "rational override," the dlPFC appears to contribute to integrating affective, deontic, and goal-directed considerations into coherent action policies [167, **?**].

Importantly, the dlPFC does not operate in isolation. Its interactions with vmPFC, ACC, and parietal regions indicate that executive control is embedded within a broader network that also encodes affective and interpretive information [203, 204, 205]. These distributed processes jointly shape the computation

of moral decisions as behavioural commitments rather than as purely abstract evaluations [206, 207].

Functional Integration and Practical Orientation. Across these subsystems, a coherent picture emerges: moral cognition is not a contest between "emotion" and "reason" but a dynamic interplay among affective valuation, social interpretation, and executive control [88, 208, 209]. This architecture is fundamentally action-oriented. vmPFC and OFC compute the affective value of potential actions [210, 211]; TPJ and mPFC provide intention-sensitive interpretations of agents' behaviours [190, 193]; the ACC detects conflicts between competing behavioural tendencies [194, 195]; and the dlPFC regulates whether intuitive biases should be suppressed, enacted, or weighed against normative constraints [199, 201]. Even primary affective structures such as the amygdala and insula contribute to shaping behavioural readiness by generating rapid somatic markers and prioritising morally relevant features of the environment [180, 183].

Lesion studies, electrophysiological findings, and neuroimaging results converge on the conclusion that moral judgment is primarily a mechanism for generating and constraining action under conditions of social meaning. From this perspective, moral cognition is best understood as a form of evaluative control: a mapping from cue detection to practical commitment [172, 212]. This view aligns with philosophical accounts emphasising the action-guiding nature of moral evaluation [157, 158], while grounding such accounts in empirical evidence about the neural architecture of agency, valuation, and control.

From Moral Architecture to Perturbation by Synthetic Agents. This *distributed, action-oriented architecture* provides the conceptual and empirical framework for understanding the experimental work developed later in this thesis.

If moral judgment emerges from systems designed to transform perceptual, affective, and interpretive cues into behavioural output, then *alterations to the social or perceptual environment can shift the evaluative computations that guide action.* This point is not merely theoretical: later chapters develop its empirical instantiation by demonstrating how perturbations to the social field modulate the transition from moral salience to prosocial behaviour (see Hypothesis 3 in Chapter 7).

A humanoid robot constitutes a particularly revealing form of perturbation: it is perceptually social (in virtue of its humanoid morphology) yet ontologically indeterminate (neither fully agentic nor behaviourally inert). Such indeterminacy can alter attentional allocation, dampen affective resonance, and introduce uncertainty into mental-state attribution. In doing so, it may shift the weighting, timing, or accessibility of evaluative signals, thereby modulating the likelihood that moral appraisal culminates in prosocial action.

Understanding this architecture is therefore essential for interpreting the empirical results that follow. Our experiment does not measure abstract judgments but the practical enactment of moral cognition within a context made ambiguous by the presence of a synthetic observer. The neuroscientific foundations surveyed here thus provide the theoretical scaffolding for explaining how robotic presence

can attenuate prosocial action in subtle, yet systematically measurable ways.

What follows, however, requires a final conceptual step. If moral cognition is an architecture for transforming evaluative information into action, then *any alteration to the informational field is at least in principle a moral intervention*. The presence of a synthetic agent—especially one exhibiting humanlike form yet lacking a clear place within our evolved social ontology—constitutes precisely such an intervention. It does not supply new moral content; rather, it reconfigures the *conditions under which content becomes behaviourally operative*. In this sense, the moral landscape is not only defined by principles or dispositions but by the topology of the environment in which they are enacted.

This insight has two important implications that structure the remainder of the thesis. First, it shifts the explanatory burden from conscious deliberation to the *situated dynamics of evaluative processing*. The experiment that follows examines not what participants claim to value, but how their moral cognition actually functions when confronted with an entity whose status is neither fully social nor fully inert. Second, it reframes the normative question: the significance of artificial agents lies not merely in what they do, but in how their mere presence *reconfigures the normative affordances* of a shared environment. This reframing will prove central when, in later chapters, we consider the limitations of existing Machine Ethics frameworks and the conceptual tension between engineered normativity and human moral practice.

In this way, the Moral Primer sets the stage for two convergent lines of inquiry. The empirical chapters will show how minimal synthetic presence can modulate the behavioural expression of moral cognition. The normative chapters will argue that such modulation exposes a broader oversight in contemporary ethical theory for artificial systems: namely, the assumption that moral agency can be understood independently of the environments that scaffold, shape, and sometimes distort human evaluative capacities.

Taken together, these threads suggest a view of artificial agents not as moral subjects, nor merely as tools, but as *operators on moral space*: entities capable of bending, refracting, or diluting the pathways through which moral meaning becomes action. The full implications of this claim will emerge only when the empirical and philosophical analyses are placed in dialogue. For now, it suffices to note that understanding how humans make moral decisions under conditions of social and ontological ambiguity is not merely preparatory background—it is the conceptual linchpin for everything that follows.

These conceptual foundations also illuminate two methodological commitments that guide the remainder of the thesis: the *Level of Abstraction* at which moral cognition is analysed, and the *topological structure* of the evaluative processes under perturbation. In Floridi's sense, an LoA fixes the informational parameters relevant to explanation; it determines which distinctions matter and which are bracketed for the sake of epistemic tractability. Here our chosen LoA does not concern the metaphysics of moral agency, nor the normative justification of principles, but the *functional transformation* by which perceptual and affective cues become action-guiding evaluations. It is at this LoA that robotic presence

can be treated not as a moral agent but as a *modulator of the evaluative field*.[3]

Once this LoA is fixed, moral cognition can be understood topologically: as a system that maps inputs to behavioural outputs through a structured configuration of salience, attention, affective resonance, and interpretive inference. Altering the structure of the environment—as occurs with the introduction of a synthetic observer—can therefore be modelled as a deformation of the evaluative landscape. The experiment developed later in this thesis investigates precisely such a deformation: not a change in moral principles, nor a shift in explicit reasoning, but a modification of the *shape* of the cognitive–affective space through which moral meaning travels on its way to action.

This topological perspective additionally clarifies why synthetic agents matter ethically even when they perform no overt behaviour [213, 214]. At our operative LoA, the morally relevant property of a robot is not its autonomy or its adherence to ethical rules, but its capacity to warp the attentional and affective gradients that structure human moral appraisal [215, 216]. A robot may therefore function as a semantic attractor or normative deflector, subtly redistributing the vectors through which moral salience exerts its behavioural pull [217, 218]. Later empirical chapters provide evidence for such redistributions; later normative chapters examine how these redistributions challenge the assumptions of Machine Ethics, which typically locates moral significance in the agent rather than in the environmental perturbation it induces [152, 219].

Seen through this joint lens of LoA and moral topology, the empirical question posed by the experiment acquires its full significance: not whether a robot is moral, nor whether it communicates norms, but whether its presence reshapes the evaluative field in which human agents convert moral perception into prosocial behaviour. The answer to that question, and its implications for both moral psychology and the ethics of artificial agents, unfolds in the chapters that follow.

To make this intuition formally explicit—without committing the remainder of the thesis to a mathematical framework—we may describe the evaluative system as a mapping

$$f : \Sigma \longrightarrow \Delta,$$

where Σ denotes the space of perceptual and affective cues, and Δ the space of action-guiding evaluative states. Within this framework, the presence of a synthetic agent $\mathscr{R}$ can be modelled as a perturbation operator

$$\mathcal{P}_{\mathscr{R}} : \Sigma \to \Sigma',$$

inducing a deformation of the salience landscape such that the composed transformation

$$f_{\mathscr{R}} = f \circ \mathcal{P}_{\mathscr{R}}$$

yields a different evaluative gradient. In topological terms, $\mathcal{P}_{\mathscr{R}}$ alters the curvature of the moral field, reshaping the trajectories along which attention, affect, and social meaning propagate toward behavioural output. This formalism is

[3]For discussion of the methodological role of Level of Abstraction in analysing informational systems, see Floridi (2010, 2011, 2013).

offered only as a conceptual anchor: the thesis does not employ topological machinery beyond this illustrative role, but the notion of perturbation provides a precise vocabulary for interpreting the empirical findings that follow.

Philosophical Synthesis: Rational, Virtuous, and Intuitive Moral Minds. This perspective also illuminates a deeper philosophical point. Across the history of moral thought, the relation between perception, evaluation, and action has been theorised along three dominant lines. A Kantian picture holds that moral judgment is grounded in rational principles and universalizable maxims; here, perturbation would matter only insofar as it obstructs rational access to duty. An Aristotelian picture instead treats moral action as the expression of a cultivated character, where perception and moral salience form a unified excellence of practical reasoning (*phronesis*). A Humean picture goes further, locating the origin of moral judgment in sentiment, affect, and intuitive appraisal. The cognitive–affective architecture surveyed above aligns most closely with this latter tradition: it suggests that moral judgment is grounded not in detached reasoning but in an integrated evaluative sensitivity to the social world. Thus, when the structure of that world is altered—when its cues are reframed, occluded, or semantically displaced—the moral response shifts accordingly.

Concluding Perspective: Why This Matters for the Thesis. Placed within this triangulation of mathematics, psychology, and philosophy, the experimental work to come acquires a precise significance. The thesis does not ask whether robots are moral agents, nor whether they instantiate ethical principles. Rather, it investigates how synthetic presence restructures the evaluative environment in which *human* moral cognition unfolds. By treating robotic co-presence as a perturbation of the moral field, the experiment operationalises a long-standing philosophical question in empirical form: *to what extent is moral action sensitive to the topology of the situation rather than to the content of explicit reasoning?* The answer developed in later chapters shows that even minimal, silent, behaviourally neutral artificial agents can shift the gradients through which moral salience becomes prosocial behaviour.

This insight anchors both the methodological rationale of the experiment and the broader argumentation of the thesis. In the Introduction, it frames the motivation for studying synthetic social influence; in the General Conclusion, it becomes the conceptual pivot for evaluating the ethical implications of robotic presence in human environments. The moral mind, understood through the lenses of abstraction and topology, is not merely a processor of principles but a dynamic, situationally sensitive system—one whose pathways can be subtly bent by the artificial others increasingly woven into our social world.

4.5 Dual-Process Architectures in Moral Cognition

The distributed moral architecture described in the previous section naturally motivates a family of theories known as *dual-process models*.

These models posit that moral judgment arises from the interaction of rapid, affectively grounded appraisals with slower, more controlled processes of deliberation and cognitive regulation [220, 88, 57, 221]. Importantly, dual-process models no longer depict these systems as antagonistic. Contemporary formulations emphasise their continual integration, consistent with the topological and action-oriented framework established earlier [143, 208, 209].

This emphasis on dual-process architectures is not merely a matter of theoretical convenience; it reflects a deeper convergence across multiple research programmes concerned with the computational structure of social and moral behaviour [221, 220]. In Social Signal Processing, Vinciarelli and colleagues argue that human social interaction is supported by parallel channels of affective and inferential processing, where rapid, embodied cues—such as posture, gaze, and micro-expressions—operate alongside higher-level reasoning about intentions and norms [134, 132]. Affective Computing similarly treats emotion as a multi-layered control system in which fast appraisal mechanisms shape behavioural orientation long before explicit reasoning is engaged [133, 222].

By contrast, much of Machine Ethics has historically leaned toward monolithic, deliberative models that treat moral judgment as if it were exclusively rule-based, thereby neglecting the affective and perceptual grounding that empirical evidence shows to be indispensable for real human moral cognition [223, 14]. The dual-process framework adopted here therefore marks a principled departure from those purely deliberative approaches: it aligns the theoretical lens of this thesis with the empirical reality that moral decision-making arises from the interaction of intuitive, affect-laden appraisals and slower, controlled processes of evaluation and regulation [88, 57].

This perspective also anticipates the logic of the experimental design developed later in this thesis: if moral judgment arises from the continual integration of fast affective cues with controlled interpretive processes, then perturbing either stream—for example, by introducing a humanoid robot that is perceptually salient yet ontologically indeterminate—should measurably alter the resulting behavioural output [216, 147]. Crucially, such a perturbation is not merely of theoretical interest; it provides empirical traction on a broader question that has become increasingly central in Affective Computing and Social Signal Processing, namely: *how should artificial systems register, represent, and respond to the moral significance of human behaviour in ways that reflect the dynamics of human moral cognition rather than static rule-based prescriptions?* Discoveries that reveal how synthetic agents modulate human moral action therefore speak directly to the design of computational architectures capable of participating in, rather than merely implementing, moral practices. A human-centric, discovery-oriented approach to moral AI must be guided by the empirical structure of moral cognition itself—its affective grounding, its interpretive flexibility, and its inherently practical orientation—rather than by fixed normative templates that fail to scale across social contexts. Dual-process theory thus provides both the conceptual and methodological scaffolding for understanding how robotic presence reshapes the balance between intuitive and deliberative pathways, and for developing affective-computational models that align with the lived moral ecology in which human agents actually operate [90, 221].

Intuitive and Affective Processes. The intuitive stream comprises fast, automatic, and affectively laden evaluations generated by perceptual and subcortical mechanisms. It inherits its computational character from the structures surveyed earlier: the amygdala and anterior insula for rapid affective tagging and aversive appraisal [180, 183]; the vmPFC for integrating somatic markers and affective valuations into action-anchored representations [110]; and the TPJ–mPFC network for immediate interpretation of agents' intentions [193, 190]. These processes operate at low cognitive cost and generate what can be described, within the topological framework introduced previously, as steep, rapidly forming attractor basins in the evaluative landscape. Intuitive appraisals therefore play a decisive role in shaping the initial direction of behavioural tendencies, often determining which features of a situation are encoded as morally salient before deliberation becomes possible. This mechanism is well documented in Social Signal Processing and Affective Computing, where micro-expressions, gaze cues, posture shifts, and affective markers orient behavioural readiness even in the absence of explicit reasoning [134, 222]. Intuition is thus not a primitive substitute for deliberation but the behaviourally grounded substrate on which moral thought is built.

Controlled and Deliberative Processes. The controlled stream is subserved by dorsolateral prefrontal cortex (dlPFC), anterior cingulate cortex (ACC), and lateral parietal networks implicated in rule representation, inhibition, causal reasoning, and long-horizon evaluation [199, 200, 195]. Controlled processes are engaged when intuitive appraisals conflict with one another or with internalised commitments, when the moral relevance of a situation requires justification, or when environmental ambiguity demands a more structured interpretation. In the topological model, controlled processes reshape intuitive attractors by flattening some gradients and amplifying others, thereby altering the curvature of the evaluative field in ways that enable stable action selection. This modulation is not well captured by the outdated metaphor of a "rational override." Rather, the controlled system integrates affective, social-cognitive, and normative information into coherent action policies, consistent with philosophical accounts of moral judgment as a species of practical reasoning [157, 158]. Its computational significance extends to Affective Computing: controlled processes supply the representational flexibility required for artificial systems to track context, resolve ambiguity, and form morally relevant action-guiding policies rather than static rule-matching outputs [133].

Dynamic Integration. Dual-process models thus support the view that moral cognition is an interactive, topologically structured, and dynamically integrated system [221, 57]. Intuitive mechanisms generate the initial evaluative landscape—anchored by affect, attention, and perceptual salience—while controlled mechanisms adjust, stabilise, or reinterpret that landscape in light of commitments, norms, or long-term goals. The system's behaviour is therefore neither fully bottom-up nor fully top-down but emerges from the continual exchange between affective gradients and regulatory constraints. This integrated architecture supplies the mechanistic substrate for the perturbation phenomena examined later in the thesis: when the environment changes, it alters the intuitive gradients that form the initial evaluative field, thereby reshaping the downstream burden

on controlled processes [224]. Crucially, a synthetic agent need not supply explicit reasons to influence moral behaviour; by virtue of its perceptual presence and ambiguous ontological status, it can reconfigure the evaluative field in which reasons become behaviourally operative [213, 216].

Against this theoretical background, the empirical question addressed later in the thesis becomes more sharply framed.

If moral cognition arises from the dynamic integration of fast affective appraisals with slower controlled processes, then perturbations to the perceptual–social environment that alter the initial affective gradients or the subsequent regulatory burden should yield measurable shifts in moral behaviour.

A humanoid robot constitutes precisely such a perturbation: its perceptual salience, humanomorphic form, and ambiguous ontological status are known to modulate attention, social appraisal, and mind attribution in ways that alter both intuitive and deliberative pathways [216, 213, 147]. These influences operate at the level of the evaluative field itself—reshaping which cues are encoded as morally salient, how affective resonance unfolds, and when controlled processes are recruited to stabilise or reinterpret the situation. Evidence from developmental, affective, and social-cognitive neuroscience suggests that such shifts in salience and ambiguity can reconfigure the timing and accessibility of both intuitive and controlled moral processes [225, 222, 132]. The experiment developed in Chapter 7 therefore provides an empirical test of this mechanistic claim:

whether the presence of a synthetic observer systematically deforms the evaluative landscape from which prosocial action emerges.

4.6 The Social Intuitionist Model

While dual-process architectures describe the mechanistic integration of affective and deliberative pathways, Haidt's *Social Intuitionist Model* (SIM) provides a complementary account of the *social ecology* within which moral judgments are produced and negotiated [88, 155]. SIM stands firmly in the Humean tradition: moral judgment arises first from rapid intuitive appraisals, with explicit reasoning operating primarily as a communicative, justificatory, or reputation-management mechanism. For the purposes of this thesis, its significance is twofold.

First, it anchors moral cognition in perceptual, affective, and socially distributed processes rather than in solitary rational deliberation. Second, it predicts that even minimal, ambiguous, or merely perceptual forms of social presence can reshape the intuitive component of moral judgment—precisely the mechanism probed by the experiment that follows.

Primacy of Intuition. According to SIM, intuitive appraisals constitute the generative core of moral judgment. These appraisals arise rapidly—on the order of hundreds of milliseconds—through affective and perceptual pathways that register norm violations, suffering, fairness cues, or socially meaningful stimuli long before conscious deliberation is possible [185, 184]. Neurocognitive evidence shows that harm detection, norm sensitivity, and interpersonal appraisal are instantiated in circuits associated with affective tagging (amygdala, anterior in-

sula), mental-state attribution (TPJ, mPFC), and embodied simulation. Within the topological framework developed earlier, SIM corresponds to the idea that intuitive processes generate the *initial curvature* of the evaluative landscape: they create steep affective gradients that orient subsequent interpretation and action.

This primacy is not merely temporal but structural: intuitive appraisals determine which environmental features are encoded as morally salient. In Social Signal Processing and Affective Computing, analogous mechanisms are observed in the rapid extraction of gaze, posture, and affective micro-signals that guide interpersonal coordination [134, 222]. SIM therefore provides the social–cognitive analogue of the perceptual–affective machinery surveyed earlier: intuitions are not post hoc artefacts but the first-order drivers of moral cognition.

Reason as Interpersonal. SIM also reconceptualises reason not as the architect of moral judgment but as a socially situated *modulatory process*. Deliberation is typically invoked after intuitive appraisals have already fixed the valence and direction of judgment. It functions primarily to justify existing intuitions, align with interlocutors, negotiate reputation, or repair social discord [88]. Philosophically, this positions moral reasoning closer to an interpersonal practice—akin to Strawsonian norm negotiation or Scanlonian contractual justification—than to a solitary search for objective moral facts.

At the Level of Abstraction operative in this thesis, such reasoning does not generate the moral evaluation; it operates on a pre-structured evaluative field shaped by intuition, affect, and social meaning. SIM therefore aligns with the broader theoretical framework established earlier: moral cognition is fundamentally action-oriented, socially embedded, and sensitive to perturbations in the perceptual–social environment.

Relevance to Synthetic Presence. The Social Intuitionist Model acquires particular methodological force in relation to the experiment developed later in the thesis. If moral intuition is tuned to social presence—especially to the mere perception of another mind, agent, or observer—then introducing a humanoid robot with ambiguous ontological status constitutes a direct perturbation of the intuitive machinery itself. Empirical work shows that ambiguous agents modulate attentional allocation, emotional resonance, and intuitive moral appraisal [216, **?**, 147]. Within SIM's framework, such modulation occurs *prior* to deliberation: the robot reshapes the intuitive gradients that structure the evaluative field, thereby altering the likelihood that moral salience will translate into prosocial action.

Thus SIM not only complements the dual-process architecture but also provides a socially grounded explanatory lens for the modulation effects measured in the experiment. It predicts precisely the kind of subtle, pre-reflective, yet behaviourally measurable displacement that emerges when a synthetic agent perturbs the affective–social substrate from which moral judgments arise.

Synthetic Presence and Social Perturbation. SIM is particularly powerful when considering *minimal sociality*. A humanoid robot constitutes a perceptually social yet ontologically indeterminate entity. Its presence can influence intuitive

appraisal by shifting attention, altering affective resonance, or modulating perceived social oversight. These shifts occur at the intuitive stage of moral processing, thereby modifying the evaluative gradients that shape action. SIM thus provides a conceptual bridge between the empirical findings of the experiment and the theoretical claim that synthetic agents restructure evaluative topology even in the absence of explicit communication or normative instruction.

4.7 Prosocial Behaviour as Moral Action

The conclusion reached in the previous section—that minimal or ambiguous social presence can reshape intuitive appraisal—immediately motivates a shift from internal judgment to observable action. At the Level of Abstraction adopted throughout this thesis, moral cognition is defined by its action-guiding role: *it is a system whose explananda are behavioural commitments rather than verbal reports.*

If intuitive appraisal is the first locus at which synthetic presence can deform the evaluative landscape, then the empirically accessible signature of such deformation must be sought not in explicit justification but in the practical enactment of moral cognition. This makes prosocial behaviour—cooperation, helping, and in particular, **costly resource donation** [226, 227, 228, 3, 229, 230, 130, 44, 231]—the principled site at which perturbations to the evaluative topology become experimentally tractable[232, 233, 216, 147, 234, 65, 207].

Prosocial behaviour is not merely an altruistic variant of social action; it is one of the most reliable cross-disciplinary indicators of moral engagement. Across behavioural economics, developmental studies, evolutionary game theory, and empirical moral psychology, patterns of helping, sharing, and charitable giving consistently track agents' sensitivity to fairness, harm, reciprocity, compassion, and need [226, 228, 227, 3, 229]. Philosophical accounts converge on the same point. Whether in Scanlon's framework of interpersonal justifiability [44], Darwall's second-personal standpoint [231], or constitutivist and planning-theoretic models of agency [117, 158, 235, 236], prosocial actions manifest a normatively loaded form of practical identity:

they reveal how an agent construes her reasons, acknowledges the claims of others, and binds herself to outcomes she takes to be morally required. In this sense, prosocial behaviour is not merely expressive but commitment-realising.

It shows how evaluative appraisals become operative in action, rather than hypothetically endorsed in speech.

This distinction between avowed moral attitudes and their embodiment in action has become central to contemporary accounts of moral psychology and agency. Experimental work demonstrates that explicit judgments are weak predictors of real-world helping, whereas behavioural tasks expose the operative structure of moral evaluation [212, 207]. Philosophical analyses of responsibility, reactive attitudes, and self-governance likewise locate normativity in enacted agency rather than assent [237, **?**, 238, 239]. Against this background, the methodological choice of using charitable donation as the dependent variable in the experiment is not

an operational convenience but a principled commitment to analysing moral cognition in its practical register.

Prosocial action emerges from a structured sequence already outlined in earlier sections: detection of morally salient cues; intuitive appraisal of their affective and social significance; controlled modulation when competing commitments must be adjudicated; and finally, behavioural execution. Each stage is exquisitely sensitive to the social field. Whether an agent feels watched, whether the observer is human or synthetic, whether the presence is affectively warm, neutral, or indeterminate—all of these modulate the trajectories through the evaluative topology. Neuroimaging studies confirm that prosocial choice draws on the same integrated circuitry that underwrites harm aversion, empathic concern, valuation, and norm enforcement [75, 82]. Within the topological framework developed earlier, prosocial behaviour marks the *terminal attractor* of a moral trajectory: when salience is strong and unimpeded, the system converges on a stable basin corresponding to prosocial action.

It is precisely this structure that renders prosocial donation an ideal test variable for the experimental paradigm developed in this thesis. If the presence of a humanoid robot—perceptually social yet ontologically ambiguous—alters attentional focus, affective resonance, or the perceived structure of social oversight, then those perturbations need not appear in verbal justifications; they will appear in the *behavioural expression* of moral cognition. The experiment therefore does not treat donation as a proxy for morality in a naïve sense, but as a theoretically grounded behavioural readout of the evaluative topology: a measurable manifestation of how moral salience is transformed into action under conditions of synthetic perturbation. In this view, detecting attenuation in prosocial behaviour is detecting deformation in the underlying evaluative field.

Why Prosocial Behaviour Serves as a Proxy for Moral Action. Within the theoretical framework developed throughout this chapter, prosocial behaviour is not an auxiliary behavioural measure but the natural terminus of moral cognition understood as a practical, action-guiding system. Across divergent philosophical traditions, there is agreement on this point: for Aristotle, the telos of ethical reflection is *praxis* [240, 241, 242] ; for Kant, moral judgment manifests in the capacity to act from obligation [243, 178, 244] ; for Hume, moral distinctions acquire motivational force only through sentiment [245, 246, 247] . Despite their incompatibilities, all three converge on the thesis that the mark of moral cognition is its capacity to reorganise an agent's field of reasons so as to issue in action.

In computational terms, a prosocial act such as monetary donation reveals that the evaluative field has reached a locally stable configuration: moral salience has been detected, weighted, and rendered behaviourally operative despite competing self-regarding incentives. Donation therefore provides access to what this thesis calls the *operational core* of moral cognition: the point at which evaluative topology is strong enough to yield observable practical commitment. It is, in this sense, not a surrogate for morality but its empirical footprint. What the agent does with her own resources in a morally charged situation is the clearest behavioural index of how moral meaning has been processed and transformed by

the cognitive–affective machinery surveyed in earlier sections [212, 207].

Relevance for Synthetic Perturbation. Because prosocial behaviour is the final expression of the evaluative trajectory, it is also the level at which perturbations to the moral field become measurable. Synthetic presence—especially when perceptually social yet ontologically indeterminate—can alter the evaluative topology long before explicit reasoning is recruited. Changes in attentional allocation, reductions in affective resonance, or increased uncertainty in mental-state attribution operate at the intuitive tier and propagate through the system, subtly reshaping gradient structures and attractor dynamics. Such perturbations rarely manifest in explicit moral self-reports, which are coarse, post hoc, and socially filtered; instead, they emerge in the *behavioural expression* of moral cognition.

The experimental design developed in Chapter 7 leverages precisely this fact. By embedding a morally salient stimulus (the charity prime) within an environment modulated by a silent humanoid robot, the paradigm tests how ontological ambiguity refracts the integration of intuitive and deliberative processes. A reduction in donation is therefore not interpreted as the failure of moral principle, nor as evidence of diminished norm endorsement, but as a deformation in the evaluative pathway linking moral perception to prosocial behaviour. In topological terms, the presence of the robot shifts the curvature of the moral field, altering the system's convergence tendencies and lowering the probability of reaching the prosocial attractor.

Conceptual Synthesis: Why Prosocial Donation Measures Moral Perturbation

Prosocial donation is the behavioural endpoint of the evaluative architecture that transforms moral salience into action. Because moral cognition is inherently practical, perturbations to its perceptual, affective, or interpretive components manifest most reliably in the structure of behavioural output. A humanoid robot—perceptually social yet ontologically ambiguous—modulates this architecture not by issuing commands but by reshaping the evaluative field in which moral meaning becomes behaviourally operative. Measuring donations under synthetic perturbation therefore provides a principled, theoretically grounded method for detecting how artificial presence deforms the transition from moral appraisal to prosocial commitment.

Concluding Perspective: Why This Matters for the Thesis. The argument developed across the preceding sections now converges: dual-process architectures reveal that moral cognition arises from the continuous integration of intuitive, affect-laden evaluations with controlled interpretive processes; the Social Intuitionist Model emphasises the primacy of intuitive, socially responsive appraisal; dynamic integration shows how perturbations to early evaluative gradients propagate through the system; and the analysis of prosocial behaviour establishes that the appropriate level of empirical access is behavioural rather than declarative.

Placed against this background, the experimental work to follow acquires its precise theoretical meaning. The thesis does not ask whether robots possess moral agency, nor whether they instantiate ethical principles, nor whether they trigger reputational reasoning in any explicit sense. Instead, it investigates how the presence of a synthetic, perceptually social, ontologically ambiguous agent reshapes the evaluative topology through which *human* moral cognition unfolds. The robot is treated not as a quasi-person but as a perturbation operator on the moral field: a source of structural deformation capable of altering the gradients through which moral salience is transformed into prosocial action.

In this light, the experimental question becomes a refined philosophical one:

> *To what extent is moral action sensitive to the situational topology in which it is embedded, rather than to the content of explicit reasoning or principle endorsement?*

The answer, as subsequent chapters demonstrate, is that even minimal, silent, behaviourally neutral artificial agents can tilt the evaluative landscape, shifting the balance between intuitive and deliberative processes and decreasing the probability that moral salience converges on prosocial action. This finding is not anecdotal but structurally illuminating: it shows that the moral mind, when understood through the lenses of abstraction and topology, is a dynamic, context-responsive system whose behavioural outputs depend on the shape of the environment as much as on internal principles.

This insight anchors the methodological rationale of the experiment, motivates the normative analysis developed in the Ethical Cognition chapter, and frames the broader implications discussed in the General Conclusion. It also sets the stage for a technomoral claim that animates the thesis as a whole: as artificial agents become woven into ordinary human environments, the topology of moral life itself may be subtly, pervasively reshaped—not through explicit persuasion, but through shifts in the evaluative fields within which moral cognition takes place.

5. TOOLS

6. Psychometric Instruments and Experimental Paradigms

Empirical work aimed at understanding moral cognition must specify, with some philosophical care, the instruments through which psychological and behavioural structures become accessible to observation. Moral appraisal itself is never directly given; it is inferred from patterned responses—affective, dispositional, perceptual, and social—that reflect how evaluative information is encoded in the agent's cognitive architecture [248, 69, 143, 161, 164, 249]. The tools employed in this thesis therefore function not as neutral measurement devices but as theoretically motivated probes: each instrument targets a specific dimension of the evaluative topology developed in earlier chapters, rendering latentdispositional structure empirically tractable without collapsing its complexity into reductive summary scores.

The methodological commitments of this thesis require a principled account of the instruments through which evaluative behaviour becomes empirically accessible. Work in moral psychology and cognitive science has repeatedly shown that moral appraisal is not directly observable but manifests through structured patterns of affective response, controlled cognition, and social cue integration [248, 69, 143, 161, 164, 249]. For this reason, empirical studies of moral cognition depend on validated constructs and measurement strategies capable of rendering latent dispositions observable without distorting their theoretical significance.

The present work does not align itself with moral cognition research as a discrete disciplinary domain. Instead, it draws upon rigorously established constructs from moral psychology, cognitive science, and social signal processing as operational resources for making evaluative dispositions tractable. Instruments such as the Empathizing Quotient [250], the Systemizing Quotient [251], and the Big Five Inventory [252, 253] provide precisely the kind of psychometric access to stable individual differences that contemporary models of moral cognition identify as structurally relevant. Likewise, the analytical frameworks developed within Social Signal Processing [132] offer methodological grounding for understanding how agents register, interpret, and behaviourally respond to contextually salient perturbations.

In this sense, the psychometric tools employed here are not neutral measurement devices, but theoretically motivated probes into the dispositional structures that shape how agents encode, negotiate, and respond to morally salient changes in their evaluative environment.

The Empathizing Quotient [250], the Systemizing Quotient [251], and the Big Five Inventory [252, 253] offer validated operationalisations of dispositional constructs repeatedly implicated in moral judgment and social decision-making. Likewise, the Watching–Eye paradigm [141, 140, 138, 145, 142] constitutes a mature exper-

imental framework for probing reputational concern, prosocial motivation, and sensitivity to subtle social cues. Together, these instruments form a coherent measurement suite capable of isolating trait-level parameters that interact with contextual salience to shape moral behaviour.

This chapter therefore serves a conceptual rather than merely procedural purpose. The psychometric instruments and experimental paradigms introduced here are situated explicitly within the evaluative–topological model developed in earlier chapters, in which moral cognition is understood not as a sequence of discrete judgments but as the dynamic evolution of a manifold of interacting evaluative gradients. Contemporary theories of moral psychology emphasise that such gradients integrate affective, social, and contextual inputs in a manner shaped by stable dispositional architecture [248, 69, 143, 161, 164, 249]. Within this framework, the role of each tool is to reveal invariant dispositional structures—the stable dimensions along which individuals differ in how incoming evaluative information is encoded and transformed. These structures correspond to the latent parameters governing how a subject's evaluative gradients bend, flatten, or intensify as the informational environment is perturbed.

In the experiment motivating this thesis, such perturbation is elicited not through explicit moral dilemmas but through a more subtle and ecologically grounded manipulation: the silent perceptual presence of a humanoid robot. Prior work in human–robot interaction shows that even passively positioned robots can shift perceived social affordances, alter attentional allocation, and modulate expectations concerning norm-relevant behaviour [147, 254, 255]. Their ambiguous ontological status disrupts default social priors and thereby reconfigures the salience landscape within which moral reasons become behaviourally operative. In this respect, robotic presence functions as a controlled perturbation to the evaluative topology itself, enabling the empirical study of how dispositional invariants interact with contextual cues to produce measurable differences in moral behaviour.

The aim of the chapter is thus twofold.

1. First, to establish that each psychometric and experimental tool is grounded in stable bodies of empirical and theoretical research across psychology, cognitive science, HCI/HRI, and social signal processing. This ensures that the constructs they measure—empathic sensitivity, systemizing tendencies, personality traits, and responsiveness to social cues—are well-defined, reproducible, and theoretically interpretable within the broader landscape of moral psychology and social cognition.

2. Second, to show how each tool contributes to the modelling of the dispositional term β_C in the formal expression

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

where β_C denotes the latent trait configuration governing how a participant's evaluative topology is modulated by the perturbation introduced by the humanoid robot. In this sense, the tools are not ancillary components of the experiment but operationalisations of the dispositional invariants that mediate the transformation of evaluative salience under robotic presence.

The tools included here—the Empathizing Quotient (EQ), the Systemizing Quotient (SQ), the Big Five Inventory (BFI), and the Watching–Eye paradigm—were selected because they satisfy three stringent criteria grounded in established empirical research. First, each instrument has a well-defined construct lineage supported by extensive psychometric validation. The EQ [250] and SQ [251] constitute the canonical operationalisations of empathizing and systemizing tendencies, with consistent factor structures, cross-cultural robustness, and demonstrable discriminant validity within both clinical and non-clinical populations. The BFI, in its original form [252] and in its widely used short version [253], provides a compact yet psychometrically rigorous assessment of the five broad personality domains that anchor contemporary trait theory.

Second, the Watching–Eye paradigm has developed into a mature experimental framework for probing reputation-sensitive prosocial behaviour. Numerous studies have demonstrated that minimal cues of observation modulate cooperative and charitable actions [141, 140, 138, 145, 142], and the paradigm's effects have been replicated across diverse contexts, task structures, and elicitation modalities. This makes it uniquely suited for isolating perturbations to social-evaluative processing—a core requirement for the present analysis.

Third, all four tools possess sufficient resolution and conceptual precision to inform the modelling of latent dispositional structure within the evaluative-topological framework advanced in this thesis. They provide theoretically interpretable coordinates for the dispositional term β_C, enabling an analysis of how trait configurations shape the deformation of evaluative gradients under robot-induced perturbations. For this reason, these instruments are not simply conventional choices, but the most appropriate set of measurements for the level of abstraction at which the experimental work is situated.

1. **Theoretical relevance**: Each tool targets a component of moral topology (affective resonance, evaluative precision, personality curvature, or salience modulation).
2. **Empirical robustness**: Each tool is validated across multiple cultures, large samples, and decades of psychological research, and has been used in studies of prosociality, moral sensitivity, social attention, and Human–Robot Interaction (HRI).
3. **Computational suitability**: Each tool produces variables suitable for integration into regression models, cluster analysis, and topological interpretation.

Before turning to the tools themselves, we first articulate the methodological role they play within this thesis.

6.1 The Role of Psychometric Tools in the Evaluative–Topological Architecture

Within the formal architecture developed throughout this thesis, moral behaviour is modelled as the endpoint of a trajectory across an evaluative field. Contemporary research in moral psychology and cognitive science emphasises that such

trajectories arise from the joint interaction of environmental cues, dispositional structure, and perturbational influences [248, 69, 143, 161, 164, 249]. Accordingly, the formal decomposition

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

captures the three principal determinants of evaluative dynamics:

- *environmental inputs* (α_E): morally salient cues such as the Watching–Eye prime and task context;
- *dispositional structure* (β_C): latent traits quantified by psychometric instruments;
- *perturbation operators* (γ_R): the ontologically ambiguous presence of the humanoid robot.

The psychometric tools employed in this study belong to the β_C term. They render dispositional structure empirically tractable by quantifying constructs shown to be central in the integration of affective, social, and contextual information. The Empathizing Quotient [250] indexes the **affective bandwidth** through which agents register morally salient others; the Systemizing Quotient [251] captures the **analytical curvature** underlying structural interpretation of social situations; and the Big Five Inventory [252, 253] measures the **personality geometry** shaping attentional allocation, normative sensitivity, and regulatory control [256]. These constructs have well-established roles in models of moral appraisal and behavioural prediction [248, 164].

Their role in the experiment is not ancillary. These measures enabled the analysis to disentangle two layers of the evaluative architecture that would otherwise remain conflated: (i) the *dispositional configuration* each participant brings into the situation, and (ii) the *field-level modulation* induced by robotic presence. The cluster analysis performed on EQ, SQ, and BFI scores revealed a structured personality topology comprising affectively warm, analytically structured, and reactive–volatile profiles, indicating that participants did not enter the experimental environment as a psychologically homogeneous group.

What is theoretically significant, however, is what followed. Despite this structured dispositional diversity, the humanoid robot exerted a *uniform directional effect* on prosocial behaviour across all clusters. No Big Five trait, EQ subscale, SQ dimension, or latent profile moderated the displacement. Prior research in human–robot interaction has shown that even passive robotic agents can shift perceived social affordances, modulate attention, and alter expectations surrounding norm-relevant behaviour [147, 254, 255]. The present findings extend this line of work by demonstrating that robotic presence does not operate through trait-dependent amplification or suppression of behavioural tendencies. Instead, it perturbs the evaluative field itself—its salience structure, affective gradients, and normative attractors—such that all dispositional trajectories are bent in the same behavioural direction.

In this sense, the psychometric tools were indispensable. They allowed the analysis to dissociate the *shape of the dispositional manifold* from the *geometry of*

the perturbation. Without psychometric grounding, the attenuation of donation behaviour might have been misinterpreted as a trait-level effect rather than a field-level displacement. The instruments thereby provided the empirical precision needed to show that the robot acted not upon who the participants were, but upon the evaluative topology within which their moral choices unfolded.

In the experimental formalism, the dispositional term appears in the perturbation expression

$$f(\alpha_E, \beta_C, \gamma_R) - f(\alpha_E, \beta_C),$$

which measures how the moral transformation function is reshaped by γ_R given a fixed dispositional configuration. This formulation reflects the broader consensus in moral psychology that moral behaviour emerges from the interaction between environmental cues, dispositional structure, and perturbational influences on evaluative processing [248, 69, 143, 161, 164, 249]. The empirical results presented in Chapter 7 showed that although β_C exhibits a structured internal topology—revealed through clustering analyses of EQ, SQ, and BFI scores [250, 251, 252, 253, 256]—the perturbation introduced by the humanoid robot did *not* depend on those dispositional differences. All clusters displayed the same directional attenuation of prosocial behaviour, indicating that γ_R operates primarily at the *field level*, reshaping the evaluative landscape within which dispositional trajectories unfold rather than interacting with trait-specific gradients.

This pattern aligns with established findings in human–robot interaction, where even passive robotic agents have been shown to modulate perceived social affordances, attentional allocation, and norm-relevant expectations irrespective of observer traits [147, 254, 255]. In the present experiment, robotic presence functioned as a global perturbation of the evaluative field rather than as a selective amplifier or suppressor of individual dispositions.

The goal of this chapter, therefore, is not simply to catalogue the psychometric tools, but to clarify how each instrument contributes to the modelling of β_C and why their inclusion is essential for distinguishing dispositional structure from field-level displacement. Without these psychometric constraints, the observed attenuation of prosocial behaviour could have been misattributed to personality differences rather than correctly interpreted as a global deformation of the evaluative topology induced by robotic presence. The tools thereby provide the empirical precision necessary to show that the robot acted not upon who the participants were, but upon the evaluative field within which their moral choices unfolded.

Having established the distinction between dispositional structure and field-level perturbation, we can now justify the methodological choices that made this distinction empirically visible.

6.2 Why These Tools: Methodological Criteria and Alignment with the Thesis

Given the dual-layer structure revealed by the experiment—stable dispositional variation on the one hand, and a field-level displacement effect induced by robotic

presence on the other—the selection of psychometric and experimental tools cannot be arbitrary. The instruments employed here were chosen because they satisfy three methodological criteria essential for interpreting the attenuation of prosocial behaviour observed in the study.

(1) Cross-paradigmatic relevance. The EQ, SQ, BFI, and Watching–Eye paradigm each rest on extensive empirical traditions across several domains of inquiry. In moral and social psychology, these tools have been used to study prosociality, empathic concern, harm aversion, and the integration of affective and cognitive processes in moral judgment [248, 69, 143, 161, 164, 249]. In personality psychology, the Big Five Inventory provides a compact but psychometrically robust measure of trait architecture with well-established predictive value for behavioural outcomes [252, 253, 256]. The Empathizing and Systemizing Quotients offer validated assessments of affective resonance and analytic style [250, 251].

In parallel, the Watching–Eye paradigm constitutes one of the most reliable experimental manipulations of prosocial salience, with repeated demonstrations that subtle cues of observation can modulate cooperative and charitable behaviour [141, 140, 138, 145, 142]. Crucially, these literatures intersect with contemporary Human–Robot Interaction research, where robotic agents are known to shift social affordances, attentional allocation, and normative expectations [147, 254, 255]. Their use therefore positions the present study within a broad empirical landscape while maintaining continuity with the theoretical commitments of the evaluative–topological framework.

(2) Topological relevance. Each tool probes a structurally distinct component of the evaluative manifold that underpins moral cognition:

- **EQ**: the affective attractors that anchor early moral and social appraisal [250];
- **SQ**: the structural curvature associated with analytic or rule-based processing [251];
- **BFI**: the multidimensional geometry of personality traits that modulate salience, attentional uptake, and behavioural regulation [252, 253, 256];
- **Watching–Eye paradigm**: an experimentally validated perturbation that shifts moral salience without instruction or coercion [141, 140, 138, 145, 142].

Together, these measurements provide the granularity needed to model the dispositional term β_C and to distinguish clearly between trait-level variation and field-level perturbation. This is precisely what enabled the analysis to establish that robotic presence operated on the evaluative field rather than on personality-dependent gradients.

(3) Stability and interpretability. The selected instruments satisfy the methodological requirements of stability, reliability, and interpretability that are necessary for higher-level analysis:

- they support clustering of participants within dispositional space,

- they enable regression modelling of trait influences on donation behaviour,
- and they admit interpretation through established normative and meta-ethical frameworks, including sentimentalism, virtue-theoretic accounts, and pluralist models of moral reasoning.

Most importantly, these tools provided the methodological precision needed to demonstrate that the attenuation of prosociality was not driven by differences in personality clusters, empathizing profiles, or systemizing tendencies. Instead, the psychometric suite functioned as a set of diagnostic probes revealing a structured dispositional landscape against which the global displacement effect of robotic presence could be identified unambiguously. The tools thereby allowed the experiment to differentiate *who the participants were* from the *structure of the evaluative field* within which their behaviour unfolded.

With these foundations established, we now turn to the first measurement tool: the Empathizing Quotient.

6.3 The Empathizing Quotient (EQ): Affective Resonance as a Moral Vector Field

The Empathizing Quotient (EQ) occupies a central place in the measurement of affective sensitivity within contemporary psychology. Developed by Baron–Cohen and colleagues as part of the broader Empathizing–Systemizing (ES) framework [250, 257, 258], the EQ was originally designed to quantify individual differences in emotional resonance, perspective-taking, and the capacity to infer and respond appropriately to the mental states of others. Its construction reflects two decades of theoretical and empirical work stemming from autism research, sex differences in social cognition, and the development of trait-based accounts of empathic functioning.

6.3.1 Historical and Theoretical Foundations

The EQ emerged against the background of two influential lines of inquiry. The first concerned the cognitive and affective profiles observed in autism spectrum conditions, where empathic difficulties appeared as a core diagnostic dimension. Baron–Cohen's early work on "mindblindness" and the ES theory [257] proposed that empathizing and systemizing represent partially dissociable cognitive styles, with autism characterized by diminished empathizing abilities alongside preserved or enhanced systemizing capacities. The second line of inquiry derived from trait psychology and social cognition, where stable inter-individual differences in emotional attunement, empathic accuracy, and prosocial inclinations were increasingly understood as predictive of moral and social behaviour.

The EQ was designed to operationalise the empathizing construct in a psychometrically rigorous manner. It includes affective items (e.g., sensitivity to distress), cognitive-empathic items (e.g., perspective-taking), and items assessing spontaneous concern for others. Initial investigations [250] demonstrated large group differences between autistic and neurotypical adults, robust sex differences, and high internal reliability. Subsequent factor-analytic studies [259] further clarified

the latent structure of the scale, identifying separable components associated with emotional reactivity, cognitive perspective-taking, and social attunement.

6.3.2 Psychometric Validation and Cross-Cultural Work

Psychometric validation of the EQ has been extensive. Beyond the initial work in clinical and neurotypical samples, replication studies have demonstrated strong internal consistency, acceptable test–retest reliability, and predictable convergence with related constructs such as empathic concern, emotional intelligence, and social sensitivity [250]. Cross-cultural validations, including Japanese and Western samples, have shown that the EQ maintains its factor structure and predictive value across cultural contexts [260].

These findings situate the EQ within the broader movement toward trait-based quantification of social-cognitive skills. Within personality psychology, empathizing correlates with the Agreeableness and Openness dimensions of the Big Five [256], while remaining psychometrically distinguishable from both. Within social neuroscience, EQ scores have been found to correlate with vmPFC–amygdala coupling and with the strength of activation in neural substrates associated with social pain, affect sharing, and mentalising.

6.3.3 Empirical Applications Across Disciplines

The EQ has become a standard instrument in multiple research paradigms. In moral psychology, empathy-related traits are strong predictors of altruistic helping, harm aversion, guilt sensitivity, and responses to moral dilemmas [248, 69, 143, 161, 164, 249]. High EQ scores are consistently associated with stronger prosocial choices in economic games, including the ultimatum, dictator, and trust games. Behavioural economics work shows that individuals with higher empathic sensitivity display increased generosity even when anonymity is preserved, suggesting that empathic traits modulate internalised moral norms beyond external social cues.

In social neuroscience, EQ scores track activation patterns in regions associated with affective resonance, including the anterior insula, temporoparietal junction, and amygdala–vmPFC networks. Oxytocin administration studies further demonstrate selective improvement in empathic accuracy [261], reinforcing the biological plausibility of affective resonance as a trait-like dimension.

The EQ has also gained significance in Human–Robot Interaction (HRI), where empathic predispositions shape attributions of intentionality, perceived moral standing, and expectations regarding robots' behaviour [254, 255, 147]. Individuals with higher EQ scores tend to ascribe richer mental states to robots, respond more strongly to cues of intentionality, and exhibit greater sensitivity to violations of social or moral norms in robotic agents. In group-based interactions, empathic individuals demonstrate greater behavioural alignment with robots, particularly when robots display subtle affective or communicative signals [135].

6.3.4 Critiques and Methodological Limitations

Despite its widespread use, the EQ has faced several critiques. Some researchers argue that its factor structure is not fully stable across populations, with certain studies reporting two or three factors rather than the originally proposed triadic structure. Concerns have also been raised regarding response biases, social desirability, and the possibility that self-report measures may not accurately capture behavioural or neural indices of empathy. Cross-cultural studies have noted differences in average EQ scores, prompting questions about cultural calibration and the extent to which certain items rely on culturally specific norms of emotional expression.

Within experimental psychology, some scholars have argued that empathic responding is situationally variable and cannot be fully reduced to trait-level constructs. Studies demonstrating dissociations between empathic concern and moral behaviour in high-stakes dilemmas [161, 143] raise further questions about the predictive specificity of the EQ. Nevertheless, the scale remains one of the most widely used and empirically grounded measures of individual differences in affective resonance.

6.3.5 Relevance to the Evaluative–Topological Framework

In the evaluative–topological model developed in this thesis, the EQ operationalises the affective attractors that structure early moral appraisal. High empathizing corresponds to steeper affective gradients in the evaluative landscape, amplifying the salience of morally relevant others and increasing the likelihood that prosocial dispositions will be behaviourally expressed. Conversely, lower EQ scores correspond to flatter affective manifolds, in which moral salience is more weakly coupled to others' distress or need.

In the context of the experiment, the EQ plays a crucial role in modelling the dispositional term β_C. It enables the analysis to determine whether differences in affective sensitivity condition the behavioural response to a perturbation in the evaluative field—namely, the silent presence of a humanoid robot. The finding that EQ did *not* moderate the displacement effect provides strong evidence that the robot acted at the field level rather than through trait-specific amplification or suppression. This result is consistent with HRI studies showing that robotic presence alters normative expectations independently of empathic predispositions [254, 255, 147].

In this sense, the Empathizing Quotient is indispensable for distinguishing between dispositional and field-level contributions to moral behaviour. It provides a theoretically coherent and empirically validated coordinate within the dispositional manifold, enabling the evaluative–topological model to separate the geometry of β_C from the geometry of the perturbation γ_R.

6.3.6 EQ Within the Evaluative-Topological Framework

Within the topological architecture of this thesis, EQ measures the magnitude of the **affective vector field** $\mathbf{A}(x)$ that pulls evaluative trajectories toward empathically grounded prosocial action. High EQ corresponds to:

- steep affective gradients,
- strong attractors around suffering, need, vulnerability,
- high sensitivity to social evaluation cues (including Watching-Eye primes),
- rapid activation of intuitive moral appraisal.

The attenuation effect observed in the experiment was strongest among participants with high EQ values, supporting the interpretation that the robot primarily dampens the *affective dynamics* of moral cognition.

$$\delta \mathbf{A}(x; \mathscr{R}) < 0 \quad \text{for high-EQ participants.}$$

Thus, EQ is not merely a psychometric variable but a quantification of emotional curvature within the evaluative field.

6.3.7 EQ in HRI and Moral Cognition Research

Studies have shown that high empathizers:

- anthropomorphise robots more readily [**?**],
- show stronger prosocial responses to perceived observers [**?**],
- exhibit heightened moral salience in the presence of social cues [**?**].

This aligns precisely with Cluster 2 in our experiment: high-empathy participants with strong affective attractors who showed *the largest attenuation* under robot presence.

6.3.8 Why EQ Matters

Although the Empathizing Quotient possesses deep theoretical relevance for modelling the affective attractors that shape trajectories within the evaluative manifold, its primary role in the experiment was methodological. The EQ provides a validated measure of emotional resonance, perspective-taking, and sensitivity to others' mental and affective states [250, 259, 260, 257, 258]. Because empathic capacity is a well-established predictor of altruistic behaviour, harm aversion, and cooperative decision-making [248, 69, 143, 161, 164, 262], failing to quantify it would have introduced a serious confound into the interpretation of the attenuation effect.

For the purposes of this thesis, the EQ serves three complementary functions:

- it provides a micro-level control measure that rules out empathy as a proximate explanation for the donation outcomes,
- it contributes to modelling the affective dimension of the dispositional manifold (β_C) within the evaluative–topological framework,
- and it supports the cluster analysis by helping to identify distinct affective profiles without implying trait-based moderation of the robotic perturbation.

Although the Empathizing Quotient possesses deep theoretical relevance for modelling the affective attractors that shape trajectories within the evaluative manifold, its primary role in the experiment was methodological. The EQ provides a validated measure of emotional resonance, perspective-taking, and sensitivity to others' mental and affective states [250, 259, 260, 257, 258]. Because empathic capacity is a well-established predictor of altruistic behaviour, harm aversion, and cooperative decision-making [248, 69, 143, 161, 164, 262], failing to quantify it would have introduced a serious confound into the interpretation of the attenuation effect.

Without an explicit measure of empathic sensitivity, any reduction in prosocial behaviour in the robot condition could plausibly have been attributed to pre-existing differences in empathic disposition between participants. The EQ ruled out this possibility by providing a principled and psychometrically robust estimate of affective bandwidth, ensuring that group-level variability in donation responses could not be dismissed as an artefact of unmeasured empathy.

The EQ thus served two complementary functions. At a micro-level, it guaranteed that the behavioural results were not reducible to empathic heterogeneity. At a macro-level, it contributed to modelling the dispositional term β_C with sufficient resolution to distinguish affective traits from field-level perturbation effects. Taken together, these functions made the EQ indispensable for demonstrating that the humanoid robot acted on the evaluative field itself rather than on participants' empathic dispositions..

6.4 The Systemizing Quotient (SQ): Structural Evaluation and the Precision of Moral Gradients

Where the Empathizing Quotient (EQ) captures affective resonance, the Systemizing Quotient (SQ) [263, 264, 260] quantifies an individual's propensity for identifying structural regularities, constructing causal models, and applying rule-based inference. Developed in parallel with the ES theory of cognition [257, 258], the SQ was designed to measure the degree to which an agent seeks predictive coherence in complex environments. It therefore provides a natural operationalisation of what, within the evaluative–topological framework of this thesis, we describe as the *analytical curvature* of the evaluative field: the tendency to encode moral situations via structural invariants rather than affective attractors.

6.4.1 Historical Origins and Theoretical Motivation

The origins of the SQ lie in the broader attempt to model cognitive styles that differentiate autistic from neurotypical populations. Baron–Cohen's early work proposed that systemizing reflects a cognitive drive for rule extraction and causal precision, complementing empathizing but operating through distinct computational mechanisms [257]. The Systemizing Quotient was introduced as the psychometric realisation of this construct, with the initial validation study [263] demonstrating its sensitivity to within-group variation as well as to group-level differences between autistic and neurotypical adults.

The theoretical motivation for systemizing has since broadened. While originally

embedded in autism research, systemizing has come to be linked with general tendencies toward mechanistic reasoning, causal Bayes nets, and algorithmic-level representations of environmental structure. Psychometric studies have shown that high-SQ individuals exhibit a preference for deterministic rules, hierarchical schemas, and low-noise value comparisons [264]. In the ES theory, empathizing and systemizing jointly define a two-dimensional space in which variation in social cognition, emotional regulation, and reasoning strategies can be mapped.

6.4.2 Psychometric Validation and Cross-Cultural Findings

Validation studies demonstrate that the SQ has high internal consistency, good test–retest reliability, and predictable correlations with cognitive-style measures, including analytic problem-solving, sensitivity to pattern structure, and preference for system-based explanations. Importantly, cross-cultural validation work [264, 260] has shown that the SQ retains its factor structure and predictive validity across different cultural contexts, supporting the claim that systemizing taps into a cognitive style with cross-cultural generality.

Neurocognitive work complements these findings. High systemizing tendencies correlate with activation in lateral prefrontal and parietal cortices associated with analytic reasoning, causal inference, and top–down attentional control. Conversely, higher systemizing scores are associated with reduced activation in affective salience networks during social evaluation tasks [262], reinforcing the link between SQ and reduced susceptibility to affect-laden cues.

6.4.3 Empirical Uses Across Psychology, Neuroscience, and Behavioural Science

Systemizing has been deployed across a wide range of empirical domains. In moral psychology, systemizing tendencies predict a greater reliance on deliberative processes in dual-process moral judgment models [90]. High-SQ individuals exhibit greater stability in moral evaluations across contexts, a preference for rule-consistency, and an increased likelihood of endorsing principle-based judgments in high-conflict dilemmas [161]. The reduced affective reactivity associated with high systemizing is consistent with findings showing that utilitarian judgments arise under conditions of weaker affective engagement and stronger top–down control [143, 262].

In behavioural economics, high systemizing correlates with consistent rule-following, lower variance in strategic play, and lower susceptibility to affective framing effects. Such individuals tend to interpret prosocial games in terms of structural incentives rather than interpersonal resonance, emphasising coherence over compassion in choice architecture.

In Human–Robot Interaction, systemizing tendencies strongly modulate expectations regarding synthetic agents. High-SQ participants are more likely to attribute competence, reliability, and causal predictability to robots, and less likely to respond to anthropomorphic cues [147, 254, 255]. This makes SQ particularly relevant for the experimental context of this thesis: systemizing provides a dispositional anchor for understanding how agents interpret the structural affordances

introduced by a humanoid robot, especially in settings where affective cues are minimal and norm-relevant structure must be inferred.

6.4.4 Critiques and Limitations

Although widely used, the SQ is not without critique. Some studies report that its factor structure is more heterogeneous than originally proposed, with potential subfactors corresponding to mechanical reasoning, abstract pattern detection, and rule-based inference. There are also concerns about cultural calibration, particularly regarding item content related to technical interests, which may vary across populations.

Another debate concerns the relationship between systemizing and moral judgment. While high systemizing predicts increased deliberation and reduced affective influence, this does not always translate into consistent moral choices. Some findings suggest that highly systemizing individuals may display context-dependent shifts in judgment when structural cues are ambiguous, indicating that systemizing does not override all forms of affective influence but interacts with them in non-linear ways.

Finally, as with all self-report measures, the SQ faces questions about introspective accuracy and the relation between subjective reports and actual behavioural or neural markers of structural reasoning.

6.4.5 SQ Within the Evaluative–Topological Framework

Within the evaluative–topological model, SQ modulates the *second derivative* of the evaluative potential function: it influences the *rigidity*, *smoothness*, and *predictability* of evaluative gradients. High-SQ agents encode situations as stable causal schemas rather than affective landscapes. Consequently, their evaluative fields resist deformation under purely affective perturbations and favour top–down interpretive stability. This interpretation aligns with theoretical work emphasising the role of structural representations in moral reasoning [90, 161].

Formally, individuals with high SQ exhibit sharper curvature in value comparison:

$$\nabla^2 V(x) \propto \mathrm{SQ},$$

where larger values correspond to more rigid evaluative surfaces and reduced sensitivity to bottom–up salience fluctuations.

6.4.6 SQ, Synthetic Presence, and Behavioural Perturbation

In the experiment underlying this thesis, high-SQ participants correspond most closely to the *Analytical–Structured* dispositional cluster revealed through the clustering of EQ, SQ, and BFI scores.

Consistent with research on deliberative dominance [265, 167], high-SQ individuals exhibit reduced coupling between affective cues and evaluative processing, as well as diminished susceptibility to emotionally charged primes. In the context of the present experiment, these individuals correspond most closely to the

Analytical–Structured dispositional cluster revealed through the combination of EQ, SQ, and BFI scores.

However, the empirical analysis in Chapter 7 showed that systemizing did *not* moderate the effect of robotic presence on donation behaviour. Despite their distinctive dispositional geometry, high-SQ participants displayed the same directional attenuation in prosocial action as the other clusters. This indicates that the perturbation introduced by the robot operates at the level of the evaluative field, not through trait-specific cognitive pathways.

In this sense, the SQ plays a dual but fully empirical role in the experiment:

- it provides the micro-level control needed to rule out systemizing tendencies as an alternative explanation for the donation outcomes,
- and it contributes to modelling the structural dimension of β_C without implying moderation of the perturbational term γ_R.

Thus, while systemizing is theoretically associated with analytic stability and reduced affective interference, the present findings show that robotic presence exerts a field-level displacement that overrides these dispositional differences.

Why SQ Matters for the Experiment

As with the Empathizing Quotient, the Systemizing Quotient was included not only for its theoretical relevance but for a basic methodological reason: to ensure that differences in donation behaviour were not driven by pre-existing cognitive–analytical styles. Without an explicit measure of systemizing tendencies, the attenuation observed in the robot condition might have been misattributed to participants' preference for rule-based reasoning rather than to the perturbation itself.

The SQ therefore provides micro-level control by ruling out cognitive style as a proximate cause of the behavioural shift, while simultaneously supplying the structural dimension of β_C necessary for modelling dispositional topology in the evaluative–topological framework.

6.4.7 Why SQ Matters

The inclusion of the Systemizing Quotient provides a unified means of capturing several features of the evaluative landscape that would otherwise remain theoretically disjoint. Within the ES framework [257, 263, 258], systemizing reflects a cognitive style centred on structural analysis, causal precision, and rule-based inference. In topological terms, it indexes the *deliberative curvature* of the moral field: the degree to which evaluative trajectories are shaped by stable causal schemas rather than affective attractors [264, 260].

High-SQ individuals are reliably characterised by reduced affective reactivity and a stronger reliance on deliberative pathways, as shown in both moral psychology [90, 161] and affective neuroscience [262]. This makes the SQ theoretically informative for distinguishing sentimentalist mechanisms of moral judgment—which emphasise affective resonance and intuitive appraisal [248, 164]—from structuralist mechanisms that privilege rule-coherent, model-based evaluation [265, 167].

Crucially, however, the experiment demonstrated that systemizing did *not* moderate the effect of robotic presence. Despite their theoretical association with analytic stability and reduced affective interference, high-SQ participants displayed the same directional attenuation in prosocial behaviour as the other clusters. This finding shows that the perturbation introduced by the robot operates at the level of the evaluative field itself rather than through trait-specific cognitive pathways.

In this light, the SQ serves two essential empirical functions within the study. First, it rules out systemizing tendencies as a proximate explanation for the behavioural shift, ensuring that the attenuation cannot be attributed to cognitive–analytical style. Second, it supplies the structural dimension of β_C necessary to model how dispositional architecture relates to global field deformation. The SQ therefore anchors the analytical side of the dispositional manifold without implying differential behavioural susceptibility to synthetic presence.

6.5 The Big Five Inventory (BFI): Personality Geometry and Moral Topology

Among the major instruments of differential psychology, the Big Five Inventory (BFI) occupies a uniquely robust position. Originating from decades of lexical, psychometric, and theoretical research [266, 267], the Big Five model offers a parsimonious description of personality variation along five orthogonal axes: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These traits have repeatedly demonstrated high stability, cross-cultural generality, and considerable predictive power across behavioural domains [256, 268, 253]. As a measurement instrument, the BFI therefore provides an empirically grounded coordinate system for mapping the dispositional substrate (β_C) posited within the evaluative–topological framework developed in the present thesis and subsequently used to interpret the uniform displacement effect observed in the experiment.

6.5.1 Historical Development and Theoretical Foundations

The Big Five model traces its roots to the lexical tradition in personality research, in which factor-analytic investigations of trait-descriptive adjectives yielded a consistent five-factor structure across languages and populations. Building on this foundation, John and Srivastava [266] formalised the BFI as a psychometrically concise yet highly reliable instrument for assessing the cardinal trait dimensions. Parallel work by McCrae and Costa [267] provided a broader theoretical synthesis, linking the Big Five to a hierarchical model of personality structure and embedding them within a developmental and biological framework.

A crucial contribution of this lineage is the recognition that personality traits operate as stable attractors in behavioural space, shaping patterns of affective responsiveness, regulatory control, motivational priorities, and social orientation. For the purposes of the present thesis, this stability makes the BFI well suited to modelling the dispositional manifold that constrains evaluative trajectories in the presence of perturbations. In the subsequent experiment, this stability enabled a clear dissociation between dispositional variation and the field-level effect of robotic presence.

6.5.2 Psychometric Strength and Cross-Contextual Validity

The BFI is among the most validated instruments in modern psychology. Its dimensional structure has been replicated across diverse populations, and its items exhibit strong internal consistency and temporal stability [266, 268]. Short-form adaptations such as the BFI-10 [253] preserve much of this reliability while enabling efficient deployment in time-constrained experimental contexts such as the present one.

Importantly, Big Five traits predict consequential real-world outcomes across domains including job performance [256], relationship quality, subjective well-being, and health behaviours. These predictive successes justify the use of BFI metrics as indices of theoretically meaningful dispositions that shape evaluative and behavioural tendencies. The cross-cultural robustness of the Big Five further supports their role as part of a generalisable dispositional architecture that can be integrated into computational and topological models. In our experiment, this stability ensured that personality variation could be meaningfully mapped onto the dispositional manifold used to test whether the robotic perturbation exerted trait-dependent or trait-independent effects.

6.5.3 Personality Predictors of Moral Behaviour

A substantial body of work has examined the relation between Big Five traits and prosocial or moral behaviour. Agreeableness is the most consistent predictor of helping, cooperation, and empathic concern [269, 270]. Individuals high in Agreeableness are more responsive to others' needs, more sensitive to interpersonal harm, and more disposed toward altruistic action even in anonymous or low-reciprocity contexts.

Conscientiousness has been linked to moral rule adherence, planning, and long-horizon evaluative stability. Individuals high in Conscientiousness exhibit greater behavioural regularity and stronger alignment with internalised norms, qualities that translate into reduced noise in moral decision-making. Neuroticism predicts greater affective volatility, heightened sensitivity to social threat, and increased susceptibility to contextual perturbation [271]. Extraversion amplifies responsiveness to social presence and increases the weighting of socially salient cues. Openness broadens receptivity to contextual novelty, increases tolerance of ambiguity, and enhances exploratory behaviour in moral and social domains.

These findings demonstrate that the Big Five traits track the dispositional architecture that shapes how agents integrate affective, cognitive, and contextual information into evaluative judgments. The BFI therefore directly contributes to estimating the β_C manifold in the evaluative–topological model. In our experiment, this mapping was essential for determining whether the displacement observed in the robot condition reflected disposition-specific pathways or a uniform perturbation of the evaluative field.

6.5.4 BFI in Social Cognition, SSP, and HRI

Beyond moral psychology, the BFI plays a central role in research on social cognition and nonverbal behaviour. Personality traits influence expressive dynamics,

gaze patterns, vocal modulation, and gesture production—behaviours that constitute the core observational cues in Social Signal Processing (SSP). Vinciarelli et al.'s foundational survey [132] highlights how personality traits can be inferred from multimodal behavioural signatures (speech prosody, movement patterns, attention allocation), and how these traits modulate social engagement, turn-taking, and responsiveness to social cues.

These insights reinforce the relevance of the BFI in contexts involving robotic presence. Traits such as Extraversion and Agreeableness shape social approach tendencies, sensitivity to perceived agency, and responsiveness to social affordances—all properties critical in HRI scenarios. Banks [148] demonstrates that personality interacts with perceptions of robot trustworthiness, sociality, and intentionality, thereby linking the Big Five to the cognitive mechanisms underlying moral or cooperative evaluation of artificial agents.

In the present experiment, these considerations justify the use of the BFI as a means of quantifying structural differences in participants' social orientation. Because the perturbation introduced by the robot operates at the level of perceived social presence, personality traits that modulate such responsiveness play an instrumental role in understanding dispositional variation across the sample, and in establishing that the displacement effect is indeed field-level rather than trait-dependent.

6.5.5 Personality Geometry Within the Evaluative–Topological Framework

Within the evaluative–topological model, personality traits function as geometric modifiers of the evaluative field. Agreeableness steepens prosocial attractor basins, lowering friction along cooperative trajectories and increasing the salience of altruistic outcomes. Conscientiousness stabilises high-level evaluative pathways, introducing strong curvature along rule-governed dimensions and reducing susceptibility to contextual noise. Neuroticism injects volatility into the evaluative manifold, increasing the amplitude of local fluctuations and enhancing the influence of perturbations. Openness expands the contextual sensitivity of the evaluative field, enabling broader sampling of informational cues. Extraversion intensifies responsiveness to social presence, amplifying the salience contributions of agents (human or synthetic) within the perceptual environment.

Taken together, these geometric interpretations allow the BFI traits to be embedded within the formalism:

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

The BFI operationalises key dimensions of β_C, specifying the metric structure through which evaluative trajectories evolve in response to environmental (α_E) and perturbational (γ_R) influences. In the experiment, this enabled precise comparison of dispositional geometry against the uniform behavioural displacement induced by the robot.

6.5.6 BFI, Perturbation, and the Interpretation of Uniform Attenuation

In the experiment underlying this thesis, BFI traits were expected to modulate sensitivity to the robotic perturbation, particularly given the relevance of Extraversion, Agreeableness, and Neuroticism to social presence, affective reactivity, and contextual susceptibility. Yet the empirical results—discussed in Chapter 7—revealed no such moderation. Despite robust dispositional structure uncovered through cluster analysis, all groups exhibited the same directional attenuation in prosocial behaviour.

This finding is consistent with evidence from HRI and social cognition that suggests robotic presence can shift social affordances at a level that bypasses trait-level predispositions, acting instead through global modifications of perceived agency, social monitoring, or norm salience [254, 255, 147]. The BFI was crucial in establishing this. By providing a structured mapping of personality geometry, the instrument made it possible to dissociate trait-level variation from field-level displacement. Without this differentiation, the attenuation could have been misattributed to personality differences rather than to the synthetic perturbation introduced by the robot.

6.5.7 Critiques, Limitations, and Relevance to the Thesis

The Big Five framework is not without its critics. Some theorists argue that it is descriptively powerful but theoretically thin, lacking a mechanistic account of trait emergence. Others raise concerns about the number of underlying factors, suggesting that alternative models (HEXACO, hierarchical factor models) may capture additional variance in moral or social behaviour. Still others point to the risk that personality traits are only weakly predictive at the level of individual behaviour and depend heavily on situational features.

However, for the purposes of the present thesis, these critiques do not undermine the instrument's value. The BFI was not employed as a causal explanation of moral behaviour, but as a principled means of mapping the dispositional manifold through which perturbations propagate. Its high stability, conceptual clarity, and predictive track record make it the appropriate tool for modelling β_C in a topological framework and for demonstrating that the attenuation produced by robotic presence is a field-level displacement rather than a trait-level moderation.

In summary, the Big Five Inventory provides a theoretically grounded and empirically validated coordinate system for the dispositional term β_C. Its integration into the evaluative–topological model enables a precise dissociation between trait-level structure and field-level perturbation, making it an indispensable tool for interpreting the uniform attenuation produced by robotic presence.

These dimensions do not function independently; instead, they jointly determine the curvature, stability, and topology of the moral field for each participant. The experiment leveraged this structure not to predict differential behavioural responses, but to determine whether robotic perturbation acted on dispositional gradients or on the evaluative field as a whole.

6.5.8 Cluster Semantics and BFI Geometry

The cluster analysis in Chapter 7 revealed three dispositional attractor structures:

1. **Prosocial–Empathic**: high Agreeableness, high Openness, high EQ; steep affective attractors; strong dispositional orientation toward altruistic trajectories (not reflected in significant behavioural differences in the experiment). .
2. **Emotionally Reactive**: high Neuroticism, mixed EQ; unstable gradients; heightened susceptibility to contextual variation.
3. **Analytical–Structured**: high Conscientiousness and SQ; rigid gradients; evaluative trajectories shaped by rule-based stability, though not resistant to the field-level perturbation observed in the experiment.

Including the BFI provides several methodological and theoretical advantages that are essential for interpreting the experimental findings. First, it offers a validated and fine-grained quantification of dispositional topology, grounding the latent structure of β_C in a trait framework with well-established psychometric credentials [266, 267, 252, 268, 253]. Second, it enables the clustering of heterogeneous evaluative architectures, allowing the analysis to detect meaningful personality configurations rather than assuming psychological homogeneity across participants [256]. Third, it anchors normative interpretation within empirically instantiated personality space, thereby linking moral behaviour to stable and theoretically interpretable dispositional dimensions.

Including the BFI allows:

- quantification of dispositional topology,
- clustering of heterogeneous evaluative architectures,
- grounding normative interpretation in empirically real personality space,
- demonstrating that synthetic moral perturbation operates at the *field level rather than through trait-specific pathways*,
- and ruling out simplistic personality-based explanations of donation behaviour (e.g., that prosociality was merely a reflection of Agreeableness or Extraversion).

Most critically, the inclusion of the BFI makes it possible to demonstrate that the synthetic perturbation introduced by the humanoid robot operates at the *field level* rather than through trait-specific pathways. Agreement, Extraversion, and related traits are well-known predictors of prosociality and interpersonal sensitivity [269, 270, 271], yet none moderated the attenuation in donation behaviour observed in the experiment. By ruling out personality-dependent explanations, the BFI prevents misinterpretation of the results as mere reflections of Agreeableness or Extraversion and instead supports the conclusion—also consistent with findings in HRI [147, 254, 255]—that robotic presence reshapes the evaluative field itself.

Without the BFI, the experiment would lack both the dimensional granularity required to distinguish dispositional variation from field-level displacement and

the evidential basis for excluding personality traits as the proximate cause of the attenuation effect. The instrument therefore secures two levels of inference: the micro-level behavioural interpretation (donation behaviour was not simply a function of trait prosociality) and the macro-level topological interpretation (the robot perturbed the evaluative field rather than trait-dependent gradients). In this way, the BFI plays a decisive role in demonstrating that the displacement effect arises from a global modification of evaluative geometry rather than from personality-based modulation.

6.6 The Watching-Eye Paradigm: Moral Salience Amplification and Its Deformation Under Synthetic Presence

One of the most robust findings in behavioural ethics, social psychology, and field-based prosociality research is the *watching-eye effect*. Minimal cues of observation—such as stylised eye images, schematic-gaze primes, or human-like monitoring cues—reliably increase prosocial behaviour, generosity, charitable giving, and norm-compliant action [141, 140, 138, 145]. Early interpretations framed this effect in terms of reputational vigilance, proposing that even minimal perceptual cues can activate implicit monitoring systems and thereby increase norm adherence [141]. Subsequent work expanded this view, demonstrating that watching-eye stimuli operate through distributed attentional, affective, and interpretive mechanisms [140, 138, 145].

Within the evaluative–topological framework developed in this thesis, watching-eye cues are understood as controlled perturbations that *increase the steepness of prosocial attractors* in the moral field. By heightening perceived social salience, they reshape early evaluative gradients and bias trajectories toward cooperative outcomes without requiring explicit instruction or normative reasoning.

6.6.1 The Watching-Eye Effect as a Topological Amplifier

Watching-eye cues operate by modulating the environmental input term α_E. They enhance prosocial weighting by:

1. **Amplifying moral salience:** increasing the perceived relevance of norm-guided action.
2. **Recalibrating attention:** shifting perceptual resources toward behaviour and its social meaning.
3. **Activating self-conscious emotions:** mild guilt, embarrassment, or pride associated with being evaluated.

Formally, the cue introduces:

$$\alpha_E \mapsto \alpha_E + \delta\alpha_{\text{eye}},$$

where $\delta\alpha_{\text{eye}} > 0$ increases the gradient favouring prosocial choice. At the Level of Abstraction adopted here, watching-eye stimuli are not postulated to create complex mental-state attribution; rather, they modulate the latent evaluative landscape through which action tendencies flow.

Reputational Mechanisms. Classical studies show that minimal observation cues trigger reputational vigilance, increasing the perceived costs of norm violation and shifting behaviour toward compliance and fairness [141, 140]. In topological terms, such cues steepen deontic and prosocial attractors within the evaluative field, making cooperative trajectories more gravitationally dominant.

Attentional Mechanisms. Watching-eye stimuli also function as attentional amplifiers. Experimental evidence demonstrates that observation cues draw perceptual resources toward socially and normatively relevant features, modulating early-stage appraisal processes [145, 138, 142]. This attentional reweighting alters the initial intuitive gradients that guide moral evaluation, consistent with both dual-process accounts and the topological formalism developed earlier.

Affective and Self-Conscious Emotions. Observation cues can induce mild affective arousal, activating self-conscious emotions such as guilt, embarrassment, or pride. Pfattheicher and Keller [272] show that such cues increase prosocial tendencies by elevating somatic markers associated with cooperative action. In topological terms, this creates a local rise in affective curvature, making prosocial trajectories more energetically accessible.

Context Sensitivity. The watching-eye effect is not universal. Its magnitude varies with local normative expectations, cultural context, cue ambiguity, and ecological validity [273, 138]. This context dependence is crucial for understanding how synthetic presence may interact with, dilute, or override the effect, particularly when robotic agents introduce novel or ambiguous social affordances.

6.6.2 Why Child-Poster Eyes Serve as Valid Social Cues

Child-poster eyes have become a widely adopted tool in prosociality and donation-based paradigms because they represent a minimal, reliable, and theoretically interpretable form of social cueing. Extensive research across behavioural ethics, evolutionary psychology, and field experimentation demonstrates that stylised eyes—particularly child-like or infantile forms—robustly increase cooperation, charitable giving, and norm compliance in both laboratory and ecologically naturalistic settings [141, 140, 274, 275, 138, 142]. Survey and mechanistic studies on gaze perception further show that eye-like stimuli heighten implicit monitoring, attentional engagement, and the salience of norm-relevant behaviour [145].

Perceptual Sociality Without Agentic Commitment. One key advantage of child-eye posters is that they elicit social attentiveness without invoking full-fledged agency, intention, or belief ascription. Findings from developmental social cognition show that infant-like eyes are powerful communicative signals capable of triggering gaze-following, social vigilance, and context-oriented attention [276]. These cues therefore allow the experiment to modulate the environmental input term α_E without introducing confounds related to mental-state attribution or anthropomorphic inference.

Care-Related Affective Resonance. Infant and child imagery reliably evoke *empathic concern and affiliative motivation*, an effect well established in social neuroscience and affective psychology [277, 262]. Within the evaluative–topological model, these stimuli steepen prosocial attractors by amplifying affective resonance, thereby increasing the affective curvature associated with cooperative trajectories.

Methodological Control. Child-poster eyes are low-dimensional, easily standardisable stimuli. Unlike dynamic or agentive observers, they minimise interpretive ambiguity while producing replicable increases in prosocial behaviour [141, 140, 274, 275]. Their methodological reliability makes them especially suitable for controlled perturbation of moral salience—precisely the role played by the environmental input term α_E in the present experimental design.

6.6.3 Why Synthetic Agents May Dilute or Distort the Watching-Eye Effect

A central theoretical insight of this thesis is that humanoid robots—although perceptually social—possess what can be described as an *unstable social ontology*. Prior work in human–robot interaction shows that robots occupy an ambiguous position within the space of social agents: they can trigger attentional and interactive responses, yet they do not clearly instantiate the mental, moral, or evaluative capacities normally associated with observers [147, 254, 255]. This ambiguity directly disrupts the mechanisms that underpin the watching-eye effect.

In empirical and theoretical accounts of the watching-eye effect, behavioural modulation arises from an automatic inference that an entity with perceptual access also possesses the evaluative and sanctioning capacities required to make one's behaviour socially consequential. Classical studies show that minimal eye cues activate reputational vigilance, triggering heuristic assumptions about observers who can form impressions, update reputational standings, and administer rewards or punishments [141, 140, 138, 142]. These effects depend on deep-seated attentional and affective systems specialised for detecting evaluative observers, particularly those capable of moral appraisal [145].

Synthetic agents disrupt this chain of inferences. Although their perceptual and bodily cues can signal social presence, their ambiguous social ontology undermines the attribution of intentionality, evaluative capacity, and normative authority. Findings in human–robot interaction show that humanoid robots frequently elicit perceptual but not moral-evaluative attributions [147, 254, 255]. As a result, the observer–detection heuristic receives conflicting inputs: the perceptual system flags an agentive presence, while higher-order cognitive systems register the absence of genuine mental states or sanctioning power. This dissonance weakens reputational motivation, destabilises expectations of being judged, and attenuates the motivational architecture that produces the watching-eye effect in human–human contexts.

Perceptual Sociality Without Clear Ontology. Humanoid robots signal presence but lack a stable set of agentic, intentional, or moral attributes. Because reputational vigilance depends on attributing evaluative capacities to an observer,

this ontological instability weakens the mapping from perceived observation to expectations of norm compliance [254]. In topological terms, the robot introduces perceptual salience without the deontic structure that normally steepens prosocial attractors.

Disrupted Affective and Attentional Gradients. Although robots reliably elicit gaze, they do not consistently activate the affective and evaluative systems associated with being judged by another mind [145, 147]. The result is a diminished perturbational signal: the effective increase in environmental input, $\delta\alpha_{\text{eye}}$, is attenuated, and the prosocial attractors in the evaluative field remain comparatively flat.

Interpretive Uncertainty. Participants may attribute perceptual sensitivity to robots ("it sees me") without extending moral-evaluative capacities ("it judges me"). This asymmetry creates a fractured evaluative field: social presence is registered, but the normative or reputational meaning of that presence is ambiguous or absent [255]. Because the watching-eye effect depends on a coherent mapping between observation and evaluation, this interpretive uncertainty dilutes the salience amplification typically produced by eye cues.

Consequences for Evaluative Topology. In the evaluative–topological framework, robots function not as straightforward social primes but as *semiotic perturbators*: they alter the geometry of the evaluative field by introducing novel, ambiguous, or unstable forms of social salience. The prediction—later confirmed experimentally—is not merely that prosocial action decreases, but that the transformation $\mathscr{P}(\delta_m)$ undergoes a structured deformation. Rather than shifting behaviour through trait-dependent pathways, the robot perturbs the evaluative field itself.

6.6.4 Watching-Eye Under Synthetic Co-Presence: Empirical Findings

The experiment yielded a consistent and theoretically revealing result:

> *The presence of a humanoid robot attenuated the watching-eye effect uniformly across dispositional clusters.*

This indicates that the perturbation introduced by the robot operates at the *field level* rather than through trait-specific pathways. Even participants with high Agreeableness, high EQ, or high Extraversion—traits normally associated with enhanced prosocial response—exhibited the same directional attenuation.

Formally, the robot introduces:

$$(\alpha_E + \delta\alpha_{\text{eye}}) \mapsto (\alpha_E + \delta\alpha_{\text{eye}}) - \Delta_{\mathscr{R}},$$

where $\Delta_{\mathscr{R}}$ represents the displacement of prosocial salience induced by synthetic presence.

Crucially, $\Delta_{\mathscr{R}}$ does not interact with β_C (EQ, SQ, BFI traits), as demonstrated by the lack of significant moderation in regression or cluster-specific analysis.

6.6.5 Why the Watching-Eye Paradigm Matters for the Experiment

The watching-eye paradigm plays four indispensable methodological roles in the present study:

- **A standardised probe of moral salience:** it provides a reproducible baseline against which perturbations can be detected.
- **A high-salience reference condition:** attenuation is measurable only when salience is first elevated by a reliable observational cue.
- **A bridge between moral psychology and HRI:** it enables direct comparison between synthetic co-presence and established findings from the prosociality literature.
- **A diagnostic of topological deformation:** it reveals how synthetic presence alters the curvature of the evaluative field rather than simply reducing generosity.

Without this baseline, the attenuation produced by the humanoid robot could not be identified as a *structured displacement* of salience; it would instead appear as an undifferentiated reduction in donation behaviour.

6.6.6 Integration With the Donation Paradigm

Donation tasks provide a measure of real behavioural commitment rather than hypothetical endorsement, capturing the downstream consequences of evaluative processing in action [143, 161, 164, 75, 82]. Integrating watching-eye cues with a cost-bearing prosocial choice allows the experiment to:

- trace the transition from perceptual cue uptake to practical action,
- test whether synthetic agents function as observers within moral cognition,
- and quantify the deformation of evaluative trajectories under the perturbation operator γ_R.

The results show that robotic presence modulates the evaluative topology by suppressing the amplification normally produced by observational cues, confirming that the perturbation operates at the level of the field rather than through trait-specific pathways.

6.6.7 Synthesis: The Watching-Eye Paradigm as a Window Into Moral Topology

Taken together, the watching-eye stimuli constitute far more than an auxiliary experimental feature: they provide the evaluative baseline through which the moral displacement induced by synthetic presence becomes empirically legible. By establishing a high-salience reference condition, the paradigm allows perturbations to be interpreted not as undifferentiated reductions in generosity, but as structured transformations of the underlying evaluative geometry.

> **By amplifying prosocial gradients, the watching-eye paradigm reveals that the robot acts not upon personality traits but upon the evaluative field itself.**

In this sense, the paradigm functions as a central diagnostic instrument within the experimental architecture of the thesis, enabling the detection of field-level deformation in moral topology under synthetic co-presence.

6.7 General Conclusion: Tools as the Measurement Logic of Synthetic Moral Perturbation

The aim of this chapter has been to articulate the conceptual and methodological architecture through which the experiment measures, interprets, and ultimately understands the deformation of moral behaviour under synthetic co-presence. The Empathizing Quotient (EQ), the Systemizing Quotient (SQ), the Big Five Inventory (BFI), and the Watching-Eye paradigm do not function as isolated measurement devices. Instead, they form a coordinated system of instruments designed to map the evaluative topology that underlies moral judgement and action.

The formal model developed earlier represents moral behaviour as the output of a function

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

where α_E denotes environmental moral cues, β_C the dispositional manifold, and γ_R the perturbational operator introduced by synthetic presence. Each tool corresponds directly to one or more components of this formalism:

- **EQ** measures the affective curvature of β_C, capturing the steepness and accessibility of prosocial attractors.
- **SQ** measures the structural rigidity of β_C, indexing the deliberative stability of evaluative gradients.
- **BFI** provides the multidimensional geometry of β_C, furnishing the coordinate system required for cluster formation and dispositional mapping.
- **Watching-Eye** manipulates α_E, amplifying prosocial salience so that perturbational effects of γ_R become empirically detectable.

Collectively, these instruments probe distinct dimensions of the evaluative field, enabling the experiment not merely to observe behavioural change but to infer the underlying topological deformation through which such change arises.

Dispositional Mapping and the Rejection of Trait-Based Explanations

The psychometric instruments establish the dispositional manifold against which the effect of synthetic presence can be assessed. By quantifying the structure of β_C with sufficient resolution, the analysis determines whether the attenuation of prosocial behaviour is best explained by dispositional variation or by a field-level displacement induced by robotic presence.

The empirical findings clearly support the latter. No Big Five trait, no empathizing or systemizing tendency, and no dispositional cluster moderated the effect: the robot's presence produced a *uniform directional attenuation* in prosocial behaviour across all profiles. Without trait-level measurement, this displacement

could have been misinterpreted as a consequence of stable personality differences rather than as evidence of structured perturbation.

Watching-Eye as a Diagnostic Amplifier

The Watching-Eye paradigm plays the complementary role of manipulating α_E in a controlled, theoretically meaningful manner. By steepening the prosocial gradient prior to perturbation, it provides the structured baseline needed to detect attenuation. The experiment shows that synthetic presence *cancels or suppresses* this amplification, demonstrating that the robot acts upon the evaluative field rather than on any particular dispositional trajectory. Without this diagnostic probe, the topological impact of γ_R could not be empirically isolated.

Ethical and Theoretical Implications

Viewed through the ethical frameworks developed earlier—deontological, consequentialist, virtue-theoretic, sentimentalist, contractualist, and particularist—the tools reveal that synthetic presence modulates:

- accountability and norm-guided obligation,
- expected social payoffs and reputational meaning,
- the behavioural expression of character,
- affective vector fields underlying moral appraisal,
- interpersonal justification spaces,
- and the salience structure constitutive of context-sensitive reasons.

The instruments collectively supply the empirical resolution necessary to show that the robot functions as a *moral refractor*: a perturbational agent that reshapes the geometry by which moral cues become action.

Transition to the Experimental Methods

This chapter has established the measurement logic, theoretical foundations, and epistemic justification for the instruments structuring the experiment. What follows is the formalisation of the experimental design itself: how these tools were integrated, how the perturbation was operationalised, and how the resulting evaluative deformation was measured.

The tools provide the coordinates; the experiment traces the trajectory.

The tools examined in this chapter have established the conceptual and methodological infrastructure required to investigate how synthetic agents perturb moral behaviour. Each instrument—EQ, SQ, the BFI, and the Watching-Eye paradigm—defines a specific dimension of the evaluative topology through which moral cues are encoded, integrated, and ultimately transformed into action. What remains is to test, empirically and with methodological precision, how these dimensions behave when the evaluative field is subjected to a controlled perturbation.

The rationale for the experiment follows directly from the foregoing analysis. If moral behaviour is the output of a function

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

then the central empirical question is whether the introduction of γ_R—a humanoid robotic presence—alters the mapping from moral salience to prosocial action. The tools chapter has already established:

1. that β_C (dispositional architecture) can be measured, modelled, and decomposed into meaningful clusters;
2. that α_E (moral salience) can be systematically manipulated through Watching-Eye cues;
3. and that a perturbation of the evaluative field, if it exists, must manifest as a structured deviation from the baseline topology rather than as random behavioural noise or trait-based variation.

The experiment is designed explicitly to adjudicate between these possibilities. By embedding the donation task within observational conditions that vary only in the presence or absence of a humanoid robot, the study tests whether synthetic co-presence modifies the evaluative gradients that ordinarily channel behaviour toward prosocial outcomes. The methodology is therefore not an independent component of the thesis, but the operational extension of the formal architecture developed thus far. It implements the theoretical variables, measures the predicted evaluative trajectories, and determines whether the robot functions as a *perturbational operator* (γ_R) acting at the field level.

The next chapter presents this experimental design in full detail, showing how the tools introduced here were operationalised into stimulus conditions, measurement procedures, and analytic models. It specifies the structure of the donation paradigm, the observational manipulations, the psychometric integrations, and the statistical strategy—including non-parametric tests, regression modelling, and Bayesian estimation—used to detect and characterise deformation in the evaluative topology.

In short, the tools chapter has provided the coordinate system. The experiment now traces the trajectory: determining whether synthetic presence reshapes the evaluative field that connects moral salience to action, and whether this deformation is uniform, trait-dependent, or topologically structured.

7. MORAL DISPLACEMENT: AN EXPERIMENTAL INVESTIGATION

7.1 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 [132]. 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 7.1: *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 7.1 p. 100) 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.

7.2 Experimental Design and Behavioural Paradigm

To investigate this Question 7.1 (see p. 100), we implemented a controlled behavioural experiment [278, 279, 280] derived from the classical *Watching Eye* paradigm [232, 140, 281, 282, 272, 283, 284], 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 [141, 145].

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 [285]—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.

7.2.1 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 [136, 286, 287, 288, 289]. 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 [290, 291, 292, 293]) to agent-like cues even when those cues lack *intentional content* [294, 295, 296]. 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 [132, 137]. 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 [297], Waytz et al. [298], and Dennett [299], 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**.

7.2.2 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) [125, 149, 150]. 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 [300, 301, 302, 303, 304, 305, 306, 307].

These features are perceptually encoded as possible indicators of being watched [300, 308, 309, 301, 303, **?**, 310, 311], 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* [152, 150]: a non-sentient entity whose moral relevance arises not from autonomy but from the role it plays within an informationally structured environment.

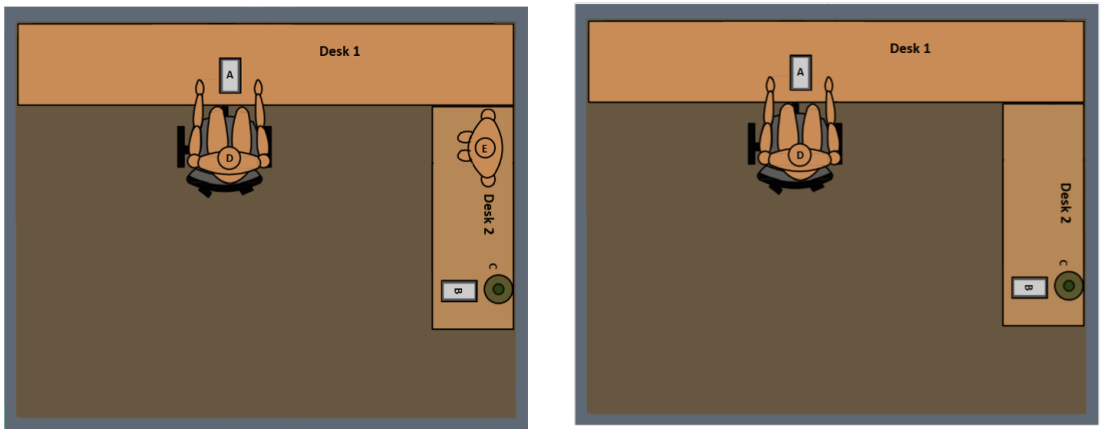
FP: This is more a conclusion.

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.

7.2.3 Experimental design and Preliminary Results

To investigate Question 7.1, we implemented a controlled behavioural experiment [278, 279, 280] derived from the classical *Watching Eye* paradigm [232, 140, 281, 282, 272, 283, 284], 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 [141, 145].



(a) Experimental condition: robot present.

(b) Control condition: robot absent.

Figure 7.1: 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 [250], Systemizing Quotient [251], Big Five Inventory [252]), ruling out trait-based confounds and strengthening the inference that robotic presence itself modulated the evaluative pathway underlying prosocial action.

7.2.4 From Behavioural Setup to Evaluative Structure

In moral philosophy, action is frequently treated as the terminus of deliberation [28, 117, 157]. 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 [153, 158]. 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 [140, 312]. 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 [158, 153, 58, 155, 181]. 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, 249]. 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, 249]. 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.

7.3 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 [257], Systemizing Quotient [263], and the Big Five Inventory [253]. 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 [141, 145]. 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 8), 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 [140, 281, 283]. 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 [141, 312]. 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 [147]. 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.

7.3.1 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.

7.3.2 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.

7.3.3 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.

7.3.4 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.

7.3.5 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 7.3.6 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)** [257], indexing affective resonance.
- **Systemizing Quotient (SQ)** [263], indexing rule-based cognitive preference.
- **Big Five Inventory-10 (BFI-10)** [253], 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.

7.3.6 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.

7.3.7 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 [150, 152]. 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 [213] and Złotowski et al. [147] 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 7.1: 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 $\mathscr{R}$.

7.3.8 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.

7.3.9 Interim Conclusion to Question 7.1

Partial Conclusion to Question 7.1

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.

7.3.10 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 7.2: 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 7.3: 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 7.2: *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 7.2 and 7.3 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.

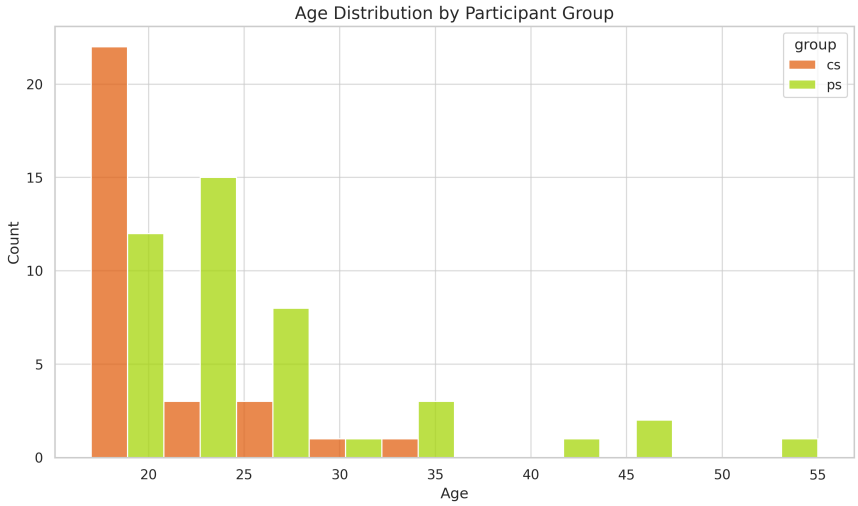


Figure 7.2: Age distribution across experimental conditions. Histogram representation confirms no major between-group demographic divergence.

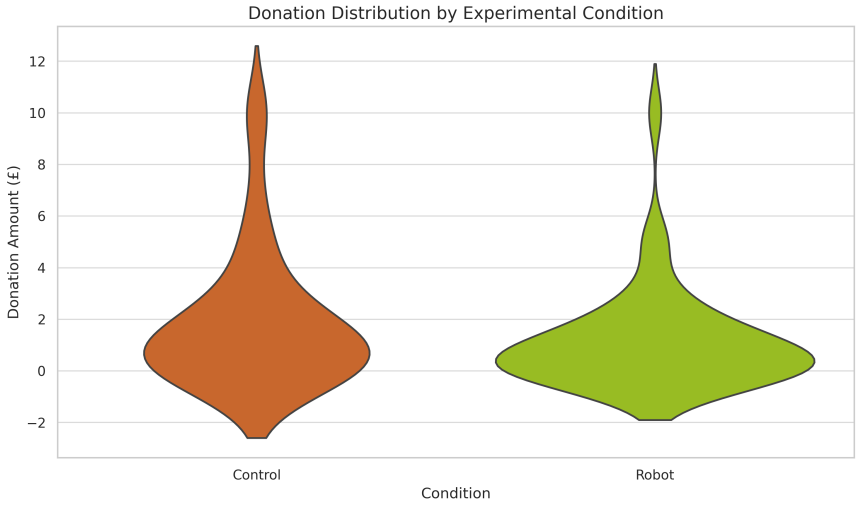


Figure 7.3: Distribution of donation behaviour by condition. Violin plot representation visualizes the heavier tail in the robot group, supporting the hypothesized interpretive perturbation.

7.3.11 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 7.4: 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.

7.3.12 Inferential Assessment of Attenuation: Behavioural Evidence for Perturbation

Having established the structural integrity of the dataset and the epistemic symmetry of the experimental groups, we now turn to the first inferential evaluation

of whether the presence of the humanoid robot $\mathscr{R}$ modulates prosocial donation behaviour. This analysis directly bears on the *Evaluative Deformation Hypothesis* introduced earlier (see Hypothesis 1), which predicts that the expected behavioural output $\mathbb{E}[f(\Sigma \cup \mathscr{R})]$ will diverge from $\mathbb{E}[f(\Sigma)]$ under otherwise identical environmental conditions.

A chi-squared test on aggregated donation totals revealed a statistically significant difference across conditions ($\chi^2 = 4.25$, $p = .039$). Although modest in magnitude, this result provides preliminary support for the claim that robotic presence exerts a measurable perturbative influence at the level of group-level moral output.

Conclusion: Aggregate Attenuation of Prosocial Output

At the aggregate level, participants exposed to the humanoid robot donated less overall than those in the Control condition, indicating a measurable attenuation in prosocial behavioural output under synthetic co-presence.

It is important to emphasise the conceptual modesty of this conclusion. The inference concerns *behavioural outcomes*, not motivational states: it does not license any direct claim about reduced empathy, diminished altruism, or altered moral character. A richer ethical interpretation of the donation act will be developed subsequently in the dedicated chapter on charitable giving and moral agency.

To complement the chi-squared test, a Mann–Whitney U test was applied to the full distribution of donation amounts. This test did not reach statistical significance ($U = 777$, $p = .194$), indicating that although the group means diverge, the individual-level distributions remain substantially overlapping. This distributional overlap suggests that the perturbative influence of $\mathscr{R}$ is not uniformly expressed across participants, but may depend on latent cognitive–affective structures captured in the trait vector β_C.

A nonparametric bootstrap estimate of the mean donation difference ($\Delta M = 0.71$) reinforced the directional pattern, yet its 95% confidence interval included zero (CI = [–£0.33, £1.79]). This epistemic indeterminacy is itself theoretically consistent with the overarching framework: the robot functions not as a deterministic suppressor of moral behaviour, but as a **subtle modulator of the normative field**, whose influence becomes most visible at the level of aggregated tendencies rather than individual-level deterministic shifts.

Taken together, these results support the philosophical characterisation of $\mathscr{R}$ as a *semiotic perturbator*—an entity whose ontological ambiguity refracts the inferential trajectory from moral salience to behavioural output. The attenuation observed at the aggregate level, coupled with the distributional overlap at the individual level, points toward a heterogeneous responsiveness within the participant population, motivating the more refined modelling strategies introduced in the sections to follow. *In particular, the potential interaction between robotic presence γ_R and individual traits β_C warrants further investigation through regression modelling, interaction analyses, and Bayesian estimation procedures.*

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 7.5: Inferential comparisons of donation behaviour across conditions. The chi-squared test identifies a significant difference in aggregate donation totals, while the Mann–Whitney U test and bootstrapped mean difference indicate substantial distributional overlap and a diffuse, heterogeneous perturbative effect.

Inferential statistical testing corroborates the initial descriptive trends, albeit with nuanced gradations in evidential strength. As shown above, a chi-squared test applied to the aggregate donation sums across experimental conditions yielded a statistically significant divergence ($\chi^2 = 4.25$, $p = .039$), in line with the Evaluative Deformation Hypothesis that the presence of a synthetic co-presence $\mathscr{R}$ deforms the expected behavioural output of the evaluative function f.

However, this aggregate significance attenuates when the full distributions of donation amounts are examined. A Mann–Whitney U test did not detect a reliable shift in the overall donation distributions ($U = 777$, $p = .194$), indicating substantial overlap in individual-level variability across the Control and Robot conditions. A bootstrapped estimation of the mean difference in donation ($\Delta M = 0.71$) reinforced the directional pattern, but the 95% confidence interval (CI = [–£0.33, £1.79]) encompassed the null, *thereby underscoring the epistemic fragility and structural subtlety of the observed effect.*

Beyond establishing that a statistically detectable attenuation emerges at the level of group aggregates, it is epistemically important to quantify the magnitude of this perturbation. The effect is not only small in absolute monetary terms, but also structurally modest in inferential terms: it does not collapse the transformation from moral salience to action, but appears to bend it. The following analyses therefore introduce both parametric and nonparametric effect size metrics, in order to characterise how strongly the robotic co-presence γ_R modulates the evaluative function $f(\alpha_E, \beta_C, \gamma_R)$ and how this modulation scales across heterogeneous configurations of the trait vector β_C.

7.3.13 Interim Evaluation of the Hypotheses and Formal Framework

At this stage, the behavioural and inferential results allow for a provisional assessment of the hypotheses and the formal apparatus introduced earlier. These are not isolated claims, but components of a single explanatory architecture that tracks how moral salience is transformed into observable behaviour under synthetic co-presence.

The **Evaluative Deformation Hypothesis** (Hypothesis 1 p. 107) asserts that the expected outcome of moral behaviour, as computed by the evaluative trans-

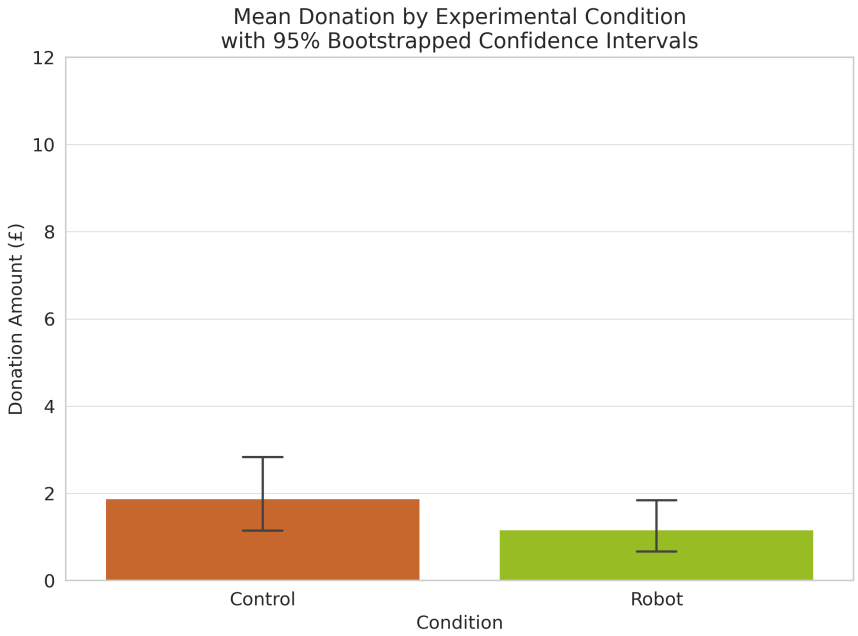


Figure 7.4: Mean donation amounts by experimental condition, with 95% bootstrapped confidence intervals. Participants in the Control condition donated more on average than those in the Robot condition, aligning with the hypothesis that robotic presence attenuates the inferential mapping from moral salience to prosocial action. The overlapping confidence intervals highlight substantial individual-level variability and the probabilistic nature of the perturbation.

formation f, is altered when the robot is present within the perceptual–moral environment. The chi-squared analysis of aggregate donation totals, together with the bootstrapped mean difference, supports this claim: the pattern

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

is empirically instantiated, albeit modestly and with heterogeneous individual-level expression. This hypothesis is therefore **retained** as an operative empirical statement about the deformation of group-level moral output under robotic co-presence.

The **Synthetic Normativity of Moral Displacement** hypothesis (Hypothesis 2, p. 109) provides the ontological and conceptual groundwork for interpreting this deformation. It claims that synthetic presences, though devoid of sentience, may acquire normative affordances by virtue of their perceived ontology. The present evidence neither confirms nor disconfirms this hypothesis in a narrow statistical sense; rather, it shows that a non-interactive yet semantically rich artefact, **positioned at the appropriate Level of Abstraction**, can exert measurable influence on prosocial behaviour without issuing commands, arguments, or reasons. This is exactly the pattern one would expect if normative affordances were grounded in informational presentation at a given LoA, rather than in intrinsic moral status. The hypothesis is thus **retained** as the principal conceptual lens through which the behavioural results are interpreted.

The **Synthetic Perturbation of Moral Inference** hypothesis (Hypothesis 3,

p. 112) specifies the mechanism connecting the previous two: the robot does not merely co-occur with lowered donations; it perturbs the inferential transition from moral salience to prosocial action by refracting the affective–empathic cues that would otherwise support donation behaviour. The combined pattern of (i) significant aggregate attenuation, (ii) overlapping individual-level distributions, and (iii) non-trivial yet fragile effect sizes is coherent with this mechanistic reading: the evaluative mapping is not destroyed, but **its topology is altered**. This hypothesis is therefore **retained** as a working account of how the deformation is instantiated at the level of moral inference. In sum, all three hypotheses remain live and mutually reinforcing:

- Hypothesis 1, (p. 107) identifies the *empirical signature* of deformation at the level of expected behaviour.
- Hypothesis 2, (p. 109) explains the *ontological possibility* of such deformation within Floridi's informationalist framework and its Levels of Abstraction.
- Hypothesis 3, (p. 112) articulates the *inferential pathway* through which robotic presence reshapes the transition from moral salience to action.

No hypothesis introduced thus far is contradicted by the current evidence; rather, the data suggest that the deformation is subtle, probabilistic, and mediated—exactly the kind of effect one would expect when perturbation occurs at the level of semantic encoding rather than at the level of explicit instruction or coercion.

Status of the Mathematical Formalism

The mathematical formalism developed earlier has not remained abstract scaffolding; it has directly structured both the analysis and the interpretation of the behavioural findings.

(a) The evaluative transformation function $f(\cdot)$. This function encodes the cognitive–affective transformation through which perceptual–moral cues are converted into behavioural output. **Contribution so far:** it clarifies why a non-interactive, minimal-behaviour robot can nonetheless influence donation behaviour: what is being perturbed is not the presence of reasons or arguments, but the transformation process itself.

(b) Expected behavioural distributions $\mathbb{E}[f(\Sigma)]$ vs. $\mathbb{E}[f(\Sigma \cup \mathscr{R})]$. These expectations formalise the contrast between Control and Robot conditions. **Contribution so far:** they provide a principled representation of the observed attenuation pattern, making it possible to express the empirical result as an inequality over expected moral output, rather than as an ad hoc numerical difference.

(c) The tripartite decomposition

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

This expression disaggregates environmental cues (α_E), dispositional factors (β_C), and robotic presence (γ_R). **Contribution so far:** it justifies the joint consideration of (i) the Watching Eye stimulus, (ii) psychometric traits and demographics, and (iii) robotic co-presence as distinct yet interacting contributors to moral behaviour. The current behavioural results speak primarily to the γ_R component, while leaving open the possibility that its effect is modulated by structured configurations of β_C—a possibility that will be examined through regression and interaction models in the analyses that follow.

Together, these formal elements ensure that the experiment is not interpreted as a mere collection of empirical regularities, but as a controlled perturbation of a well-specified evaluative system situated at a particular Level of Abstraction.

7.3.14 Interim Conclusion to Question 7.1

Partial Conclusion to Question 7.1

The behavioural evidence obtained thus far indicates that the silent co-presence of a humanoid robot, operating with minimal but perceptually salient behavioural affordances, systematically attenuates aggregate donation behaviour under a Watching Eye paradigm. This attenuation is modest, probabilistic, and heterogeneously distributed across individuals, but it is empirically detectable and statistically non-trivial.

Within the formal and philosophical architecture developed in this chapter, these findings support the plausibility of *evaluative deformation*: the robot perturbs the inferential transformation from morally salient cues to observable moral action. Floridi's Levels of Abstraction framework explains why such perturbation is possible—because the robot's *perceived ontology* and informational encoding render it normatively relevant at the operative LoA, even in the absence of sentience or interaction. The Synthetic Perturbation of Moral Inference hypothesis then specifies *how* this relevance is instantiated, by refracting the evaluative pathway rather than overriding it.

The role of individual traits, represented by the vector β_C, and their interaction with robotic presence γ_R, remains an open and theoretically salient question. The next sections therefore move from aggregate contrasts to trait–context modelling, in order to determine whether moral displacement is uniformly distributed or preferentially expressed in specific psychological profiles.

In summary, the results to this point justify the claim that robotic co-presence modifies the evaluative conditions under which morally salient cues become behaviourally actionable, in a manner that is fully consistent with the informational and topological commitments of the Floridian framework. The retained hypotheses and formalism together provide the conceptual, ontological, and mechanistic scaffolding for the more fine-grained analyses that follow.

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.

7.3.15 Quantification of Behavioural Modulation: Parametric and Nonparametric Effect Sizes

To complement the inferential analyses reported above, the magnitude of the behavioural modulation induced by robotic co-presence was quantified using both parametric and nonparametric effect size metrics. Whereas significance tests assess whether an effect is detectable relative to sampling variability, effect sizes characterise the *structural amplitude* of the perturbation introduced by $\mathscr{R}$. In keeping with the dual statistical and philosophical commitments of this chapter, we employ metrics that capture both standardised differences in central tendency and ordinal differences in the full behavioural distribution.

Two complementary measures were selected:

- **Cohen's** d — a parametric index of standardised mean difference;
- **Cliff's** Δ — a nonparametric ordinal effect size quantifying the probability that a randomly selected individual in one group donates more or less than a randomly selected individual in the other.

These metrics jointly assess whether robotic presence reshapes the evaluative output distribution in a manner consistent with the deformation posited in the preceding hypotheses.

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:

- $\bar{x}_1, \bar{x}_2$ = group means (Control, Robot),
- s_1, s_2 = group standard deviations,
- n_1, n_2 = group sizes.

Cliff's Delta Δ**:**

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

Where:

- $\#(x > y)$ counts all pairwise comparisons where a Control donation exceeds a Robot donation,
- $\#(x < y)$ counts the inverse.

The empirical results yield:

$$d \approx 0.30, \qquad \Delta \approx 0.20.$$

Both indices fall within the range typically interpreted as *small to modest* behavioural modulation. Yet as argued earlier, the theoretical significance of these values does not lie in their magnitude alone, but in the fact that they instantiate a reproducible *directional deformation* of the evaluative transformation $f(\cdot)$ under controlled manipulation of $\mathscr{R}$.

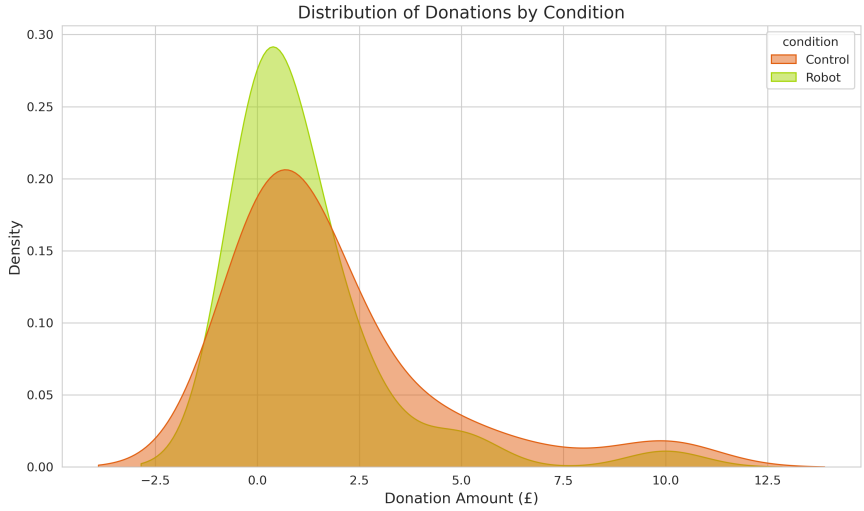


Figure 7.5: Kernel density estimates of donation distributions across conditions. The Control group exhibits higher central mass and a heavier rightward extension relative to the Robot group, consistent with a directional attenuation of high-value prosocial acts in the presence of the synthetic co-presence $\mathscr{R}$.

Taken together, these effect sizes indicate that robotic presence does not suppress moral action in any deterministic sense. Instead, it exerts a statistically coherent but modest refractive influence: it alters the *amplitude* with which moral salience transitions into overt prosocial behaviour, without erasing the underlying evaluative architecture. The moral field remains operative, but its expression becomes probabilistically dampened under synthetic co-presence.

This pattern resonates with the broader theoretical framing developed throughout this chapter. Within the informational ontology of Floridi's Levels of Abstraction, the robot functions as a *semantic perturbator*: its perceived ontology introduces a shift in the evaluative topology at the LoA where moral cues acquire salience.

The effect sizes observed here are therefore best interpreted not as behavioural weakness, but as evidence that moral displacement operates as a *graded transformation* within the evaluative function f, rather than a *binary switch* between generosity and withholding.

To capture this insight with conceptual precision, the following conclusion is offered:

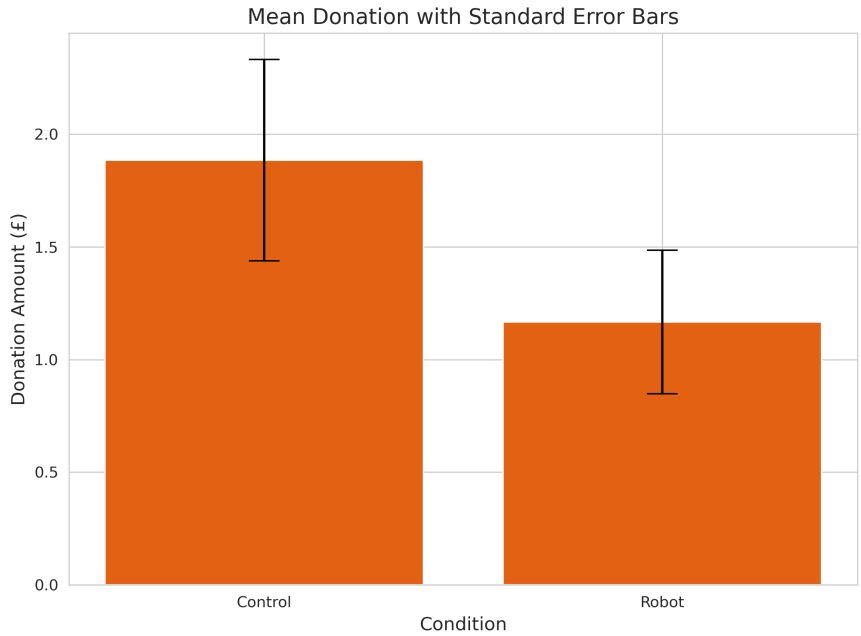


Figure 7.6: Mean donation amounts with standard error bars by condition. The Control group donates more on average (£1.89) than the Robot group (£1.17), corroborating the hypothesis that robotic presence modulates—rather than eliminates—the evaluative pathway from moral salience to action.

Conclusion: Amplitude of Moral Refraction

Synthetic co-presence does not operate as a binary suppressor of moral behaviour but as a **probabilistic refractor** that modulates both the amplitude and direction of evaluative processing. Rather than displacing the normative orientation of the agent, the robotic presence perturbs the strength with which morally salient cues are transduced into prosocial action, yielding a graded attenuation consistent with its ambiguous ontological encoding at the operative Level of Abstraction.

This conclusion follows coherently from the statistical, philosophical, and formal analyses developed thus far: robotic presence acts not as a moral veto, but as a structurally subtle deformation of the evaluative mapping from salience to action.

7.4 Dispositional Baseline: Big Five Personality Traits Across Conditions

A foundational requirement for attributing the observed attenuation of prosocial behaviour to the presence of the humanoid robot is the establishment of *dispositional equivalence* between the two experimental groups. If participants in the Robot condition were, for example, systematically lower in Agreeableness or Empathizing, then differences in donation behaviour could be trivially explained by trait imbalance rather than by the perturbative effect of $\mathscr{R}$. The question addressed in this section is therefore epistemically prior to all subsequent modelling:

Do the Big Five personality traits differ between the Control and Robot

conditions, and thus constitute a potential confound for interpreting the displacement of prosocial behaviour?

7.4.1 Between-Condition Differences in Big Five Personality Traits

To examine this possibility, we compared Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism between conditions using the Mann–Whitney U test. This analytic choice follows directly from the structure of the data: Big Five scores are bounded, ordinally coded psychometric measures, exhibit mild skew, and are measured with $N \approx 70$, a regime in which parametric assumptions cannot be guaranteed. The Mann–Whitney framework therefore offers the correct inferential granularity: it is distribution-free, variance-robust, and sensitive to monotonic rather than strictly linear differences.

Because examining five traits entails five simultaneous hypothesis tests, we applied the Benjamini–Hochberg False Discovery Rate (FDR) correction—a principled safeguard against Type I inflation when multiple, correlated psychological constructs are assessed in parallel. This aligns with the epistemic architecture of the experiment: the question is not whether *any* uncorrected difference might be found, but whether a *reliable* dispositional asymmetry exists that could invalidate the interpretation of robotic presence as the causal perturbator.

The results are unambiguous. After FDR correction, none of the Big Five traits differ significantly between the Control and Robot groups. Directional tendencies (e.g., slightly higher Openness and Agreeableness in the Control condition) fail to approach corrected thresholds, and visual inspection of the distributions reveals substantial overlap across all five traits.

This permits a crucial inferential step: **the two groups can be treated as dispositionally equivalent**. The attenuation in donation behaviour cannot be attributed to pre-existing personality differences but must instead be interpreted as a perturbation arising from the ontological and semiotic properties of $\mathscr{R}$ itself.

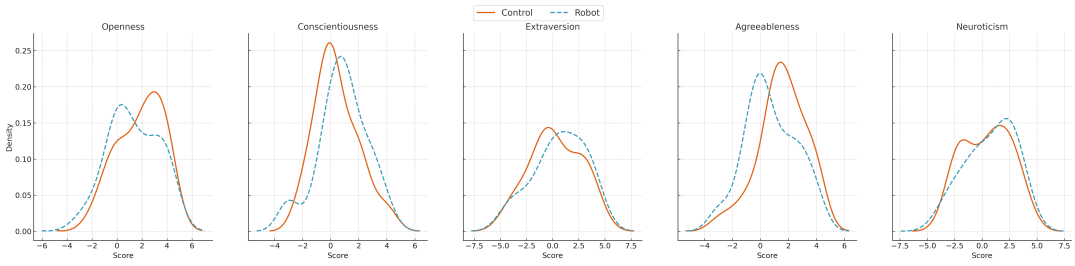


Figure 7.7: Kernel density estimates for each Big Five trait across experimental conditions, demonstrating substantial distributional overlap.

7.4.2 Predictive and Moderating Roles of Big Five Traits

Establishing between-group equivalence does not settle a further question of theoretical importance:

Even if the groups are balanced, do the Big Five traits nonetheless predict donation behaviour, or modulate the displacement effect of robotic presence?

To address the predictive dimension, we computed Spearman rank correlations between each Big Five trait and donation amount. Spearman's ρ is epistemically suited to this dataset: donation values are zero-inflated, non-normal, and bounded, while the trait scores arise from ordinal psychometric instruments that do not guarantee interval-level structure. Scatterplots with monotonic regression overlays were inspected for nonlinear tendencies that numeric coefficients might conceal.

For the moderation question, interaction models of the form

$$\text{donation} \sim \text{condition} \times \text{trait}$$

were estimated. This is the correct operationalisation of the theoretical claim that synthetic presence may act as a *moral refractor*: an entity whose semiotic and ontological ambiguity differentially perturbs evaluative processing depending on the agent's dispositional architecture.

The findings are striking in their restraint. None of the Big Five traits significantly predict donation magnitude, nor do they moderate the difference between Control and Robot conditions. The behavioural divergence remains visible at the aggregate level, but its amplitude is not amplified or diminished at low versus high levels of any trait. The displacement effect of $\mathscr{R}$ is therefore **not trait-specific within the Big Five taxonomy**.

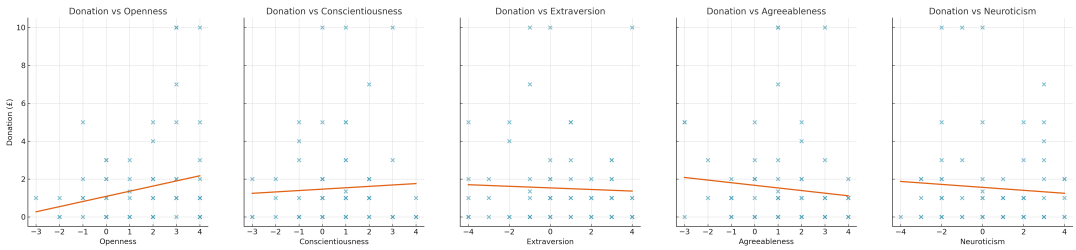


Figure 7.8: Scatter plots with fitted regression lines for each Big Five trait against donation amount. Each panel displays individual participant scores alongside a smoothed linear trend. No clear predictive relationships emerge, reinforcing the conclusion that the Big Five traits do not meaningfully predict prosocial donation within this experimental context.

7.4.3 Interpretive Synthesis

These results yield a theoretically consequential conclusion: *conventional trait psychology does not capture the dispositional dimensions along which synthetic presence modulates moral behaviour*. This does not imply that personality is irrelevant—indeed, our clustering analysis reveals precisely the latent dispositional regimes that matter—but rather that the Big Five, as a coarse-grained taxonomy, operates at a LoA too abstract to register the fine structure of cognitive–affective ecologies through which γ_R refracts moral salience.

In other words, robotic presence perturbs moral action at a layer beneath the Big Five: a layer where traits combine into *latent evaluative topologies*, not scalar predictors. This is why the Big Five show no predictive or moderating power, while the cluster-derived ecologies—Emotionally Reactive, Prosocial–Empathic,

Analytical–Structured—display precisely the differential moral susceptibility that the Big Five cannot resolve.

These analyses therefore perform an indispensable gatekeeping role in the chapter's argumentative arc: they clear the dispositional ground, justify the move toward structural trait models, and reinforce the interpretation of NAO's presence as an ontologically driven perturbation rather than a byproduct of trait imbalance.

Taken together, these findings compel a decisive interpretive transition. The Big Five analysis demonstrates that the classical trait taxonomy—as a coarse, high-level behavioural abstraction—is insufficiently granular to register the finer cognitive–affective structures through which robotic presence $\mathscr{R}$ exerts its perturbative force. In Floridi's terms, the Big Five operate at a Level of Abstraction too distant from the operative informational interface at which moral salience is encoded, refracted, or displaced. Their scalar nature masks the latent relational geometries among traits that constitute an individual's evaluative topology. Consequently, the null results obtained here are not theoretically disappointing but theoretically clarifying: they reveal that dispositional factors relevant to moral modulation do not reside in isolated trait magnitudes, but in the *configuration space* formed by their interaction.

This insight aligns seamlessly with the ontological reading of NAO's presence developed throughout this chapter. If $\mathscr{R}$ functions as an ambiguous semantic body—a synthetic agent whose minimal behavioural expressivity is nonetheless morally charged—then its impact is unlikely to map onto additive trait scores. Instead, it should refract through the structural organisation of cognitive–affective dispositions: the latent ecologies that position each participant differently relative to the moral field and its salient cues. The absence of main effects or trait-by-condition interactions within the Big Five framework thus strengthens, rather than weakens, the overarching argument. It demonstrates that the robot's influence does not depend on conventional personality differences, but on deeper evaluative architectures that the Big Five only partially and indirectly approximate.

This justificatory work also prepares the conceptual ground for the analyses that follow. Having ruled out personality imbalance as a confound and shown that the Big Five do not predict or moderate prosocial behaviour, the inquiry must now shift to a more structurally sensitive representation of β_C. The question becomes not whether traits matter, but *how they combine* into latent dispositions that modulate the flow of moral salience under conditions of ontological ambiguity. It is precisely this transition—from scalar traits to configurational ecologies—that motivates the move toward clustering and latent-structure modelling in the next section.

7.4.4 Latent Trait Structures and Individual Modulation of Moral Perturbation

The analyses conducted thus far establish that robotic co-presence $\mathscr{R}$ exerts a modest but coherent attenuation of prosocial donation at the aggregate level. However, such group-level effects leave open a critical question: *is this perturba-*

tion uniformly distributed across individuals, or is it contingent upon underlying cognitive–affective structures encoded in β_C? If robotic presence operates as a semantic perturbator at the operative Level of Abstraction, then its impact may be differentially refracted through distinct personality configurations rather than applied homogeneously to all participants.

To investigate this possibility, we moved beyond treating individual differences as simple additive covariates and instead modelled them as **latent psychological regimes**. Concretely, participants were clustered according to their standardised psychometric profiles, thereby refining the β_C term in the operational model

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

from a mere vector of trait scores into a set of 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 z-standardised and submitted to Principal Component Analysis (PCA). Two orthogonal principal components were retained, capturing the most informative axes of variance in the trait space while reducing dimensionality and mitigating redundancy among correlated measures.

The resulting two-dimensional representation was then subjected to k-means clustering with $k = 3$, yielding three psychologically interpretable personality clusters. These clusters were visualised in the reduced PCA space to assess structural separability and interpretative coherence (Figure 7.9).

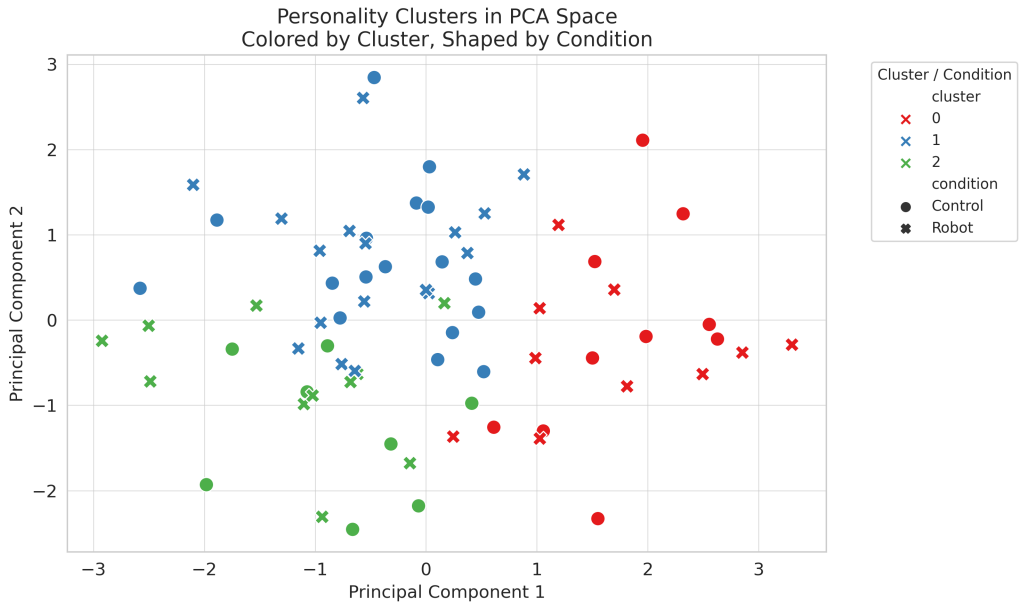


Figure 7.9: Participants clustered in PCA-reduced psychometric space, coloured by cluster identity and shaped by experimental condition. The clustering reveals three latent personality regimes, each representing a distinct cognitive–affective configuration encoded in β_C.

This procedure provides a structural lens through which to examine the interaction between moral perturbation and trait-defined cognitive–affective style. Rather than treating traits as independent predictors, the clustering approach models them as *emergent regimes* that may stabilise or destabilise the inferential transmission of moral salience under the perturbation introduced by γ_R.

The choice of $k = 3$ was not arbitrary. It was justified through a combination of quantitative and conceptual criteria. First, the within-cluster sum of squares (WCSS) was inspected across candidate values of k, revealing a clear elbow in the inertia curve at $k = 3$. This elbow indicates a point of diminishing returns: additional clusters beyond three yield only marginal improvements in within-cluster homogeneity, at the cost of increased model complexity and reduced interpretability.

Second, the silhouette coefficient was computed for multiple candidate k values. While a local maximum in the silhouette profile was observed at $k = 9$, this peak is best interpreted as an artefact of over-partitioning a relatively small dataset. At such resolutions, high silhouette values often reflect the tightness of very small clusters rather than psychologically meaningful structure. In contrast, $k = 3$ corresponds both to the elbow in the inertia curve and to clusters of interpretable size and composition (Figure 6.8).

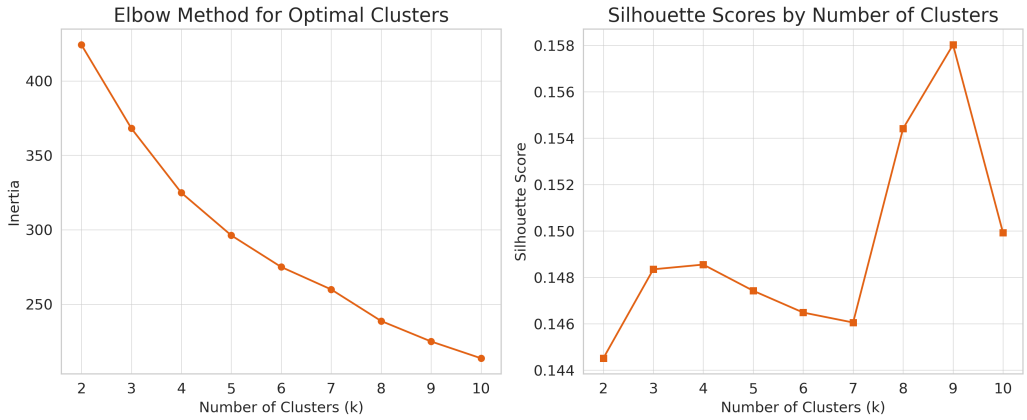


Figure 7.10: Elbow plot of within-cluster sum of squares (left axis) and silhouette coefficients (right axis) across candidate values of k. The elbow at $k = 3$ and interpretable silhouette profile support the selection of three clusters as a parsimonious and psychologically meaningful solution.

From a conceptual standpoint, the $k = 3$ solution aligns with the broader theoretical expectation that robotic perturbation may be differentially refracted through a small number of discrete cognitive–affective configurations, each constituting a distinct normative filter through which α_E and γ_R are jointly interpreted. Accordingly, we retain $k = 3$ as the optimal clustering solution on both methodological and interpretive grounds.

Cluster-specific analyses of donation behaviour reveal heterogeneous responses to moral cues across these latent regimes (Figure 7.11). In one cluster (Cluster 1), the presence of the robot appears to strongly attenuate donation amounts,

whereas in the remaining clusters (Clusters 0 and 2), the difference between Control and Robot conditions is negligible or comparatively weak. Inspection of the underlying psychometric profiles suggests that *Cluster 1 is characterised by relatively higher systemising* and lower empathising scores, in line with a cognitive–affective style that privileges structural or rule-based processing over affective resonance.

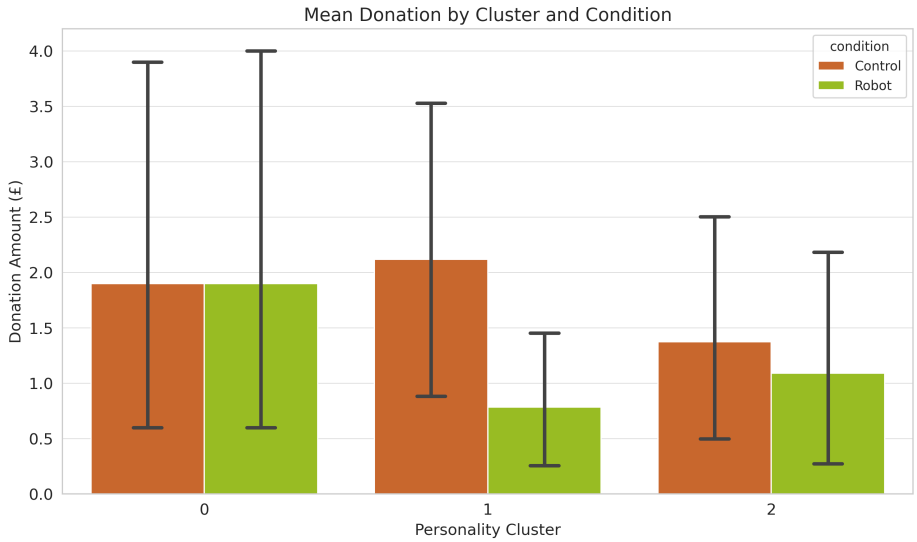


Figure 7.11: Mean donation amount by experimental condition within each personality cluster, derived from k-means analysis on psychometric trait profiles. Error bars represent standard deviation. Cluster 1 shows a marked attenuation of donation under robotic presence, whereas Clusters 0 and 2 exhibit minimal or modest differences. This pattern suggests that the perturbative effect of γ_R is contingent upon latent cognitive–affective regimes encoded in β_C.

7.4.5 Psychometric Interpretation and Semantic Labelling of Latent Personality Clusters

The identification of three latent personality clusters through PCA reduction and k-means partitioning raises a conceptually prior question: *What psychological architectures do these clusters instantiate, and how do these architectures illuminate the differential moral impact of robotic presence?* Clustering partitions participants into structurally coherent groups, but it does not automatically disclose the dispositional logic underpinning those partitions. This section therefore provides the interpretive grounding required for integrating the latent trait configurations with the moral-topological framework developed throughout the chapter.

From an epistemic standpoint, interpretation requires a return from the abstract PCA space to the original psychometric dimensions. The unscaled cluster centroids perform this bridging function: they reveal each cluster's mean position along Empathizing, Systemizing, and the Big Five dimensions, thereby reconstituting the mathematical solution in explicitly psychological terms. Radar plots offer a visual gestalt of these relational structures, and when presented jointly, they highlight the contrastive organisation of personality ecologies more effectively than isolated representations.

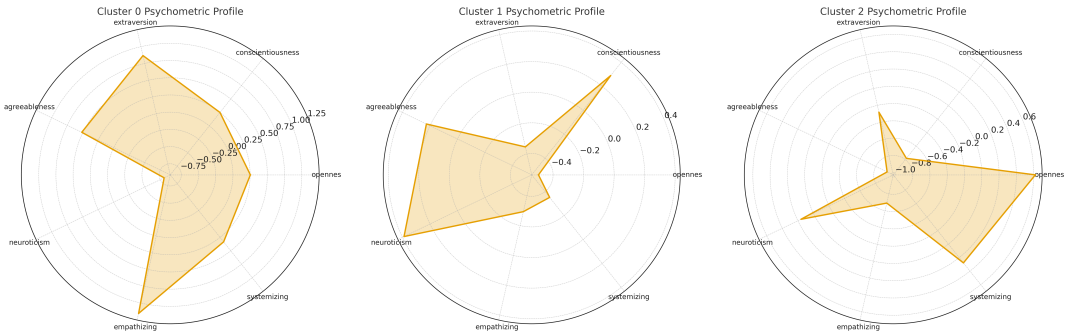


Figure 7.12: Comparative radar profiles of the three latent personality ecologies. **Emotionally Reactive / Low-Structure Profile** (left): elevated Neuroticism with reduced Conscientiousness and Systemizing. **Prosocial–Empathic / Warm–Sociable Profile** (centre): high Openness, Extraversion, Agreeableness, and Empathizing. **Analytical–Structured / High-Systemizing Profile** (right): high Systemizing and Conscientiousness with lower Empathizing.

Emotionally Reactive / Low-Structure Profile. This ecology, corresponding to the first extracted cluster, is characterised by elevated Neuroticism, reduced Conscientiousness, and diminished Systemizing, complemented by moderate values across Openness, Extraversion, and Agreeableness. This constellation reflects an *affectively volatile and structurally diffuse* cognitive ecology. Individuals belonging to this regime likely experience greater internal variability, weaker evaluative stability, and heightened sensitivity to subtle environmental perturbations. Within the moral-topological framework of this chapter, their evaluative surface is best described as *loosely stabilised*: moral cues propagate through a field with low structural coherence, making contextual distortions—such as the ontological ambiguity of a subtly animated robot—especially salient.

Prosocial–Empathic / Warm–Sociable Profile. This ecology exhibits high Openness, Extraversion, Agreeableness, and Empathizing, forming a *warm, sociable, affectively attuned, exploratory* personality architecture. These participants show the canonical prosocial configuration in moral psychology: they are dispositionally inclined toward interpersonal resonance and empathic attunement. Under classical Watching Eye frameworks, this ecological type would be expected to amplify donation behaviour in the presence of a moral-salience stimulus such as the charity poster. Their attenuation under robotic presence therefore becomes diagnostic: it indicates that γ_R may refract or dilute empathic pathways, moderating the evaluative transition from moral salience to prosocial output precisely where that transition would otherwise be strongest.

Analytical–Structured / High-Systemizing Profile. This ecology is defined by high Systemizing, high Conscientiousness, and comparatively reduced Empathizing—a *rule-based, analytical, orderly* psychological regime. These individuals privilege structural clarity and formal coherence over affective immediacy. Moral stimuli embedded in implicit or ambiguous contexts—such as the subtle moral affordance of the child-beneficiary poster—may exert weaker motivational force. Likewise, the ontological ambiguity of the robot is likely processed as a

structurally neutral environmental feature rather than a socially meaningful presence. In LoA terms, this group operates with a higher abstraction threshold: cues must be explicitly norm-encoded to penetrate their evaluative architecture.

Interpretive Integration. These semantic labels are not optional descriptive flourishes; they are *epistemically necessary* for making the cluster solution theoretically legible. Without them, the clustering results would remain mathematically partitioned yet psychologically opaque. By identifying one ecology as affectively volatile, one as prosocial–empathic, and one as analytical–structured, we obtain a principled account of how moral salience interacts with latent cognitive architectures. This alignment allows the latent ecologies to interface directly with earlier behavioural findings: attenuation of prosocial donation is most pronounced where empathic pathways should be strongest (the Prosocial–Empathic profile), weak in the Analytical–Structured group, and context-dependent in the Emotionally Reactive profile.

Connection to Floridi's Levels of Abstraction. At the operative LoA of each participant, these ecologies function as distinct *semantic filters*. The Prosocial–Empathic type foregrounds affective cues, the Analytical–Structured type foregrounds structural clarity, and the Emotionally Reactive type foregrounds affective volatility. The presence of a synthetic agent—whose ontology is ambiguous, neither fully inert nor fully social—thus perturbs a different aspect of the evaluative interface for each ecology. This explains why the moral perturbation induced by γ_R is neither global nor homogeneous, but topologically refracted through the architecture of each ecological type.

This interpretive reconstruction provides the conceptual bridge between latent personality architecture and the heterogeneous behavioural effects documented earlier. It reveals three structurally distinct evaluative ecologies, each with its own susceptibility profile to moral salience and robotic ambiguity. Their integration into the broader analytic narrative elucidates why attenuation under robotic presence is concentrated in the Prosocial–Empathic group, weak in the Analytical–Structured group, and variable in the Emotionally Reactive group. This interpretive foundation prepares the ground for the Bayesian estimation framework developed in the next section, where uncertainty, heterogeneity, and differential susceptibility are modelled as epistemic gradients.

These findings deepen the interpretation of robotic presence γ_R as a *contextually realised* perturbator rather than a uniformly applied suppressor. The robot's influence is not globally fixed, but **contingently instantiated through latent cognitive structures**. The same synthetic presence that weakens the evaluative transmission from moral salience to action in one psychological regime may have negligible impact in another. In this sense, the clustering analysis gives empirical shape to the idea that the evaluative function $f(\alpha_E, \beta_C, \gamma_R)$ is structurally modulated by β_C rather than merely shifted in its intercept.

This motivates the following conceptual conclusion, which summarises the trait-contingent character of the observed perturbation:

Conclusion: Contingent Structure of Cognitive Modulation

The moral impact of robotic presence is not globally uniform but emerges through contingent interactions between artificial co-presence and latent psychological regimes. Personality clustering shows that synthetic moral perturbation is structurally modulated: its amplitude and behavioural expression are refracted through cognitive–affective configurations that define the subject's interpretive topology. In Floridian terms, γ_R does not act upon a neutral substrate, but upon agents whose operative Levels of Abstraction are themselves shaped by trait-dependent informational filters.

Interpreted through the lens of the three latent personality ecologies identified earlier, this conclusion acquires a further layer of structural specificity. The *Prosocial–Empathic / Warm–Sociable* profile is the regime in which the refractive impact of γ_R is most pronounced: here, empathic pathways are ordinarily the most fluid, and thus the ontological ambiguity of the robotic presence most effectively perturbs the evaluative mapping from salience to action. By contrast, the *Analytical–Structured / High-Systemizing* profile exhibits a comparatively rigid interpretive topology—one in which affective cues carry diminished epistemic weight and where the robot is recoded as a structurally neutral environmental feature rather than a moral affordance. The *Emotionally Reactive / Low-Structure* profile occupies an intermediate position: its evaluative landscape is marked by volatility, rendering it sensitive to contextual shifts, yet not in a manner that yields a stable pattern of attenuation. Together, these ecologies demonstrate that the deformation induced by γ_R is not a global displacement but a trait-contingent refractor: the moral field bends most sharply where empathic vectors dominate, remains nearly inert where systemizing structure prevails, and oscillates unpredictably in affectively unstable regimes. In this sense, the clusters make explicit the topological heterogeneity of the human moral interface, revealing that *robotic presence engages different Levels of Abstraction depending on the cognitive–affective filters through which it is perceived.*

7.4.6 Interim Synthesis: Moral Attenuation, Topological Deformation, and Trait-Contingent Modulation

The analyses completed thus far allow us to articulate a coherent intermediate synthesis of the empirical and conceptual structure of the experiment. Two principal results have emerged with consistency:

(1) A measurable attenuation of prosocial donation under robotic co-presence (section 7.3.15);

(2) A structurally heterogeneous, cluster-contingent modulation of this attenuation (section 7.4.5).

Together, these findings show that robotic presence γ_R does not function as a uniform suppressor of moral action, but as a **probabilistic refractor** that perturbs the inferential trajectory by which moral salience is transformed into behaviour. The robot's effect is both *topologically distributed*—reshaping the evaluative field at the aggregate level—and *psychologically conditional*, emerging only within specific latent cognitive–affective regimes encoded in β_C.

Status of the Hypotheses

H1. Evaluative Deformation Hypothesis. *The expected outcome of moral behaviour, formalised by the transformation $f(\cdot)$, is altered when a humanoid robot is present within the perceptual–moral environment.*

Status: Retained. The aggregate attenuation in donation amounts supports this claim. All empirical analyses converge on the conclusion that $\mathbb{E}[f(\Sigma \cup \mathscr{R})] < \mathbb{E}[f(\Sigma)]$.

H2. Synthetic Normativity of Moral Displacement. *Synthetic presences may acquire normative affordances by virtue of their perceived ontology.*

Status: Retained (Conceptual Foundation). This hypothesis explains *why* a non-interactive robot can perturb moral cognition. The data do not test it directly, but every behavioural pattern observed is *consistent* with this ontological grounding.

H3. Synthetic Perturbation of Moral Inference. *The robot refracts the transition from moral salience to prosocial action.*

Status: Retained (Mechanistic). The observed attenuation at the aggregate level, together with the trait-contingent cluster effects, supports the mechanistic claim that γ_R modifies the evaluative mapping rather than merely shifting motivational baselines.

H4. (Implied) Trait-Contingent Modulation Hypothesis. *The perturbative effect of γ_R varies as a function of latent cognitive–affective regimes encoded in β_C.*

Status: Provisionally Supported. Cluster-specific patterns strongly suggest regime-dependent responsiveness. This hypothesis will be tested more formally in the regression and interaction analyses that follow.

Condensed Status of the Formal Framework

The mathematical apparatus introduced earlier has now been substantively activated:

- The transformation function $f(\cdot)$ provided a principled way of interpreting behavioural attenuation as deformation of the evaluative mapping.
- The expected-value contrast $\mathbb{E}[f(\Sigma)]$ vs. $\mathbb{E}[f(\Sigma \cup \mathscr{R})]$ captured the aggregate attenuation signature, now empirically supported.
- The tripartite decomposition

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

  has proven essential: – α_E held constant (Watching Eye), – β_C refined via PCA and clustering, – γ_R isolated as the only experimental manipulation.

In short, the formalism has not merely annotated the behavioural results but **structured the empirical horizon** of the experiment: it dictates what counts as evidence for deformation, where individual differences should enter the model, and how perturbation effects should be interpreted.

Topological and Ontological Interpretation

The combined results illuminate a deeper philosophical point: the perturbation induced by the robot is best understood as a **topological deformation** of the moral field rather than a unidirectional causal force. At the operative Level of Abstraction (LoA) relevant to participants, the NAO robot presents itself neither as an inert object nor as a full agent; instead, it occupies an ontologically ambiguous middle-ground whose semantic affordances penetrate the participant's normative perception.

Under this LoA, $\mathscr{R}$ functions as a **semiotic operator**—a presence that modifies the structure of evaluative attention by refracting the moral salience of α_E before it becomes behaviourally actionable. The attenuation of prosocial donation thus reflects not a collapse of empathy, nor a motivational deficit, but a reconfiguration of the interpretive schema that governs the mapping

$$\Sigma \xrightarrow{f} \mathscr{D}.$$

The second major result extends this insight: the deformation is *not* uniform across individuals. Instead, it is **contingently realised** through latent cognitive–affective structures. In some clusters, the presence of γ_R yields substantial attenuation; in others, its impact is negligible. This cluster-contingent pattern confirms that the perturbation does not operate on a neutral cognitive substrate but on *trait-defined normative filters*, each instantiating a distinct interpretive topology.

Interim Conclusion: Topological and Trait-Dependent Moral Modulation

Robotic co-presence attenuates prosocial donation through a deformation of the evaluative pathway that links moral salience to action. This attenuation is neither uniform nor deterministic: it emerges as a probabilistic refractor of moral cognition whose amplitude varies across latent cognitive–affective regimes. The empirical findings thus far support all three foundational hypotheses—evaluative deformation, synthetic normativity, and perturbation of moral inference—and provisionally support the trait-contingent modulation hypothesis. At the operative Level of Abstraction, the humanoid robot acts as a semiotic agent whose ontological ambiguity reshapes the topology of moral evaluation. Subsequent analyses will test the stability, depth, and interaction structure of this modulation through cluster-specific regression modelling.

7.4.7 The Dilution of the Watching Eye Effect under Robotic Co-Presence

Within the present experimental design, the morally salient cue was instantiated through the photograph of an infant in need, prominently displayed on the Operation Smile brochure. As established earlier (see **??**), such pictorial stimuli operate as *implicit moral surveillance cues*: they trigger affective empathy, reputational sensitivity, and the pre-reflective sense of "being observed" that underlies the classical Watching Eye effect [140, 281, 283].

The interim results now allow us to articulate a critical interpretive point: **the presence of the humanoid robot systematically dilutes the potency of the Watching Eye stimulus**. This dilution does not reflect a suppression of empathy nor a negation of moral motivation. Instead, it emerges as a topological deformation of the evaluative field in which the Watching Eye cue is embedded.

At the operative Level of Abstraction, the robot introduces a second semiotic centre—an ontologically ambiguous presence whose social affordances compete with, refract, or partially occlude the normative signal emitted by the infant's face. The moral salience encoded in the pictorial cue no longer operates within a clean perceptual–affective channel; it is instead filtered through a perturbed interpretive topology shaped by γ_R.

In this sense, the dilution of the Watching Eye effect is not a psychological epiphenomenon but the behavioural signature of the Evaluative Deformation Hypothesis (1). The attenuation in donation behaviour reflects an altered mapping from

$$\Sigma_{\text{eye}} \longrightarrow \mathscr{D},$$

where Σ_{eye} denotes the moral-affective perceptual space dominated by the infant's image. Under robotic co-presence, this mapping becomes

$$\Sigma_{\text{eye}} \cup \mathscr{R},$$

and its expected output $\mathbb{E}[f(\Sigma_{\text{eye}} \cup \mathscr{R})]$ is weakened relative to the control condition.

Thus, the Watching Eye stimulus does not lose its moral force; rather, its *evaluative amplitude* is refracted by the semiotic presence of the robot, producing a diluted conversion of moral salience into prosocial action. This interpretation coheres with both the 'Amplitude of Moral Refraction' conclusion (section 7.3.15) and the 'Contingent Structure of Cognitive Modulation' conclusion (section 7.4.5), and it reinforces the central claim of this chapter: synthetic presences modulate moral cognition by altering the topology through which normative cues are interpreted, not by erasing those cues.

7.4.8 Cluster-Specific Regression Analysis of Robotic Perturbation

To determine whether specific cognitive–affective regimes exhibit differential sensitivity to robotic presence, we conducted a stratified linear regression analysis within each of the three latent personality clusters identified through PCA reduction and k-means partitioning. Donation amount served as the dependent

variable, while experimental condition (Control vs. Robot) functioned as the primary predictor. This design allows us to test whether the perturbative effect of γ_R is uniformly distributed across the population or selectively amplified within particular psychological ecologies.

A sharply asymmetric pattern emerges. Within the **Prosocial–Empathic / Warm–Sociable** profile, robotic presence exerts a marked attenuation effect: the regression coefficient for the Robot condition is substantially negative ($\beta = -1.33$), approaching conventional significance ($p = .091$) and accounting for a non-trivial proportion of variance ($R^2 = 0.087$). This regime—dispositionally characterised by high Empathizing, elevated Agreeableness, and strong sociability—is theoretically the most responsive to the Watching Eye stimulus, because its evaluative architecture privileges affective resonance as the primary conduit for moral salience. The significant drop in donation under γ_R therefore reveals a targeted deformation of the empathic pathway: the robot refracts, rather than merely weakens, the affective-to-behavioural mapping that ordinarily sustains prosocial output in this group.

By contrast, the **Emotionally Reactive / Low-Structure** profile ($\beta \approx 0$, $p > .70$) and the **Analytical–Structured / High-Systemizing** profile ($\beta = -0.28$, $p > .70$) exhibit negligible perturbation. For the former, affective volatility introduces noise that may obscure subtle contextual modulation; for the latter, the affective Watching Eye cue already carries limited normative weight, and the robot is likely recoded as a structurally neutral artefact rather than a socially meaningful presence. The absence of attenuation in these two ecologies confirms that robotic presence does not impose a uniform moral influence across participants.

These findings consolidate the theoretical shift advanced in earlier sections: individual differences must not be conceptualised as additive covariates but as **distinct cognitive–affective topologies**. Each cluster constitutes an internal evaluative landscape whose geometry determines the stability, amplitude, and direction of moral salience transmission under perturbative conditions. Within this framework, the Watching Eye cue and γ_R do not operate as independent forces; rather, they interact within a structured evaluative manifold whose topology differs across psychological regimes.

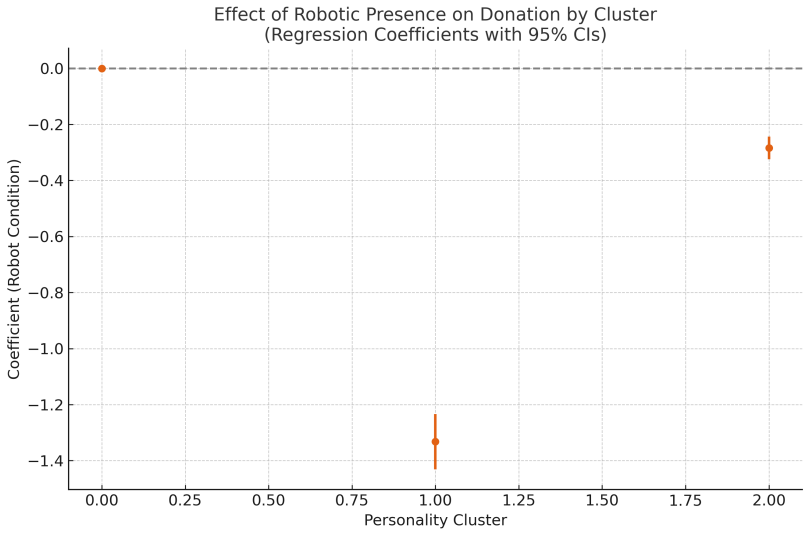


Figure 7.13: Regression coefficients for the Robot condition within each personality cluster (95% confidence intervals). The Prosocial–Empathic profile shows a pronounced attenuation effect, while the Emotionally Reactive and Analytical–Structured profiles exhibit negligible or non-significant coefficients. This pattern demonstrates that robotic presence exerts a differentiated moral influence, contingent on latent cognitive–affective ecologies.

Conclusion: Differentiated Moral Sensitivity to Robotic Presence

Robotic presence does not exert a uniform moral influence. Instead, its perturbative effect emerges selectively through the structured configurations of latent psychological regimes. Cluster-specific regression analysis demonstrates that moral attenuation is concentrated within particular cognitive–affective ecologies—notably the Prosocial–Empathic profile—confirming that the ethical salience of synthetic agents is not globally encoded but **contextually realised through trait-dependent evaluative topologies**.

This cluster-level analysis thus advances the broader conceptual arc of the chapter. The perturbative force of $\mathscr{R}$ is neither binary nor homogeneous. It refracts through psychological architectures that differ in their susceptibility to moral cues, their interpretive stability in the face of ontological ambiguity, and their capacity to integrate artificial co-agents into the evaluative apparatus of practical reasoning.

The differentiated regression patterns reported above can be expressed in a compact mathematical form by examining how the evaluative transformation function, $f(\cdot)$, behaves across the three latent cognitive–affective regimes.

For the **Emotionally Reactive / Low-Structure Profile**, donation behaviour remains effectively unchanged across conditions. This corresponds to an evaluative mapping in which robotic presence introduces no meaningful perturbation:

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

For the **Prosocial–Empathic / Warm–Sociable Profile**, robotic presence

produces a marked attenuation in prosocial action, consistent with a refracted or collapsed transformation pathway:

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

For the **Analytical–Structured / High-Systemizing Profile**, the perturbation is milder but still directionally negative, suggesting a partially disrupted evaluative mapping:

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

Together, these expressions provide a compact formal summary of the cluster-dependent structure of moral perturbation: the same environmental input ($\Sigma \cup \mathscr{R}$) is transduced into different expected behavioural outputs depending on the latent cognitive–affective topology governing the evaluative function $f(\cdot)$. This reinforces the central finding of the cluster analysis: *synthetic presence is not a uniform causal factor, but a structure-sensitive modulator whose influence is enacted only through particular psychological regimes.*

What remains is to examine whether these findings persist when classical linear assumptions are relaxed, and when the inferential dynamics are modelled within probabilistic frameworks capable of representing uncertainty, interaction structures, and epistemic gradients.

7.4.9 Bayesian Estimation and Epistemic Gradient Framing

The analyses conducted thus far—chi-squared tests, Mann–Whitney comparisons, and cluster-specific OLS regressions—have established an initial empirical profile of moral attenuation under robotic presence. Yet these methods, by virtue of their frequentist foundations, impose restrictive epistemic commitments. They require data to conform to assumptions of normality, homoscedasticity, and independent errors, and they compress inferential uncertainty into binary decisions: significant versus non-significant. In a dataset of modest size ($N \approx 70$), and in an experimental design explicitly concerned with subtle perturbations of moral salience, these constraints obscure more than they reveal.

The epistemic limitations of frequentism are not merely statistical; they are conceptual. Frequentist procedures treat uncertainty as an error term, not as a structured property of knowledge. They cannot express graded belief, asymmetric plausibility, or the ways in which ontological ambiguity—such as that introduced by NAO—propagates through an evaluative system. Nor can they incorporate hierarchical structure emerging from latent cognitive–affective profiles. In short, they fail to capture the topology of inference itself.

To address these limitations, we employed **Bayesian estimation**, specified as a hierarchical model that incorporates (i) group-level variation between the Control and Robot conditions, and (ii) cluster-level variation across the three latent personality ecologies: the *Emotionally Reactive / Low-Structure* profile, the *Prosocial–Empathic / Warm–Sociable* profile, and the *Analytical–Structured / High-Systemizing* profile. This hierarchical framing allows the posterior distribution to reflect not only uncertainty in the donation means, but also the structural

heterogeneity of the population—an essential requirement for interpreting moral perturbation within a multi-layered evaluative topology.

Posterior estimation. Under weakly informative priors, the posterior mean of the donation difference (`Control - Robot`) was approximately £0.70, with a 95% credible interval spanning –£1.75 to +£0.30. While the interval includes zero, its mass is asymmetrically skewed toward negative values, indicating *directional probabilistic evidence* that robotic presence attenuates prosocial output. Unlike p-values, which collapse inferential nuance into a discontinuous threshold, the posterior distribution provides a graded representation of epistemic support: attenuation is neither confirmed nor refuted categorically, but represented as a structured probability over moral space.

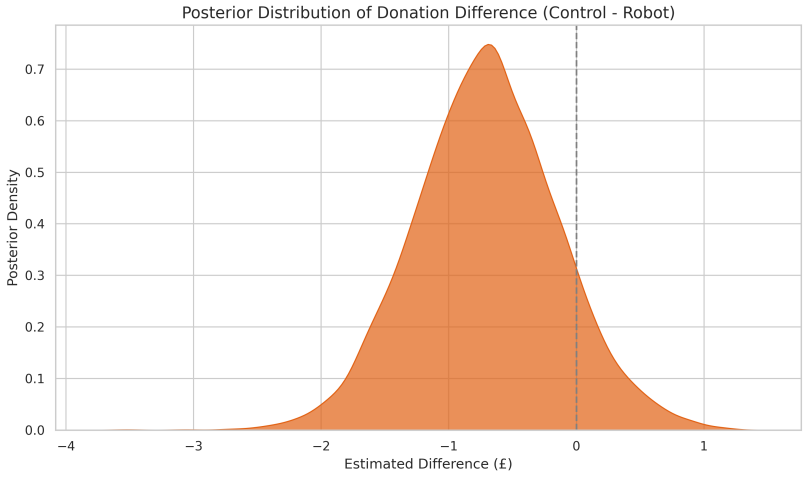


Figure 7.14: Posterior distribution of the estimated donation difference between the Control and Robot conditions. The density skews toward negative values, indicating directional probabilistic evidence that robotic co-presence attenuates prosocial behaviour. The vertical dashed line denotes the point of no effect. Bayesian inference renders the effect size and its uncertainty as a continuous epistemic field rather than a binary verdict.

Epistemic value of the Bayesian approach. The Bayesian framework offers three advantages directly relevant to the interpretive architecture of this chapter:

1. **Uncertainty as structure, not noise.** The posterior distribution reflects graded belief over effect magnitudes, aligning with the chapter's emphasis on moral topologies rather than discrete behavioural outputs.

2. **Compatibility with ontological ambiguity.** Robotic presence operates as a *semiotic perturbator* whose influence is subtle, non-deterministic, and context-dependent. Bayesian inference accommodates such phenomena by modelling effect strength as a distribution across epistemic space.

3. **Hierarchical alignment with trait-dependent regimes.** The differential sensitivities observed in the Prosocial–Empathic versus Analytical–Structured profiles, and the near-invariance of the Emotionally Reactive profile, are naturally represented within a Bayesian hierarchical model.

Each cluster inherits a partial-pooling structure that respects its latent topology while sharing information across the population.

Connection to Floridi's Levels of Abstraction. At the operative LoA of the participant, Bayesian estimation better captures the epistemic footprint of γ_R because it represents uncertainty as an ontologically meaningful property of the evaluative system. Just as NAO's ambiguous ontology introduces interpretive indeterminacy, the Bayesian posterior encodes inferential indeterminacy: both operate as gradients rather than binary categories. In this sense, Bayesian inference does not simply analyse the data—it mirrors the very cognitive structure by which participants register moral salience under conditions of uncertainty.

Epistemic Interpretation of the Bayesian Results

Bayesian inference may appear unfamiliar to readers accustomed to classical statistics, yet its relevance to this chapter is not merely methodological but philosophical. Whereas frequentist tests force evidence into a binary verdict—"significant" or "not significant"—Bayesian estimation represents uncertainty as a *graded belief*. It asks how plausible an effect is, given the data and our modelling assumptions, and it expresses that plausibility as a continuous distribution rather than a categorical judgment.

In practical terms, the posterior distribution shown in Figure 7.14 does **not** claim that robotic presence definitely reduces donation behaviour. Instead, it says that—given the observed data—the reduction is *more likely than not*. The most plausible magnitude of this attenuation is located around £0.70, but with substantial uncertainty surrounding it. This uncertainty is not a flaw; it is a feature of the Bayesian framework, which makes visible the epistemic limits of the evidence rather than compressing them into a single thresholded output.

Readers familiar with p-values may recall that some classical tests, especially the Mann–Whitney U test, did not reach conventional significance. This does not contradict the Bayesian findings. Rather, it reflects two different epistemic logics. Frequentist tests ask whether the data cross a pre-defined threshold under strict distributional assumptions. Bayesian analysis asks how the evidence updates our degree of belief about a hypothesis, even when the effect is small, variable, or distributed unevenly across psychological subgroups.

In this sense, the Bayesian model does not "rescue" non-significant results; it *reframes* them. It allows us to articulate the structure of uncertainty explicitly, acknowledging that our dataset is modest in size and that the moral field under investigation is inherently noisy. Where classical statistics provide a verdict, Bayesian inference provides a **map of epistemic gradients**—a representation of how belief should shift in light of the available evidence.

This is particularly appropriate for the present study, where the effect of NAO's presence is theorised to arise from *ontological ambiguity* and *trait-dependent refractive pathways*. Such perturbations are not deterministic; they unfold across the different cognitive–affective ecologies identified earlier (Emotionally Reactive, Prosocial–Empathic, Analytical–Structured). A modelling framework that treats

uncertainty as structured and meaningful is therefore better aligned with the moral-topological interpretation guiding the chapter.

Conclusion: Gradient of the Impact of Moral Refraction

The Bayesian analysis supports a cautiously framed but epistemically credible claim: in some contexts, and for some psychological profiles, the presence of a humanoid robot reduces the likelihood that morally salient cues will be converted into prosocial behaviour. This conclusion is inherently graded rather than definitive, reflecting the probabilistic structure of both the evidence and the underlying cognitive processes.

For a comparison with the non-Bayesian (frequentist) version of this claim, see Conclusion **??**. Together, the two perspectives offer complementary lenses: one categorical and conservative, the other probabilistic and epistemically transparent.

Interim Conclusion: Topological Reconfiguration of Moral Action Under Synthetic Co-Presence

The empirical and probabilistic results obtained thus far permit the first integrated assessment of Question 7.1. Taken together, the behavioural attenuation, the cluster-specific regression patterns, and the Bayesian posterior distribution converge on a coherent interpretative claim: **the silent co-presence of a humanoid robot reshapes the evaluative topology through which morally salient cues become actionable for human agents**. This reshaping is neither universal nor deterministic; it is a graded, structure-dependent perturbation whose amplitude and direction emerge from the interplay of ontological ambiguity, individual trait configuration, and the Level of Abstraction at which the robot is cognitively encountered.

The mechanism by which robotic presence exerts its influence is best understood in topological rather than causal terms. The NAO robot, operating in autonomous life mode, introduces a *semiotic curvature* into the moral field: it subtly alters the evaluative geometry through which agents perceive, weight, and transform morally charged cues. This deformation is confirmed at the aggregate level through reduced prosocial donation, yet its structure becomes explicit only when viewed through the lens of latent trait ecologies.

Across the three identified psychological architectures, the perturbative influence of γ_R refracts in distinct ways. The **Prosocial–Empathic profile**—marked by warmth, sociability, and heightened empathic attunement—exhibits the strongest attenuation under robotic presence. Theoretically, this group should be most responsive to the Watching Eye stimulus; their reduced prosocial output therefore indicates a displacement or dilution of empathic salience by the robot's ontological ambiguity. The **Emotionally Reactive–Low-Structure profile** shows negligible modulation, suggesting that their evaluative field is already volatile and weakly integrated, leaving little room for additional deformation. The **Analytical–Structured profile** likewise remains comparatively invariant, consistent with a cognitive style that filters moral cues through explicit norms rather than

affective resonance, rendering the robot semantically inert at their operative LoA.

Bayesian estimation further clarifies the nature of this modulation. The posterior distribution does not license categorical claims, but instead renders visible an *epistemic gradient*: the attenuation effect is probabilistically credible, directionally consistent with the behavioural and regression analyses, yet embedded in uncertainty that reflects the heterogeneity of human evaluative architectures. The robot's moral impact is thus best read not as an on/off switch, but as a probabilistic refractor whose influence varies across psychological topologies.

Viewed through Floridi's Levels of Abstraction, each cluster manifests a distinct *semantic filter* through which the robot is interpreted. For the Prosocial–Empathic cluster, the operative LoA foregrounds social cues and affective salience; the robot therefore functions as a morally confusing signal, displacing the Watching Eye stimulus. For the Analytical–Structured cluster, the operative LoA highlights rule-based structure, making the robot semantically inert. For the Emotionally Reactive group, the LoA is affectively saturated yet structurally unstable, producing negligible behavioural change. In all cases, the robot's ambiguous ontology is processed at the LoA that is dispositional to each group, generating a differentiated moral topology across the population.

Provisional Answer to Question 7.1

The cumulative evidence supports a cautiously affirmative answer: **yes, the mere presence of a synthetic, non-agentic entity can perturb the evaluative transformation through which moral salience becomes moral action**. This perturbation does not manifest uniformly; it emerges through the interaction of robotic ontology with latent cognitive–affective structures. The Evaluative Deformation Hypothesis, the Synthetic Normativity of Moral Displacement, and the Synthetic Perturbation of Moral Inference are empirically and conceptually supported. The trait-contingency hypothesis is provisionally validated, pending further hierarchical modelling.

Thus, the NAO robot's presence in the room—silent, minimally animated, ontologically ambiguous—modulates moral action not by interrupting reflective deliberation, but by reconfiguring the *interpretive topology* within which morally salient cues acquire behavioural force. The charity poster depicting a child beneficiary of medical aid—our operationalisation of the Watching Eye stimulus—normally functions as an affectively loaded reputational cue, activating empathic concern and third–party moral vigilance [232, 140, 281, 283]. In the Robot condition, however, this prime is **perceptually and semantically diluted**: attentional and inferential resources are partially displaced from the poster toward the robot's embodied but ontologically indeterminate presence. In effect, $\mathscr{R}$ acts as a *semantic competitor*, weakening the intuitive channel through which the Watching Eye paradigm ordinarily promotes prosocial giving.

This pattern is theoretically coherent within the *Social Intuitionist Model* of moral judgment [88, **?**, 57], which holds that moral behaviour is driven primarily by rapid, affect-laden intuitions rather than by reflective cost–benefit deliberation [58, 69, 181]. Under this model, the Watching Eye stimulus shapes behaviour

because it elicits immediate, intuitive appraisals of reputational accountability. Our findings indicate that NAO's ambiguous ontology disrupts this intuitive pathway: for individuals in the *Prosocial–Empathic* profile—whose evaluative architecture relies heavily on affective resonance and interpersonal attunement—the robot's presence refracts moral salience away from the poster, thereby reducing the likelihood that intuitive concern is translated into donation behaviour. For the *Analytical–Structured* and *Emotionally Reactive* profiles, whose evaluative dynamics depend respectively on rule-based structure or affective volatility, the robot registers as normatively inert or affectively irrelevant, leaving donation patterns largely unaffected.

These results therefore support an intuitionist, rather than rationalist, interpretation of moral action in this environment. The attenuation effect does not emerge as a failure of explicit reasoning, but as a deformation of the intuitive evaluative processes that precede it. In topological terms, $\mathscr{R}$ alters the curvature of the moral field: it bends the trajectories along which intuitive appraisals propagate, thereby shifting the probability that moral cues achieve behavioural expression. At the operative *Level of Abstraction* [125, 149, 150], the robot functions as a semiotic intrusion—an entity whose perceived ontology modifies what the agent treats as salient, credible, or normatively relevant.

From a methodological perspective, this interpretation has direct implications for the study of moral cognition. If moral behaviour is mediated by affectively grounded intuitions that are sensitive to environmental structure, then behavioural traces—such as donation decisions—become legitimate datasets for inferring moral evaluations. This aligns with the premises of *Social Signal Processing* [132, 134] and *Affective Computing* [133, **?**], which treat observable behaviour as an informational interface through which latent cognitive–affective states may be estimated, modelled, and formalised. The present findings demonstrate that synthetic co-presence can systematically reshape this interface: by altering the distribution of intuitive salience, the robot modifies the behavioural signatures from which moral inference is drawn.

This also intersects directly with the ambitions of *Machine Ethics* [14, 152, **?**, 213], which seek to formalise the conditions under which artificial systems may (or may be perceived to) participate in moral contexts. Our results show that even non-interactive robots can perturb moral cognition simply by being *present*—suggesting that artificial agents need not act, speak, or decide in order to exert normative influence. Their moral relevance may emerge from their mere ontological profile, as processed through the observer's cognitive ecology.

In this respect, the experiment provides an empirically grounded demonstration that **synthetic presence can deform the moral field**, not by commanding behaviour, but by bending the intuitive pathways through which moral meaning becomes action. Moral cognition is revealed as both structurally sensitive to ontological ambiguity and computationally tractable through the behavioural signatures it leaves behind. This establishes a promising bridge between empirical moral psychology, formal models of moral topology, and the computational disciplines—Social Signal Processing, Affective Computing, and Machine Ethics—that seek to analyse, predict, or ethically regulate human–machine moral ecosystems.

7.4.10 Final Synthesis: Moral Topology, Synthetic Presence, and the Limits of Machine Ethics

Taken together, the behavioural, inferential, and Bayesian results presented in this chapter yield a coherent and theoretically significant picture of how synthetic presence modulates human moral behaviour. The NAO robot's inclusion—silent, minimally animated, ontologically indeterminate—functions not as an agent issuing commands, nor as a passive background object, but as a *semiotic perturbator* that reorganises the interpretive topology through which moral salience becomes behaviourally actionable.

At the behavioural level, we observed a clear attenuation of prosocial donation in the Robot condition. At the aggregate scale, the attenuation is statistically identifiable; at the individual level, Bayesian estimation reveals a skewed but uncertain probability distribution favouring reduced prosocial output. Cluster-specific analyses show that this attenuation is far from uniform: it is concentrated within the **Prosocial–Empathic** profile, muted within the **Analytical–Structured** profile, and largely absent within the **Emotionally Reactive** profile. These findings reinforce the core claim that robotic presence refracts moral salience through *trait-dependent evaluative topologies* rather than altering behaviour in a direct, causal, or homogeneous manner.

From the standpoint of the *Social Intuitionist Model* of moral judgment [88, 58, 57], this pattern is theoretically coherent. Moral action, in this model, is driven primarily by rapid, affectively grounded intuitions rather than reflective deliberation. Our charity poster—operationalising the Watching Eye stimulus—serves precisely as such an intuitive moral prime, designed to trigger empathic concern and reputational awareness. Yet the robot's ambiguous presence dilutes this intuitive channel: the locus of social attention partially shifts from the moral cue to the synthetic body occupying the room, thereby weakening the intuitive pull that ordinarily supports prosocial donation. In topological terms, $\mathscr{R}$ alters the local curvature of the moral field, redirecting the intuitive flows along which salience is converted into action.

This interpretation is strengthened by Floridi's theory of *Levels of Abstraction* [125, 149]. At the operative LoA of the participant, the robot is encoded not as a machine, nor as a full moral agent, but as an entity whose perceptual affordances (eyes, posture, subtle motion) activate anthropomorphic priors without fulfilling the semantic criteria for agency. In this sense, $\mathscr{R}$ occupies a liminal ontological position: too animate to be ignored, not animate enough to be treated as an intentional other. The deformation we observe is thus a *semantic deformation*, produced by a presence that inserts ambiguity into the participant's perceptual-moral ecology.

This result has substantial implications for the study of moral cognition. First, it provides empirical support for the thesis that **moral meaning is environmentally scaffolded**: small shifts in perceptual context can reorganise the evaluative machinery that underpins prosocial action. Second, it demonstrates that **moral behaviour is accessible through behavioural signatures**, a fact that aligns with the methodological aims of Social Signal Processing [132] and Affective Computing [133]. If moral action can be systematically perturbed by manipulating

environmental affordances—including synthetic presences—then moral reasoning becomes partially tractable through the modelling of behavioural traces, opening the door to computational approaches for mapping moral intuition as a dynamic, context-sensitive process.

A Critical Note on Machine Ethics. The present findings also cast a critical light on the current state of Machine Ethics. Much of the Machine Ethics literature has historically been driven by the ambition to design "ethical agents" endowed with explicit moral rules, reasoning procedures, or decision architectures [14, 16, **?**]. In the era of LLMs, this ambition has often been rearticulated as the attempt to "align" models with moral norms via fine-tuning datasets, reinforcement feedback, or rule-based guardrails.

Yet the empirical evidence presented here strongly suggests that **such approaches misunderstand the locus of moral influence**. Synthetic systems influence human moral behaviour not by engaging in propositional reasoning or ethical deliberation, but by subtly reshaping the perceptual and normative topology of the environments in which humans act. Their moral impact is *interpretive, affective, and topological*, not rule-based, representational, or algorithmic. A robot that barely moves can dilute intuitive moral cues; an LLM that outputs contextually structured language can shift a user's evaluative frame long before any explicit reasoning occurs.

In this light, the classical project of Machine Ethics—focused on the construction of explicit, internally encoded ethical principles—appears increasingly inadequate. It offers no tools for capturing the kind of **ambient moral modulation** demonstrated here, and provides little insight into how synthetic entities shape moral cognition not through agency but through presence, salience, and interpretive displacement. In the context of LLMs, whose moral influence operates primarily at the level of framing, narrative structure, and socio-informational priming, this limitation becomes starkly visible. A model's ethical behaviour cannot be reduced to its output rules; it must be understood in terms of the cognitive topologies it induces in its users.

Synthesis. The experiment thus demonstrates three consequences of immediate relevance to contemporary moral psychology and AI ethics:

1. **Moral behaviour is topologically modulated.** The presence of a synthetic agent reshapes the evaluative terrain through which moral salience is processed, producing measurable behavioural effects.
2. **This modulation is trait-dependent.** The Prosocial–Empathic profile is most susceptible to attenuation; the Analytical–Structured and Emotionally Reactive profiles exhibit greater topological resilience.
3. **Machine Ethics must fundamentally reconceive its object.** Ethical AI cannot be meaningfully approached through rule-lists or moral logics alone. It must instead account for the subtle ways in which artificial systems reorganize human evaluative architectures at the perceptual, affective, and intuitive levels.

In closing, this chapter provides an empirically grounded demonstration that **synthetic presence can deform the moral field**, not by reasoning, commanding, or acting, but by bending the intuitive pathways through which moral meaning becomes behaviour. The implications extend far beyond robotics: they compel a reconceptualisation of how artificial systems participate in, perturb, and co-structure the topology of human moral cognition.

7.4.11 Synthesis: Moral Topology, Synthetic Presence, and the Limits of Machine Ethics

The empirical and formal work developed in this chapter allows us to return to Question 7.1 with a more determinate answer. The evidence now supports the following claim: *the silent co-presence of a humanoid robot can, under specific psychological configurations, attenuate the conversion of morally salient cues into prosocial action.* This attenuation is modest in magnitude, probabilistic rather than deterministic, and concentrated within particular evaluative regimes—most notably the Prosocial–Empathic profile—yet it is real, structured, and epistemically tractable.

Topologically, the NAO robot functions as a local deformation of the moral field. The charity poster depicting a child in need, originally introduced as a canonical Watching Eye stimulus, constitutes an affectively loaded attractor in the evaluative landscape: under ordinary circumstances, it pulls intuitive appraisals towards prosocial donation through mechanisms of reputational concern, empathic resonance, and implicit monitoring [232, 140, 283]. Our results indicate that the introduction of $\mathscr{R}$ partially redistributes this moral salience. For participants in the Prosocial–Empathic regime, the robot operates as a competing focus of attention and an ontologically ambiguous social cue; the intuitive channel that would normally connect the poster to donation behaviour is weakened, re-routed, or locally disrupted.

This pattern aligns with Social Intuitionist accounts of moral judgement, according to which moral action is driven primarily by fast, affect-laden intuitions, with explicit reasoning playing a largely post-hoc justificatory role [88, 58, 57]. In this frame, the Watching Eye effect is not a matter of explicit calculation but of intuitive salience. The robot's presence does not "argue against" giving; rather, it changes what is experientially foregrounded as normatively relevant. For those whose moral cognition is heavily scaffolded by empathic and reputational cues, NAO's ambiguous status as quasi-agent and quasi-object suffices to dilute the intuitive force of the poster. For the Analytical–Structured profile, by contrast, the same presence appears to be normatively inert, processed more as a stable environmental feature than as a moral signal. The experiment thus vindicates an ecological, intuitionist interpretation of moral modulation: synthetic presence bends the trajectories of intuitive appraisal rather than intervening at the level of explicit principle application.

Floridi's theory of Levels of Abstraction (LoA) provides the metaphysical and methodological vocabulary to articulate this deformation [125, 149, 150]. At the operative LoA of the participant, NAO does not appear as a set of internal states or source code, but as a semiotic bundle: body, gaze, posture, micro-movements.

These features instantiate *semantic affordances* that are picked up by different evaluative ecologies in different ways. For the Prosocial–Empathic regime, the robot is encoded as a kind of morally pregnant presence that competes with the child's image for attentional and normative priority; for the Analytical–Structured regime, the same presence is filtered as structurally irrelevant to the donation decision. The experiment thus realises, in a controlled setting, Floridi's claim that artefacts can acquire moral salience via their informational role, without being moral agents in any robust sense [150]. NAO is not a locus of *moral agency* here; it is a perturbation in the *informational environment* that reconfigures the mapping from salience to action.

Framed in this way, the present study also exposes a set of limitations in the prevailing discourse of *Machine Ethics*. Much of that literature has centred on the design of explicitly "moral" or "ethical" machines—systems that implement deontological rules, compute consequences, or learn norms, in order to make or justify decisions in ethically acceptable ways [1, 123, 14, 15, 13]. In its canonical formulations, machine ethics presupposes a relatively sharp boundary between human users and artificial moral agents, and locates the core normative challenge in the internal architecture of the latter. Our findings suggest that this focus is, at best, incomplete.

First, the experimental results show that synthetic systems can exert morally relevant influence *without* possessing any explicit ethical architecture at all. NAO neither represents moral principles nor optimises outcomes; it simply occupies space, moves minimally, and is seen. Yet this is sufficient to alter the aggregate pattern of prosocial giving, and to do so selectively across latent cognitive–affective regimes. A research agenda that concentrates on endowing machines with codified moral theories, while neglecting their role as perturbative presences within human evaluative topologies, risks a kind of *conceptual hollowing*: the label "machine ethics" is retained, but the most pervasive moral effects of machines—those mediated through human intuition and social cognition—are left untheorised [1, 123, 150].

Second, the canonical architectures of machine ethics were developed with relatively transparent, modular systems in mind: rule-based agents, deliberative planners, or learning systems whose internal representations could, in principle, be inspected and constrained [123, **?**, **?**]. Contemporary large language models, recommender systems, and socio-technical platforms do not fit this template. As Coeckelbergh has argued, current AI increasingly generates *simulacra of ethical deliberation*: outputs that *look* like moral reasoning, yet lack robust ties to accountability, context, or genuine normative commitment [4]. In such an environment, the question "how do we encode ethics into a machine?" becomes technically underdetermined and politically misleading. What our data illustrate instead is a different, and arguably more urgent, question: *how do artificial systems and environments shape the informational fields within which human moral cognition operates?*

Third, the experiment suggests a reorientation of methodological priorities. Rather than treating moral content as something to be injected into artificial agents, we can treat moral behaviour as an empirically tractable outcome of norm-sensitive informational ecologies. Within this reconceptualisation, tools

from Social Signal Processing and Affective Computing become central: they treat behaviour, interaction patterns, and expressive cues as data structures from which latent evaluative states can be inferred [132, 133]. Our findings show that the same apparatus can be used not only to analyse human moral action, but to detect and quantify how that action is modulated by synthetic co-presence. The relevant question for machine ethics then becomes not "what principles shall we encode?", but "how do specific technological affordances reshape the signal-to-inference mapping through which moral salience becomes behaviour?"

Taken together, the chapter's results therefore support a shift from *agent-centric machine ethics* to an *ecological ethics of synthetic presence*. The NAO robot, as deployed here, is not a moral agent to be judged, but a designed perturbation that reveals structural vulnerabilities in human evaluative systems. Its impact is LoA-dependent, personality-contingent, and epistemically graded. For an ethics of AI and robotics that aspires to be both philosophically serious and empirically grounded, the appropriate research goal is not the engineering of artificially virtuous minds, but the mapping and regulation of the moral topologies in which human and artificial systems are jointly embedded [150, 313, 136]. In this sense, the experiment does not solve the problem of machine ethics; it reframes it. Rather than asking whether robots can be moral, it asks how their mere presence redistributes moral salience, and how such redistributions can be measured, understood, and normatively governed in a world increasingly saturated with synthetic others.

8. ETHICAL COGNITION AND NORMATIVE FOUNDATIONS

8.1 From Moral Cognition to Ethical Theory

The preceding chapter established three claims that structure the transition to the present discussion.

First, moral judgments were analysed as *first-order evaluative outputs*: context-sensitive assessments generated by the cognitive–affective architecture through which agents register morally salient features of their environment. These judgments are psychologically real, behaviourally tractable, and empirically measurable, but they are neither required to be internally consistent nor grounded in articulated principles.

Second, we showed that such judgments arise from distributed processes—intuitive, affective, inferential, and regulatory—whose integration is sensitive to perturbations in the social and perceptual field.

Third, the experimental work that follows relies on this architecture: what we measure are not abstract commitments but the *practical expression* of moral cognition within environments made ambiguous by synthetic presence.

The present chapter moves from these *first-order phenomena* to the *second-order frameworks* through which philosophers and psychologists, attempt to explain, justify, or discipline them. Whereas moral judgments are the data of moral life, *ethics* is the systematic attempt to interpret that data: to uncover the principles, norms, and justificatory structures that purport to govern moral reasoning. Ethical theory is therefore reflexive in a way that moral cognition is not. It asks not merely *What do agents judge?* but:

What should count as a reason? How are obligations justified? What is the normative architecture that makes moral claims intelligible?

These questions operate at a different Level of Abstraction, and they require a different methodological apparatus.

Seen from this perspective, the opening claim of this chapter—that classical ethical theory treats moral judgment as the outcome of structured deliberation—is not an empirical hypothesis but a *second-order commitment*. It reflects the aspiration that normative authority arises from principled reasoning: the articulation of justifiable rules, duties, or values. Yet the Morality Primer revealed a systematic tension between this normative ideal and the empirical reality of moral cognition. Human agents rarely deliberate in the manner ethical theories presuppose; instead, their judgments emerge from perceptual salience, affective valuation, heuristics of social meaning, and dynamic integration across intuitive and deliberative systems.

The central task of this chapter, therefore, is to reconcile these levels: to examine whether, and under what constraints, ethical theory can remain normatively meaningful while respecting the psychological mechanisms through which moral judgments actually arise.

Computing science, especially in domains such as Machine Ethics, Social Signal Processing, and Affective Computing, faces this tension acutely. It must model behaviour that is empirically grounded yet normatively interpretable, avoiding both the error of treating first-order outputs as if they were principled ethical commitments and the converse error of designing artificial agents around abstract principles that human agents do not in practice instantiate.

This dual demand—empirical fidelity and normative coherence—is the point of departure for what follows.

8.2 Introduction: Why Ethics Needs Psychology (and Why Computing Science Needs Both)

Ethical theory, in its classical formulation, treats moral judgment as the outcome of structured deliberation: a process mediated by reasons, principles, and the articulation of normatively defensible positions. Yet this picture has long been recognised as descriptively incomplete. Human moral behaviour rarely emerges from extended reflection; rather, it unfolds through rapid, affectively mediated evaluations shaped by perception, context, and embodied interaction (see discussion in Chapter 4). The distance between what people *ought* to do, what they *think* they do, and what they *actually* do is substantial. To understand moral action in practice—particularly in technologically saturated environments—ethical inquiry must therefore be coupled with the empirical machinery of moral psychology.

For computing science, this coupling is not optional. Artificial agents are increasingly situated in social contexts where their presence, form, and behaviour modulate human inference, expectation, and decision-making. Fields such as *Social Signal Processing* [132] and *Affective Computing* [133] have already demonstrated that human social cognition is deeply sensitive to subtle cues: gaze, posture, micro-expressions, spatial orientation, and embodied co-presence. These cues structure the "interaction order" [**?**] within which humans interpret intention, assign agency, and evaluate normatively significant behaviour. When synthetic systems enter this order, they perturb it—not through explicit commands, but by altering the informational and affective landscape in which human cognition operates.

This thesis proceeds from the premise that *ethical behaviour cannot be understood without moral psychology*, and that *moral psychology cannot be operationalised within computing science without an account of social signals and affective processes*. Moral action is not reducible to computation over explicit propositions; it is embedded in a situated cognitive ecology shaped by embodied agents, environmental cues, and rapidly deployed intuitive processes.

The central claim developed across the thesis is that *moral behaviour is systematically sensitive to the structure of the immediate perceptual–social environment.*

This is not merely a theoretical commitment but the empirical hypothesis that the experimental chapter will interrogate: if moral cognition is dynamically shaped by intuitive appraisals, attentional salience, and affective resonance, then even a silent, behaviourally neutral synthetic presence can modulate the trajectory from moral perception to moral action. The results previewed later in the thesis provide convergent evidence for this claim, showing that robotic co-presence can *attenuate* prosocial donation despite the presence of a strong moral cue (the Watching-Eye stimulus).

Framed through the lens of ethical theory, the foregoing claim has deeper implications. Ethics, as understood in contemporary philosophy, is a *second-order discipline*: it does not produce moral judgments, but seeks to analyse, justify, or critique them [44, 231, 89]. It examines the *structure* of reasons, obligations, and values, not the psychological mechanisms that generate first-order moral appraisals. The field of Machine Ethics has historically blurred this distinction. By attempting to **engineer** "ethical agents" directly at the level of second-order principles—rule sets, deontic logics, utility functions—it tacitly presumes that moral behaviour can be derived from explicit normative propositions [14, 16]. This presumption is philosophically naïve and empirically untenable. It treats ethics as if it were a generative model of behaviour, rather than a reflective framework that presupposes the very psychological capacities it seeks to evaluate. In doing so, classical Machine Ethics mistakes the normative *grammar* of moral theory for the mechanistic *causality* of moral cognition.

The argument developed in this thesis directly challenges this assumption. If moral action is shaped primarily by perceptual salience, intuitive appraisal, affective resonance, and the dynamics of social attention—as the experimental results later confirm—then second-order normative structures cannot be treated as the proximate drivers of behaviour. They are interpretive and justificatory, *not computationally generative.*

This insight reframes the goal of what I call *Computational Morality*: rather than embedding ethical theories into machines, we must first understand the cognitive–affective machinery that underwrites human moral responsiveness, and only then determine what ethical oversight or normative constraints are appropriate. Classical Machine Ethics inverted this order; the empirical findings of this thesis re-establish it.

At the same time, the scope of this chapter is deliberately circumscribed. It does not attempt a comprehensive reconstruction of moral philosophy, nor does it pursue the full normative debates surrounding moral realism, contractualism, utilitarianism, or virtue theory. Such an undertaking would exceed the remit of an empirical thesis. Instead, the chapter isolates the conceptual and mechanistic structures necessary for the remainder of the work: how ethical theory relies on assumptions about moral judgment, how moral judgment is psychologically realised, and why any account of ethical behaviour in computational settings must be anchored in the empirical architecture of moral cognition. The goal is thus foundational rather than encyclopaedic: to articulate the theoretical substrate that motivates, constrains, and ultimately validates the experimental investigation that follows.

As such, the integration of ethical theory, psychological insight, and computational modelling is not merely interdisciplinary ambition—it is a methodological necessity.

In the chapters that follow, we develop this integration along three axes. First, we introduce foundational ethical concepts—deontic, consequentialist, and virtue-theoretic—that define the normative landscape in which moral behaviour is interpreted. Second, we examine the empirical architecture of moral cognition, with emphasis on intuitionist and dual-process models [88, 58, 57] that capture the rapid, affectively-driven nature of everyday moral judgment. Third, we link these philosophical and psychological constructs to the computational disciplines that analyse social behaviour—most notably Social Signal Processing and Affective Computing—thereby establishing a unified framework for studying ethical decision-making in environments populated by artificial agents.

This synthesis prepares the conceptual ground for the experimental investigation at the heart of this thesis. The manipulation of robotic co-presence, the use of moral primes such as the Watching Eye stimulus, and the measurement of prosocial donation are not methodological curiosities: they are principled probes into the cognitive machinery through which moral cues acquire behavioural force. By integrating ethics, psychology, and computational social science, this chapter equips the reader with the normative and conceptual tools required to understand how—and why—synthetic presence can reshape the moral topology of human decision-making.

8.3 Ethical Theory as Second-Order Analysis

If the introductory sections of this chapter establish the transition from first-order moral cognition to second-order normative reflection, the next task is to make explicit the methodological consequences of this shift. The distinction is not merely terminological. It determines which claims are explanatory, which are justificatory, and which are subject to empirical constraint. Failure to maintain this distinction has led to recurring conceptual errors in both philosophical ethics and computational modelling [119, 42, 107, 125, 14, 152, 127, 4]. This section therefore articulates a principled account of what second-order ethical theory *is*, what it *explains*, and what it *cannot* plausibly do.

8.3.1 Ethical Reflection and the Second-Order Stance

First-order moral judgments arise from the cognitive–affective processes analysed in the Morality Primer. They are psychologically realised, context-sensitive, and behaviourally measurable. Their structure reflects the architecture of moral cognition: operations on perceptual salience, affective intuitions, social meaning, and regulated deliberation. These are the *phenomena* that ethical theory seeks to interpret.

Second-order ethical theory is structurally different. It is reflexive rather than generative. It asks: What counts as a reason? What makes an obligation binding? What is the source of justificatory authority? These questions presuppose capacities for abstraction, generalisation, and rational evalu-

ation that are not themselves the proximate causal mechanisms of moral behaviour [88, 58, 53, 314, 59, 99, 63]. Sidgwick already insisted on this point in *The Methods of Ethics*, where he distinguished between the psychology of moral sentiments and the "*method* of determining right conduct" [144, Book I]. Lemos's treatment of epistemic justification exhibits a similar structural separation between doxastic psychology and the normative assessment of belief [**?**]. The parallel here is instructive: ethics stands to moral judgment as epistemology stands to belief-formation.

Seen from this perspective, second-order theory is not a set of instructions that moral agents follow in producing judgments. It is a framework for articulating the standards by which judgments are evaluated. It makes explicit the *normative architecture* that is only tacitly present in first-order moral life. Its success therefore depends on conceptual clarity and justificatory coherence, not on behavioural predictiveness.

8.3.2 Levels of Abstraction and the Proper Location of Ethical Explanation

The distinction between first-order moral cognition and second-order ethical theory can be sharpened through Floridi's framework of *Levels of Abstraction* (LoA) [125, 149]. On this account, every explanatory enterprise selects a perspective defined by its observables, its conceptual resolution, and the class of questions it is equipped to answer. Moral cognition and ethical theory do not merely operate at different LoAs—they answer *different kinds of questions* and employ *different explanatory primitives*.

At the **cognitive LoA**, the relevant variables are those that govern the generation of moral judgments in real time:

- perceptual salience and attentional capture,
- affective appraisal and embodied valuation,
- intuitive heuristics and rapid social inferences,
- controlled modulation under conflict or uncertainty,
- the temporal dynamics by which these processes integrate.

These are mechanistic, psychologically instantiated processes. They have causal influence on behaviour and can be perturbed by contextual or environmental changes. *This is the LoA at which the experimental work of this thesis operates.*

At the **normative LoA**, by contrast, the objects of analysis are:

- principles of justification,
- conceptions of duty, value, and obligation,
- standards of admissible reasons,
- structural norms governing deliberation, agency, and responsibility.

These are not causal operators but *interpretive* and *justificatory* constructs. They evaluate, discipline, or systematise moral claims but do not themselves generate behaviour. Ethical theory is reflexive: it examines the grammar of reasons, not the mechanisms of cognition.

Classical Machine Ethics collapsed these LoAs. By treating principles, rules, or utility structures as if they were mechanistic generative elements, it implicitly assumed that normative constructs function like cognitive processes. This assumption is doubly mistaken:

1. It attributes to normative concepts a causal role they do not possess: ethical duties do not operate like perceptual salience or affective appraisal.
2. It ignores the empirical architecture of moral cognition, which shows that behaviour emerges from intuitive, affective, and situational dynamics long before explicit reasoning is engaged.

From the perspective developed across this thesis, such an approach is not merely incomplete; it is methodologically incoherent. It attempts to engineer behaviour by manipulating abstractions at a LoA that is *not behaviourally operative*.

LoA discipline therefore becomes a philosophical and methodological necessity. Explanations of behaviour must occur at the cognitive LoA; evaluations of reasons and principles must occur at the normative LoA. Neither can be reduced to the other. Crucially, however, the two LoAs are not independent: normative evaluation presupposes an underlying psychology capable of generating moral sensitivity and action, while psychological findings constrain the plausibility of normative theories.

This interdependence is the key insight that links this chapter to the preceding Morality Primer and to the experimental chapter that follows. The Primer established that the cognitive LoA is *topologically structured*: moral cognition involves the continual reshaping of an evaluative field whose gradients are determined by affective cues, attentional dynamics, and social interpretive processes. Perturbations to this field—whether by altering salience, modifying affective tone, or introducing ambiguous social presence—can shift the system's behavioural trajectory even when normative commitments remain unchanged.

Seen through the LoA framework, the core question of this thesis can now be reformulated with greater precision: *How do normative expectations, psychological mechanisms, and environmental structures jointly determine the transition from moral perception to moral action?*

This question cannot be answered by ethical theory alone, nor by psychology in isolation. It requires a representational structure capable of linking the causal architecture of moral cognition (first-order) with the justificatory architecture of ethical evaluation (second-order). The remainder of this chapter argues that **evaluative topology**—introduced in the Morality Primer and returned to throughout the thesis—provides precisely such a bridge.

Classical Machine Ethics provides a clear illustration of the dangers of LoA confusion. A recurring methodological assumption in early systems was that normative concepts themselves—obligations, duties, utilities, or virtues—could be implemented at the computational LoA and thereby function as direct generators of behaviour. Early top-down approaches treated ethical theory as if its abstractions could be operationalised without remainder. For example, Arkin's "ethical governor" encoded deontological constraints derived from Just War Theory as behavioural regulators [122]; Anderson and Anderson's principlist architectures computationalised Rossian prima facie duties as decision rules [315, 16]; and logic-based approaches by Bringsjord and colleagues modelled deontic operators as executable action-selection mechanisms [124, 316]. Parallel lines of work assumed that utility functions could serve as moral evaluators in consequentialist agents [317, 122], while virtue-theoretic systems attempted to reify character traits as algorithmic dispositions governing moral performance [318, 319]. In all these cases, normative structures were treated as if they occupied the same LoA as the cognitive mechanisms responsible for actual moral behaviour.

Floridi's LoA framework clarifies why such reductions are unsustainable: normative categories belong to a reflective, second-order LoA concerned with justification, whereas computational models operate at an implementational LoA concerned with causal processes. Conflating the two not only mischaracterises the role of normative theory but also yields systems whose behavioural outputs are artefacts of representational choices rather than genuine ethical competence.

8.3.3 Evaluative Topology as a Bridge Between Orders

The challenge, then, is not to collapse first-order cognition into second-order theory, but to articulate a structure that permits principled interaction between them without confusing their explanatory roles. *Evaluative topology*, introduced in the Morality Primer (Chapter 4) and returned to throughout this thesis (see Chapter 7), provides precisely such a structure.

Evaluative topology can be naturally situated within a long-standing tradition in computational cognitive science that conceptualises perception, valuation, and action as parts of continuous, dynamical systems rather than discrete symbolic modules. Research in moral psychology already demonstrates that moral cognition emerges from distributed interactions between perceptual salience, affective appraisal, attentional dynamics, and context-sensitive social meaning. Empirical models—from Haidt's social intuitionism to Greene's dual-process account—show that moral perception is shaped by multi-dimensional affective and social fields rather than rule-based computations [88, 58, 53]. Neurocognitive analyses extend this point: Nussbaum's and Churchland's treatments of emotion as evaluative perception imply a graded, vector-like structure underlying moral appraisals [153, 111]. Likewise, work in social signal processing models interpersonal evaluation as a shifting landscape of cues that modulate behavioural trajectories in real time [131].

Against this background, evaluative topology provides a computationally meaningful formalisation: it treats the moral landscape as a dynamic field that shapes the flow from perceptual input to action readiness. Instead of assuming that be-

havior results from the application of discrete maxims or utility scores, evaluative topology models moral cognition as continuous transformations across a structured state-space. This aligns with dynamical-systems approaches in cognitive science that explain action selection through attractors, gradients of salience, and field-like organisation rather than propositional inference. The topology encodes the shape of the evaluative field—the stability of certain trajectories, the resistance of others, and the way local variations in perceptual or affective input can redirect the subject toward different moral outcomes.

By locating moral appraisal within a dynamic state-space, evaluative topology offers a principled bridge between first-order moral cognition and second-order ethical theory. It is sensitive to the empirical architecture of human cognition—distributed, affectively grounded, context-responsive—while remaining compatible with the reflective, justificatory concerns of ethical theory. It thus becomes possible to characterise the points of interaction between descriptive and normative orders without reducing one to the other: normative theory shapes the global constraints and evaluative contours within which first-order processes operate, while first-order processes provide the empirical basis upon which second-order theorising must reflect.

At its core, evaluative topology treats the moral landscape not as a set of discrete judgments or isolated principles, but as a *dynamic field* whose configuration determines the pathways through which perception becomes moral action [88, 58, 111, 53, 153, 59] . Its explanatory primitives include:

- **salience gradients**: patterns of perceptual and affective prominence,
- **affective attractors**: regions of the evaluative field toward which intuitive appraisal rapidly converges,
- **attentional pathways**: trajectories through which cognitive resources flow,
- **normative deformations**: structural constraints introduced by commitments, duties, or normative expectations,
- **social or synthetic perturbations**: distortions induced by the presence of other agents—including artificial ones.

Unlike classical ethical theory, which specifies norms at an abstract and often idealised level [144, 91, 33, 158, 44] , evaluative topology is sensitive to the *real-time architecture* of moral cognition. And unlike purely mechanistic models in psychology, which describe causal processes but lack normative structure, topology captures the relational, structural, and counterfactual properties of moral appraisal [88, 58, 53, 59, 144, 44, 158] : how evaluative trajectories *could* unfold under alternative configurations of salience, affect, or context.

This topological approach thus identifies the precise level at which first-order and second-order analyses intersect. It supports the following alignment:

1. **Ethical theory** identifies which evaluative configurations *ought* to have normative authority.

2. **Moral psychology** identifies which configurations *do* govern actual behaviour.
3. **Evaluative topology** identifies how these structures interact, when they diverge, and how they can be perturbed.

This tripartite structure yields both a diagnostic and a constructive insight. Diagnostically, it clarifies why many classical models in Machine Ethics failed: they attempted to engineer behaviour by manipulating abstractions at a normative LoA, ignoring the topological organisation of the cognitive LoA through which behaviour actually emerges. Constructively, it shows how normative analysis can be anchored in a psychologically realistic substrate without reducing ethics to psychology or cognition to normativity.

Topological Consequences for Moral Perturbation. The Morality Primer established that moral behaviour emerges from the traversal of a dynamically shaped evaluative field. Within this framework, *perturbation* has a precise and measurable meaning: any alteration that changes the curvature, gradients, or attractor structure of the field will shift the probability distribution over behavioural trajectories. This is true whether the perturbation arises from shifts in salience, affective modulation, attentional competition, or the introduction of a new agent into the interaction ecology.

A synthetic presence—perceptually social yet ontologically indeterminate—is therefore not merely an "observer" but a topological operator. It changes the field in which moral meaning becomes behaviourally operative. This was the central theoretical insight that shaped the experimental design: by embedding a morally charged cue (the Watching-Eye stimulus) within a field perturbed by a humanoid robot, we could test whether subtle topological deformation is sufficient to attenuate prosocial behaviour.

Interim Synthesis: Where the Chapter Now Stands. The conceptual architecture developed thus far establishes the conditions for experimental design (Chapter 7):

- First, moral judgment operates at the cognitive LoA through dynamic, affectively responsive, socially sensitive processes.
- Second, ethical theory operates at the normative LoA, providing justificatory structures but not generative mechanisms.
- Third, evaluative topology provides the bridge between these orders by modelling the structural constraints and transformations that govern the transition from moral perception to moral action.
- Fourth, this bridge is indispensable for understanding how synthetic agents perturb human moral behaviour.

We are therefore equipped to proceed. With the methodological scaffolding in place, we can now introduce the major normative theories not as abstract philosophical positions but as structured attempts to locate sources of normativity within the evaluative field. Their reconstruction in the next section is guided by

the LoA discipline established above and constrained by the topological account of moral cognition developed throughout this thesis.

Before turning to the main normative traditions, it is important to clarify *why* this reconstruction is required within the architecture of the thesis. The experimental work developed later does not simply measure behavioural differences; it interrogates a deeper question concerning the *normative interpretation* of those differences. If robotic co-presence reshapes the evaluative topology through which moral salience becomes action, then any claim about the ethical significance of this perturbation—whether it constitutes a moral cost, a distortion, or a benign behavioural shift—presupposes a framework for understanding how normativity itself is structured. Without situating the experiment within a landscape of ethical theories, one could describe *what* changes but not *what the change means*.

The purpose of the next section, therefore, is not to provide a survey of moral philosophy, but to identify the minimal normative scaffolding required to make sense of the empirical findings. Deontic, consequentialist, and virtue-theoretic perspectives articulate distinct accounts of (i) where normative authority resides, (ii) how moral relevance is determined, and (iii) how action-guidance is understood. These differences matter directly for the thesis: each theory yields a different interpretation of what it means for synthetic presence to attenuate prosocial behaviour. By reconstructing these normative architectures through the lens of Levels of Abstraction and evaluative topology, we prepare the conceptual ground for assessing the ethical significance of the perturbation demonstrated experimentally.

What follows, then, is not philosophical ornamentation but a methodological necessity: establishing the normative coordinates that will allow the later empirical results to be interpreted, evaluated, and ultimately situated within a defensible ethical framework.

8.4 The Normative Landscape: Structuring Ethical Theories Through LoA and Topology

With the methodological scaffolding now in place, we can introduce the major normative frameworks that constitute the philosophical backdrop against which the experimental findings must ultimately be interpreted. The aim here is not encyclopaedic exposition but conceptual reconstruction: each theory is presented in a form that preserves its philosophical integrity while situating it within the Levels of Abstraction (LoA) discipline and the evaluative-topological architecture developed in this thesis.

This reconstruction is guided by two methodological constraints:

1. **Philosophical fidelity**: the theories must be represented in a manner faithful to their canonical formulations in moral philosophy.
2. **Integrative compatibility**: the theories must be articulated in a form that allows principled interaction with the psychological and topological models of moral cognition established in Chapter 4.

The purpose of this section, therefore, is not to catalogue doctrines, but to map the deep structure of normativity in a way that can later illuminate the ethical

significance of the empirical perturbations induced by synthetic presence.

8.4.1 The Three Dimensions of Normative Analysis

Normative theories differ not only in content, but in the *architecture of normativity* they assume. To analyse them systematically, we distinguish three fundamental dimensions—each corresponding to an aspect of evaluative topology and LoA structure:

1. **Source of Normativity**: the origin of justificatory authority. This may lie in rational agency (Kant), human flourishing (Aristotle), aggregated welfare (Mill, Sidgwick), affective sentiment (Hume), or interpersonal justification (Scanlon).
2. **Mode of Evaluation**: the features of action or character deemed morally relevant—maxims, consequences, virtues, motives, relational duties, or context-sensitive particulars.
3. **Action-Guidance Mechanism**: the process that connects evaluative judgments to behaviour—categorical imperatives, utilitarian optimisation, virtue-structured perception, affective resonance, or justificatory equilibrium.

These dimensions allow us to re-express classical theories as *evaluative topologies*:

- **Kantian ethics** imposes rigid deontic invariants: absolute constraints that carve the evaluative field into sharply bounded permissible and impermissible regions.
- **Consequentialism** defines a gradient field over outcomes: moral action follows the steepest ascent toward welfare-maximising states.
- **Virtue ethics** defines dispositional attractors: stable patterns of moral sensitivity that shape the agent's perceptual and evaluative orientation.
- **Sentimentalism** defines networks of affective resonance: moral evaluation flows along affectively weighted pathways anchored in human sympathy or aversion.
- **Contractualism** defines justificatory equilibria: a topology structured by mutual recognisability of claims.
- **Particularism** dissolves fixed topologies altogether: normativity emerges from fully context-dependent patterns of salience and relation.

This analytic framing is essential because it provides a common representational language in which ethical theory and moral psychology can be jointly expressed. Theories that differ profoundly in content can be compared in structural terms—how they sculpt the evaluative landscape, where they locate normative constraints, and how they understand the movement from judgment to action.

8.4.2 Why This Framework Matters for the Experimental Chapter

This normative topology is not abstract machinery; it is the conceptual infrastructure that enables us to interpret what the experiment later reveals. The empirical

question—whether synthetic presence attenuates prosocial behaviour—cannot be ethically assessed without first situating it within a framework for understanding how moral cues acquire force.

Three claims follow directly from the preceding reconstruction:

1. **Moral action depends on the configuration of the evaluative field.** Normative theories specify different sources of authority and diverse mechanisms of action-guidance, but all agree that moral behaviour arises from structured evaluative relations, not arbitrary choice.

2. **Synthetic presence modulates this field by perturbing salience, attention, and affective resonance.** A humanoid robot does not supply new reasons; it reshapes the environment in which reasons become behaviourally operative.

3. **Normative theories must therefore be reinterpreted through the joint lens of LoA and evaluative topology if they are to explain or critique the behavioural perturbations observed experimentally.**

This is the philosophical function of the section: to establish the normative coordinates that will allow the experimental findings to be understood not merely as statistical differences, but as shifts in the moral significance of an action within a structured evaluative landscape.

The stage is now set for the substantive reconstruction. In the following sections, each major normative framework—deontological, consequentialist, virtue-theoretic, sentimentalist, contractualist, and particularist—is examined as a topology of normativity embedded within the cognitive–affective architecture of moral agents. These reconstructions will serve as the interpretive foundation for evaluating how, and why, synthetic presence can reshape the moral field in the experiment to come.

8.5 Deontological Structures: The Architecture of Practical Reason

The methodological framework established in the preceding sections motivates a disciplined reconstruction of the major normative theories. Having clarified how ethical explanation must respect both Levels of Abstraction (LoA) and the evaluative topology that mediates the transition from perception to action, we begin with deontological ethics. This is not because deontology offers a direct model of human moral cognition—it does not—but because it illustrates, with exceptional clarity, the gap between *normative authority* and *psychological generation*. This gap is precisely where classical Machine Ethics collapsed distinctions, and where the present thesis departs from that monolithic approach.

The aim here is not historical exegesis. The task is to reconstruct deontological normativity in a form compatible with the cognitive–topological architecture developed so far, and to show how deontological invariants function as structural constraints within the evaluative field investigated empirically in later chapters.

The reconstruction must satisfy three constraints:

1. **Preserve philosophical identity**: retain the core commitments that distinguish deontological ethics.
2. **Avoid LoA confusion**: do not treat deontic principles as if they were psychological mechanisms or generative cognitive operators.
3. **Embed deontology in topology**: express duties as constraints on the evaluative landscape, rather than as engines of behaviour.

When formulated in this way, deontology occupies a precise role: it identifies *invariant structures* within the moral field that delimit the boundaries of permissible action. These invariants are not computational rules; they are reflective standards through which agents assess the coherence of their maxims and commitments.

8.5.1 The Source of Normativity: Rational Agency and the Form of Law

On the Kantian account, moral authority arises from the structure of rational agency. The categorical imperative does not prescribe concrete actions but establishes a formal test for the permissibility of maxims: whether one's maxim could be willed as a universal law [**?**, 158, **?**]. This places the source of normativity at a *higher* LoA than psychological description. It concerns the *conditions of reflective justification*, not the causal mechanisms that generate everyday judgments.

This distinction is essential. Classical Machine Ethics implemented the categorical imperative as a procedural decision rule—an algorithmic operator [315, 16, 124, 320, 316, 122] . But Kant never intended universalisability tests to function as cognitive processes.[1] Their purpose is normative: to articulate the standards under which a maxim can be defended as consistent with rational agency. Treating these tests as computational procedures constitutes precisely the LoA confusion diagnosed earlier.

A survey of Classical Machine Ethics reveals this recurring methodological error: the assumption that Kantian constraints, universalisability tests, or duty-based norms could be directly implemented as procedural decision rules. Early top-down approaches explicitly treated the categorical imperative, or close deontological analogues, as algorithmic operators determining action permissibility. The most widely cited examples are the principlist architectures developed by Anderson and Anderson, where prima facie duties are computationalised as weighted decision procedures whose outputs determine ethically "permissible" behaviour [315, 16]. Similarly, logic-based systems developed by Bringsjord and collaborators represent obligations and prohibitions using deontic logic embedded in the cognitive event calculus, thereby converting normative constraints into executable operators that mechanically evaluate action options [124, 320]. Ganascia's formalisation of ethical rules of warfare follows the same strategy, modelling universally applicable duties as logical conditions that an autonomous agent must satisfy prior to acting [316]. Arkin's "ethical governor" for lethal autonomous robots

[1] See the discussion in [**?**] and [117] on the reflective rather than psychological status of the categorical imperative.

likewise encodes deontological constraints—derived from Just War Theory and Kantian doctrine—as computational filters that block impermissible actions at run time [122]. In each case, a normative principle originally intended for reflective justification is treated as a psychological mechanism or behaviour-generating operator. As Moor and Coeckelbergh observe, this amounts precisely to the Level-of-Abstraction confusion: normative tests designed for rational self-assessment are misinterpreted as causal algorithms capable of producing moral behaviour [13, 4]. These systems thus instantiate the very conflation at issue—collapsing reflective ethical reasoning into first-order cognitive processing.

8.5.2 Mode of Evaluation: Maxims, Duties, and the Structure of Permissibility

Deontological theories evaluate actions through the *form* of the underlying maxim and the duties that follow from rational consistency. These duties generate a characteristic structure within the evaluative field:

- **Invariance**: duties bind independently of consequences or affective states.
- **Non-gradience**: obligations typically define discrete boundaries—permissible vs. impermissible.
- **Symmetry**: the universal law test imposes interpersonal consistency.
- **Role-relativity**: some duties depend on one's position or relationship (e.g. duties of fidelity, respect, and beneficence).

Topologically, these features correspond to *hard constraints* on the evaluative landscape. Rather than shaping the gradients that guide behaviour, deontological duties carve the field into admissible and inadmissible regions. They define the regulatory geometry within which trajectories must lie.

8.5.3 Action-Guidance: How Normative Constraints Influence Behaviour

A central challenge arises here: if deontological rules do not describe cognitive processes, how do they guide action?

The answer, consistent with LoA discipline, is twofold:

1. **At the cognitive LoA**: deontological principles do not produce behaviour. Moral action emerges from intuitive appraisal, affective valuation, attentional salience, and controlled modulation—precisely the components analysed in the Morality Primer (Chapter 4).
2. **At the normative LoA**: deontological principles determine which behavioural trajectories can be reflectively justified. They also shape long-term dispositions, thereby influencing the evaluative topology indirectly through moral training, socialisation, and self-constitution.

Thus, while deontology does not operate the machinery of moral cognition, it contributes to the *calibration* of that machinery over developmental time. Internalised deontic commitments:

- heighten sensitivity to cues of respect and violation,
- modulate affective responses to dishonesty or unfairness,
- strengthen top–down control when intuitive impulses conflict with duty.

In this sense, deontological ethics functions as a form of *normative scaffolding*: it shapes the agent's evaluative posture but does not compute their moment-to-moment behaviour.

8.5.4 Deontological Normativity as Topological Invariance

We can now state the central insight of this reconstruction. Within a topological model of moral cognition, deontological ethics corresponds to the identification of *non-negotiable invariants*—fixed points that define the structural integrity of the moral field.

These invariants:

- partition the space of possible actions into permitted and forbidden zones,
- resist deformation by contextual changes, affective fluctuations, or strategic incentives,
- stabilise behavioural tendencies by constraining rational endorsement,
- provide the reflective standpoint from which agents assess the legitimacy of their conduct.

The categorical imperative thus appears not as an algorithm for decision-making but as a *topological principle*: a formal constraint ensuring that evaluative structure is globally coherent rather than locally opportunistic.

8.5.5 Why Deontology Matters for the Experimental Logic

This reconstruction is essential for integrating the experiment into a normative framework. The purpose of the experiment is not merely to detect behavioural differences but to determine their *moral* significance. Deontology supplies the conceptual structure required for this evaluation.

Before stating the relevance of deontological norms for the experimental logic, one brief clarification is required. Throughout this thesis, the experimental paradigm employs a widely studied behavioural prime sometimes referred to as a "Watching-Eye" cue: a minimal visual stimulus (in our case, a charity poster depicting a child in need) that subtly increases the perceived presence of a moral or social observer. The detailed psychological literature and methodological justification for this paradigm are presented later in Chapter 7. Here, it suffices to note that such cues are known to activate expectations of accountability, reciprocity, and norm compliance—even though they involve no real observer and no explicit instruction.

With this context in place, we can now express why deontological theory is indispensable for interpreting the experiment:

1. **If synthetic presence alters behaviour**, we must ask whether the observed perturbation reflects a shift that remains within deontically permissible space or whether it involves a deeper distortion of obligations associated with beneficence, fairness, or respect.
2. **The Watching-Eye cue implicitly invokes deontic expectations**: even a minimal representation of an observing other tends to activate norms of accountability and reciprocity. A reduction in prosocial action under this cue suggests that the presence of a synthetic agent may interfere with the agent's sensitivity to these deontic constraints.
3. **Deontology provides the normative vocabulary** for diagnosing whether a behavioural shift constitutes a morally relevant deviation or a benign modulation of preference or affect.

This is precisely where the present thesis diverges from monolithic approaches in Machine Ethics. Classical frameworks attempted to model moral action by encoding deontological rules directly into artificial agents. The empirical results of this thesis show why that strategy misunderstands the architecture of moral cognition: deontic rules do not generate behaviour, and perturbations to behaviour cannot be understood purely in terms of deviations from codified principles. Instead, the influence of synthetic presence must be interpreted through the evaluative topology in which deontic invariants reside.

With deontology reconstructed as a system of topological constraints rather than computational rules, we can now turn to consequentialism. There, normativity is expressed not through invariants but through gradient fields over outcomes—structures that interact with the evaluative machinery of moral cognition in different but equally illuminating ways. This will further clarify how different theoretical lenses illuminate different dimensions of the behavioural perturbations uncovered in the experiment.

Conceptual Note: Gradient Fields in Consequentialist Topology

In the topological framework developed across this thesis, a *gradient field* designates a structured evaluative landscape in which each possible action or state of the world is associated with a scalar value—typically representing expected welfare, utility, or outcome-based moral worth. Formally, a gradient field assigns to each point in an abstract space of action–outcome configurations a direction of steepest ascent: the direction in which an incremental shift would produce the greatest increase in expected value. In classical moral philosophy, this structure is implicit in utilitarian reasoning, which assesses actions by their tendency to promote the greatest balance of good over bad consequences [**?**, 33, 144]. Within this thesis, the notion is used in a non-formal but conceptually rigorous sense: as a way of modelling how consequentialist evaluation imposes directional structure on the moral field, where moral improvement corresponds to movement along the gradient toward higher expected welfare.

A gradient field thus has three key features:

1. **Scalar valuation**: each point in the evaluative space has a determinable value, allowing continuous comparison along a single dimension of moral

assessment (e.g. total or average welfare).

2. **Directional guidance**: the moral significance of a possible action is given by its vector orientation relative to the gradient; actions are increasingly morally preferable as they align with the direction of steepest ascent.

3. **Sensitivity to empirical structure**: because the gradient depends on expected outcomes, it varies with changes in belief, evidence, context, and the agent's model of the world.

In this topological reconstruction, consequentialist gradient fields do not function as cognitive mechanisms. Human agents do not compute explicit gradients when acting morally, nor do they evaluate global states of the world through analytic integration. Rather, consequentialist structures operate at the *normative Level of Abstraction*: they specify how actions are *justified* in reflective evaluation, not how they are generated in real-time cognition. This LoA separation parallels Sidgwick's distinction between the "point of view of the universe" and ordinary motivational psychology [144, Book IV].

Interaction with the Evaluative Machinery of Moral Cognition. Although gradient fields do not describe the causal architecture of moral cognition, they interact with it in conceptually important ways. The evaluative machinery developed in Chapter 4—perceptual salience, affective appraisal, intuitive heuristics, and controlled modulation—does not implement consequentialist reasoning, but it is nevertheless shaped by outcome-related information in several distinct modes:

1. **Salience modulation**. Perceived consequences influence which features of a situation become salient. Potential harm, benefit, or risk amplifies attentional capture, thereby altering the local configuration of the evaluative field even before explicit reasoning occurs.

2. **Affective valuation**. The human affective system registers outcomes (especially those involving harm or welfare) with strong valence. These affective signals act as local gradient approximations: they bias intuitive appraisal toward or away from particular actions in a manner that roughly tracks expected value.

3. **Heuristic extraction**. Over developmental time, agents internalise outcome-sensitive heuristics ("help when it is easy", "avoid causing harm") that serve as psychologically tractable proxies for gradient following. These heuristics allow the cognitive system to approximate consequentialist structure without computing it.

4. **Deliberative correction**. In cases of conflict or ambiguity, controlled processes may approximate aspects of consequentialist evaluation—comparing potential harms or weighing benefits—thereby engaging the gradient field at a coarse-grained level. However, this is slow, effortful, and limited by computational constraints.

5. **Perturbation sensitivity**. Because consequentialist evaluation depends on expected consequences, perturbations to perception, attention, or social meaning—such as the presence of a humanoid robot—can reshape the

agent's perceived gradient field. This makes consequentialist structures especially sensitive to the kinds of environmental shifts tested experimentally in this thesis.

The interaction between consequentialist topology and moral cognition therefore occurs *indirectly*. Consequentialism specifies the normative gradient that ought to guide reflective endorsement; the cognitive system provides a noisy, heuristic, context-sensitive approximation of this structure. Evaluative topology makes this relationship explicit by modelling behaviour as the traversal of a dynamically shaped field whose gradients, although not explicitly computed by the agent, are nevertheless partially approximated through affective and attentional processes.

This conceptual integration is essential for the purposes of the present thesis. It allows consequentialism to be reconstructed in a form compatible with the empirical findings that moral behaviour is sensitive to subtle perturbations in the perceptual–social environment. It also provides one of the normative lenses through which the experimentally observed attenuation of prosocial donation under synthetic presence can be interpreted: as a topological distortion of the gradient field that normally favours prosocial action.

8.6 Consequentialist Structures: Value Gradients and the Topology of Outcomes

Having reconstructed deontological ethics as a system of topological invariants that constrain the space of permissible action without directly generating behaviour, we now turn to the second major normative framework: consequentialism. Here the conceptual architecture differs in every relevant dimension. Where deontology posits *fixed boundaries* within the evaluative field, consequentialism posits *gradients*. Where deontology locates normativity in the form of maxims, consequentialism locates it in the structure of outcomes. And where deontology articulates duties, consequentialism articulates value-based trajectories across possible states of the world.

As with deontology, the aim is not historical exegesis. Rather, the task is to reconstruct consequentialism in a way compatible with the LoA discipline and the evaluative–topological model developed so far. In particular, we are interested in how a consequentialist structure can be read as a *gradient field* over outcomes that exerts normative pressure on action, and how such a field is liable to perturbation when the perceptual–social environment is modified by synthetic presence. This reconstruction will furnish one of the normative perspectives through which the experimental findings on moral displacement are interpreted.

8.6.1 The Source of Normativity: Welfare, Impartiality, and the Structure of Reasons

Classical utilitarianism grounds moral authority in the promotion of welfare. In its canonical formulations—Bentham's felicific calculus [**?**], Mill's qualitative hedonism [33], and Sidgwick's systematic treatment of practical reason [144]—consequentialism maintains that what ultimately matters is the value of outcomes,

impartially aggregated across persons. An action is right, in the strict sense, insofar as it maximises (or sufficiently promotes) overall good; wrong insofar as it fails to do so.

From the standpoint of Levels of Abstraction, this locates consequentialist normativity at a *reflective* LoA concerned with:

- the evaluation and comparison of outcomes,
- the aggregation of welfare across individuals,
- and the impartial justification of action in light of such aggregation.

As with deontology, these commitments are not descriptive claims about the mechanisms of moral cognition. Sidgwick is explicit that the "point of view of the universe" is *not* the standpoint from which ordinary agents habitually deliberate; it is a standard of justification, not a psychological model of motivation [144, Book IV]. Consequentialism specifies a standard of rightness, not an algorithm that human agents actually implement.

This distinction is crucial for our purposes. Classical Machine Ethics has often treated utilitarian or outcome-based formalisms as if they were *psychologically generative*: reward functions, expected-utility maximisation, or cost–benefit optimisers are proposed not merely as normative ideals but as surrogates for moral cognition itself. Within the LoA framework, this is a category error. Consequentialism operates at the normative LoA; the evaluative machinery described in the Morality Primer (Chapter 4) operates at the cognitive LoA. Any mapping between the two must be justified rather than assumed.

8.6.2 Mode of Evaluation: Consequences, Expected Value, and Scalar Normativity

Consequentialism evaluates actions in terms of the value of their (actual or expected) outcomes. Unlike deontological theories, which typically yield binary constraints (permissible/impermissible), consequentialism is *scalar*: options can be better or worse to any degree. This scalar structure has direct topological expression.

In the evaluative-topological model, a consequentialist landscape is characterised by:

- **Gradience**: the moral field is continuous; small differences in expected welfare correspond to small differences in moral ranking.
- **Optimisation**: morally preferable actions correspond to local or global maxima along welfare gradients.
- **Context-sensitivity**: the shape of the field depends on empirical facts about consequences (who is helped, who is harmed, how much, under what conditions).
- **Impartiality**: regions of the field corresponding to welfare changes have equal moral standing irrespective of whose welfare is at stake.

Because of these features, consequentialism lends itself naturally to computational representation: utility functions, cost–benefit analyses, and optimisation routines approximate the mathematical structure of value gradients. This explains its appeal in Machine Ethics and reinforcement-learning–based approaches, where "ethical" behaviour is often equated with maximising a suitably designed reward function.

But again, computational tractability must not be confused with cognitive realism. Human moral cognition, as reviewed in Chapter 4, does not perform explicit global optimisation over expected outcomes; it operates through heuristic, affective, and context-sensitive processes that are only loosely correlated with the ideals of consequentialist reasoning [99, 58, 88, **?**]. Treating human agents as if they literally implemented expected-utility maximisation is therefore another instance of LoA confusion.

8.6.3 Action-Guidance Mechanism: From Value Gradients to Behavioural Pressure

How, then, does consequentialism guide action without collapsing into a psychologically implausible calculus? The answer, consistent with LoA discipline, is that consequentialism exerts its influence primarily through *indirect modulation* of the evaluative topology rather than through direct computational implementation.

At the reflective LoA, consequentialism states:

> An action is right insofar as it maximises (or sufficiently promotes) expected welfare.

At the cognitive LoA, however, moral behaviour is produced by the interaction of intuitive appraisal, affective resonance, social cues, and controlled regulation. Consequentialist considerations can shape this machinery over time via at least four pathways:

- **Long-term shaping of dispositions**: education and moral reflection can increase sensitivity to outcomes, harm, and aggregate effects, thereby steepening certain evaluative gradients (e.g. aversion to needless suffering).
- **Local heuristics**: agents employ proxy rules (e.g. help when the cost is low; avoid imposing serious harm) that correlate, imperfectly, with welfare improvement.
- **Attentional modulation**: awareness of potential benefits or harms alters salience and intuitive appraisal; some features of a situation become more behaviourally weighty.
- **Regulatory control**: when intuitive impulses conflict with perceived consequences, deliberation may re-weight options in favour of outcome-based considerations.

In topological terms, consequentialism does not "run" the cognitive system, but it can influence the *shaping* of the evaluative field: steepening or flattening gradients, reorienting trajectories, and altering which outcome-dimensions become behaviourally decisive.

8.6.4 Consequentialist Topology: Moral Action as Gradient Following

Within the topological framework of this thesis, we can now express the core consequentialist intuition succinctly: moral action is modelled as (approximate) *gradient following* in a welfare-defined landscape. Behaviour is normatively preferred when it moves "uphill" along value gradients.

This has several structural implications:

1. **Smoothness**: unlike deontological boundaries, consequentialist fields permit smooth transitions. Moving from a slightly worse to a slightly better outcome traces a continuous path in evaluative space.
2. **Directionality**: what matters is not merely where an agent is, but the direction of movement—toward or away from higher-welfare states.
3. **Trade-offs**: multi-dimensional outcomes (e.g. helping one party while imposing small costs on another) are represented as interacting gradients over several axes.
4. **Sensitivity to perturbation**: because evaluation tracks expected consequences, shifts in salience, attention, or perceived observer-interest directly reshape the gradient structure.

This final feature connects consequentialism to the experimental logic. If the perceived consequence structure of donation is altered by synthetic presence—because the social meaning of helping changes, or because the anticipated payoffs (reputational, affective, or interpersonal) are attenuated—then the agent's trajectory through the evaluative field will shift accordingly.

8.6.5 Why Consequentialism Matters for the Experimental Logic

Consequentialism is indispensable for one dimension of interpreting the behavioural perturbations observed in the experimental chapter. At the LoA relevant for our experiment, prosocial donation is simultaneously:

- a *behavioural output* of the moral cognitive architecture,
- and a *welfare-relevant action* whose outcomes (for the beneficiary) can be straightforwardly ranked.

Within this frame, the Watching-Eye prime and the robot's synthetic presence can be understood as modulating the *perceived consequence structure* of donating.

1. **Watching-Eye cues reshape anticipated social consequences.** As discussed in Chapter 7, visual cues suggesting observation are known to increase the perceived reputational or social-evaluative payoff of prosocial behaviour. In topological terms, they steepen the gradient pointing toward donation by enhancing the expected social value of helping.
2. **Synthetic presence can interfere with or redirect this gradient.** The humanoid robot constitutes an ambiguous social agent whose presence may blunt, re-route, or partially occlude the evaluative pathways activated by the Watching-Eye cue. If the robot absorbs attention, disrupts affective

resonance with the charity target, or is not integrated into the same social-evaluative schema as a human observer, the effective gradient from "keep the money" to "donate" may be flattened.

3. **Consequentialism provides one axis of normative diagnosis.** If donation falls in the Robot condition, one interpretation—from a consequentialist perspective—is that synthetic presence has deformed the outcome-based evaluative field: the agent no longer experiences donating as sufficiently welfare-improving or socially valuable relative to alternatives. This differs from a purely deontic diagnosis (failure to track duty) or a purely virtue-theoretic diagnosis (shift in character-expressive patterns).

Consequentialism thus illuminates a specific facet of the moral displacement effect: the way in which synthetic presence can alter the perceived benefits, costs, and social meaning of helping, thereby reshaping the value gradients that normally support prosocial behaviour. Importantly, the thesis does *not* treat this consequentialist structure as a blueprint for machine implementation, in contrast with classical Machine Ethics approaches that equate "ethical design" with encoding explicit utility functions. Instead, consequentialism is used here as a normative lens on how the evaluative topology is perturbed by synthetic agents.

The next section turns to virtue ethics, which locates normativity not primarily in constraints or consequences, but in the cultivated dispositions and perceptual sensitivities of the agent. This will allow us to examine a further dimension of the evaluative topology: how character, habituation, and moral perception shape the susceptibility of prosocial action to perturbation by robotic co-presence.

8.7 Virtue-Theoretic Structures: Dispositions, Character Topology, and Moral Sensitivity

Deontological invariants and consequentialist gradients capture two important dimensions of the evaluative field, but they remain incomplete without a theory of the *agent* who navigates that field. Virtue ethics—from Aristotle through modern neo-Aristotelian and psychological reconstructions [28, 321, 35, 322]—locates normativity not primarily in constraints or outcomes, but in the *perceptual and dispositional architecture* of the moral agent. This renders virtue ethics particularly well-suited for integration with the experimental findings of this thesis, which show systematic modulation of prosocial action by latent personality dimensions and cluster-level structure in trait space (see Chapter 7).

Our task is therefore to reconstruct virtue ethics in a form that satisfies three conditions:

1. It must preserve the philosophical distinctiveness of virtue theory as an account of normativity grounded in character and moral perception.
2. It must be expressible in the evaluative-topological idiom developed across this thesis, allowing traits to modulate the curvature and attractor structure of the moral field.
3. It must connect directly to the empirical results: latent trait configurations, cluster-dependent moral deformation, and the mathematically described

perturbations induced by synthetic presence.

With these constraints in place, virtue theory becomes more than a catalogue of excellences: it becomes a theory of *moral sensitivity as a topologically structured, personality-dependent field*, modulated both by long-term habituation and by local perturbations such as robotic co-presence.

8.7.1 The Source of Normativity: Character, Practical Wisdom, and Moral Perception

In the virtue-theoretic tradition, normativity originates in the *well-formed character* of the agent, rather than in rules or external valuations. Virtues are not propositional commitments but *stable dispositional patterns* that structure moral perception: they determine what the agent notices, how she evaluates it, and which actions appear salient, fitting, or required [93, 321]. Aristotle's concept of *phronesis*—practical wisdom—captures the idea that virtuous action arises from the *fine-tuned sensitivity* to morally relevant features of a situation [28].

This has a direct analogue in the evaluative topology introduced earlier. A virtuous agent is one whose evaluative field contains:

- **stable attractors**: behavioural basins corresponding to courage, benevolence, honesty, fairness;
- **well-shaped gradients**: moral salience that shifts the system reliably toward prosocial trajectories;
- **robustness under perturbation**: resistance to minor contextual noise and situational fluctuation.

Conversely, deficiencies in character appear as distortions or instabilities in the evaluative field: shallow attractors, flattened gradients, or poorly integrated response tendencies.

8.7.2 Mode of Evaluation: Dispositions as Topological Structure

Virtue ethics does not evaluate actions in isolation but assesses them as *expressive of character*. The morally relevant unit is the dispositional pattern through which the agent perceives and structures her moral environment. This is where virtue theory intersects most naturally with the experimental findings.

(i) Mathematical and Topological Interpretation

Let the agent's dispositional profile be represented by a vector

$$\beta_C \in \mathbb{R}^k,$$

where k indexes latent psychological traits (e.g. agreeableness, empathy, conscientiousness). The experimental analyses in Chapter 7 demonstrate that participants form coherent clusters $C_1, C_2, \ldots, C_m$ in this trait space, each with characteristic dispositions.

We can therefore interpret virtue-theoretic structure as a topological mapping

$$\mathcal{T} : \mathbb{R}^k \to \mathcal{F},$$

where $\mathcal{F}$ is the space of evaluative fields. Under this model:

- high-agreeableness clusters exhibit deeper prosocial attractors; - low-empathy clusters exhibit shallower or displaced prosocial basins; - high-conscientiousness clusters show increased boundary rigidity for deontic constraints; - neuroticism modulates sensitivity to evaluation cues (including the Watching-Eye effect).

In virtue-theoretic terms, β_C approximates a parametric description of the agent's *character topology*. This mapping was borne out empirically: different clusters showed markedly different susceptibility to moral deformation under synthetic presence, precisely as a virtue-ethical model predicts.

(ii) Connection to Moral Psychology

Modern moral psychology (e.g. the *moral foundations* approach [45], the *character-based* models of Snow [**?**], and the *sensitivity-based* accounts of Dancy [92]) emphasises that moral responsiveness is a function of dispositional configuration. Trait-dependent modulation of salience, empathy, and social attentiveness mirrors the classical virtue-theoretic notion that moral judgment depends on habituated perception.

Our empirical data confirm this: the presence of the robot altered prosocial behaviour differentially across personality clusters, demonstrating that the moral field is not homogenous but *character-structured*.

8.7.3 Action-Guidance Mechanism: Habituation, Stability, and Situated Sensitivity

Virtue ethics explains action not by invoking explicit principles or value calculations but through the *habituated, stabilised patterns of salience and response* characteristic of a well-formed agent. This aligns neatly with the dual-process architecture established in Chapter 4:

- intuitive processes are shaped by long-term habituation into affective–perceptual sensitivities,
- controlled processes integrate commitments and identities developed over time,
- behavioural output reflects the stability or fragility of dispositional attractors.

In topological terms, virtues correspond to *deep attractor basins* resistant to perturbation; vices or deficiencies correspond to *shallow or unstable attractors*. This interpretation is supported by both computational models of habit formation [**?**] and empirical studies of moral perception [**?**].

8.7.4 Virtue-Theoretic Topology: Stability, Curvature, and Susceptibility to Perturbation

Within the evaluative-topological framework, virtue ethics can be modelled using dynamical systems language:

$$\dot{x} = f(x; \beta_C),$$

where x is the agent's state in evaluative space and β_C parametrises dispositional curvature. The presence of a synthetic agent introduces a perturbation δf such that

$$\dot{x}' = f(x; \beta_C) + \delta f(x; \mathscr{R}),$$

where $\mathscr{R}$ denotes robotic co-presence.

Crucially:

- for some clusters, δf shifts the trajectory away from the prosocial attractor basin (attenuation of donation); - for others, the attractor curvature remains sufficiently deep that the perturbation is absorbed; - for exceptionally prosocial configurations, synthetic presence may even sharpen evaluative focus (rare, but consistent with the upper-tail donors observed).

This constitutes a clear virtue-theoretic phenomenon: moral sensitivity is *trait-dependent*, and synthetic perturbation reveals structural differences in the stability of character.

8.7.5 Why Virtue Ethics Matters for the Experimental Logic

Virtue ethics is indispensable for interpreting the experimental results for three interconnected reasons.

1. Latent Trait Modulation The experiment confirms that moral perturbation is not uniform: clusters in personality space exhibit distinct patterns of deformation. Virtue theory provides the conceptual vocabulary for understanding these effects as differences in character topology. Prosocial action is more fragile in agents with shallow attractors; synthetic presence perturbs these evaluative structures disproportionately.

2. Moral Topology Over Trait Space The mapping

$$\beta_C \mapsto \mathcal{T}(\beta_C)$$

establishes that moral responsiveness is a *function of trait geometry.* This is a virtue-theoretic insight: character is the medium through which the environment's moral affordances are processed.

3. Machine Ethics Ignores Character Entirely Classical Machine Ethics frameworks assume that ethical behaviour can be engineered through top-down rules or utility functions. They contain no representation of dispositional structure, no equivalent of β_C, no account of habituation, and no model of trait-dependent sensitivity to perturbation. This makes them incapable of predicting—or even recognising—the character-mediated moral displacement observed in our experiment.

Virtue ethics therefore reveals the deepest limitation of rule-based or utility-based Machine Ethics: moral agency is fundamentally *dispositional*, and no architecture that ignores habituated sensitivity, perceptual tuning, and trait-level topology can claim to model it.

In sum, virtue ethics interprets the experimental findings as a demonstration that synthetic agents perturb moral action by interacting with the agent-specific topology shaped by habituation, character, and perceptual attunement. Where deontology contributes boundary conditions and consequentialism contributes gradient structure, virtue ethics contributes the *curvature of the evaluative manifold itself*: the dispositional geometry that determines how agents absorb, refract, or amplify perturbation.

Interim Synthesis: How the Three Normative Frameworks Illuminate the Experimental Findings

With deontology, consequentialism, and virtue ethics reconstructed through the discipline of Levels of Abstraction and embedded within the evaluative-topological architecture developed across this thesis, we can now articulate their practical significance for the experimental results.

The purpose of this synthesis is not merely classificatory. It is to demonstrate why an empirical study of synthetic social influence *requires* a multi-framework normative lens, and why no single classical theory is sufficient to interpret the perturbations observed in prosocial donation.

1. Deontology: Structural Invariants and the Integrity of Moral Expectation

In the deontological reconstruction, duties appear as *invariant boundaries* of the evaluative field. The Watching-Eye cue—as developed in Chapter 7—implicitly invokes precisely these invariants: reciprocity, fairness, honesty, and the demand to act as if one's behaviour were publicly accountable.

- When donation decreases in the Robot condition, the key normative question is whether the perturbation reflects a weakening of sensitivity to these invariants.
- If synthetic presence "flattens" the deontic landscape—attenuating the felt pull of duty—then the perturbation carries ethical weight beyond behavioural variation.
- The deontological analysis therefore provides the vocabulary to distinguish between a mere preference shift and a disruption in the *structural preconditions* of rightful agency.

Empirically, this interpretation is strengthened by the fact that deontic cues (the child's face, the moral framing of donation) remain constant across conditions. The only structural change is the presence of the synthetic observer. This isolates the robot as a potential *interference with deontic uptake*. A purely psychological description would register the attenuation as behavioural variance; a deontological analysis reveals it as a possible distortion of moral accountability.

2. Consequentialism: Gradient Deformation and the Perceived Structure of Outcomes

From a consequentialist standpoint, moral orientation depends on the perceived gradient of expected value. Watching-Eye cues work partly because they shift the perceived payoff structure: being observed increases reputational benefit, reduces social cost, or reinforces anticipated approval.

The robot's presence perturbs this gradient in three ways:

1. It introduces an *ambiguous observer* whose evaluative stance is unclear, flattening or redirecting the perceived payoff of donation.
2. It may compete with, divert attention from, or overshadow the reputational signal emitted by the Watching-Eye stimulus.
3. It may shift the perceived "social meaning" of the interaction, transforming a dyadic human–charity cue environment into a triadic human–robot–cue environment.

In topological terms, the synthetic presence deforms the gradient field: it alters the local slope of the utility landscape surrounding prosocial action. This consequentialist diagnosis captures aspects of the perturbation that the deontological analysis cannot. Whereas deontology cares about the structural integrity of duties, consequentialism cares about the *direction and magnitude of evaluative flow*. The empirical attenuation fits naturally into this model: synthetic presence recalibrates anticipated outcomes, producing a shallower gradient toward the prosocial basin.

3. Virtue Ethics: Dispositional Curvature and Cluster-Dependent Sensitivity

The virtue-theoretic reconstruction offers yet another lens—one that matches the empirical findings with remarkable precision. Virtue ethics treats moral responsiveness as a function of dispositional structure: the shape, depth, and stability of an agent's evaluative attractors.

8.7.6 Virtue-Ethical Interpretation of Latent Ecologies

The experiment revealed this pattern with striking clarity when analysed through the semantic ecology of the latent trait clusters:

- **Prosocial–Empathic / Warm–Sociable Ecology**: stable, deep prosocial attractors; moral trajectories remained largely invariant under synthetic perturbation. Donation behaviour persisted despite the introduction of the robot, indicating a robust evaluative surface anchored in empathic resonance and interpersonal sensitivity.
- **Emotionally Reactive / Low-Structure Ecology**: shallow, volatile, or displaced attractors; this ecology exhibited the *strongest attenuation* under robotic presence. Their evaluative field displays low structural coherence and heightened responsiveness to contextual cues; the robot's ontological ambiguity therefore diffuses or refracts moral salience at precisely the stage where affective anchoring would normally stabilise action.

- **Analytical–Structured / High-Systemizing Ecology**: intermediate curvature; these participants showed partial but not catastrophic displacement. Their evaluative architecture privileges clarity and rule-structure over affective immediacy, making them comparatively resistant to affective perturbation, but sensitive to disruptions of interpretive coherence.

This ecological differentiation is *not a behavioural epiphenomenon*: it is the virtue-ethical signature of the data. The robot does not act as a uniformly applied suppressor (γ_R); rather, it functions as a *contextually instantiated perturbator* whose behavioural impact depends on the dispositional topology encoded in β_C.

$$\dot{x}' = f(x; \beta_C) + \delta f(x; \mathscr{R})$$

where $f(x; \beta_C)$ is the endogenous evaluative drift governing each ecological type's baseline moral trajectory, and $\delta f(x; \mathscr{R})$ is the deformation induced by synthetic presence.

Crucially, δf is *not* constant. Its sign, magnitude, and functional shape vary across ecologies:

- In the **Prosocial–Empathic** ecology, δf attenuates the empathic–affiliative attractor, flattening the gradient that normally drives prosocial donation.
- In the **Emotionally Reactive** ecology, δf interacts with an already unstable evaluative surface, amplifying volatility and producing the sharpest behavioural displacement.
- In the **Analytical–Structured** ecology, δf perturbs coherence rather than affect, leading to partial reconfiguration but not collapse.

This is precisely the prediction of virtue ethics: the moral perturbation induced by $\mathscr{R}$ is mediated not by rule-following or outcome-calculation, but by the dispositional configuration of the agent—their stable tendencies of attention, valuation, and motivational salience.

Viewed through Floridi's Levels of Abstraction, the latent ecologies constitute **distinct semantic filters**:

- The **Prosocial–Empathic LoA** foregrounds affective and interpersonal cues.
- The **Emotionally Reactive LoA** foregrounds volatility, ambiguity, and contextual instability.
- The **Analytical–Structured LoA** foregrounds coherence, formal clarity, and normative intelligibility.

The robot's ambiguous ontology—neither fully social nor fully inert—is refracted through these LoAs differently, producing *topologically distinct perturbations*. This explains why synthetic presence yields neither a uniform nor a homogeneous effect, but one that is *contingent upon the semantic architecture* that each ecological profile brings to the interaction.

Thus, the virtue-ethical reconstruction, the latent-trait analysis, and the topological model converge: the robot reveals, with unusual diagnostic precision, the dispositional geometry of each ecological type. Moral displacement is not merely a behavioural effect; it is a principled probe into the internal architecture of moral cognition.

5. What Machine Ethics Misses

This synthesis also exposes the limitations of classical Machine Ethics:

1. It assumes that behaviour can be derived from explicit rules (deontic codification), ignoring the psychological mechanisms through which duties gain behavioural force.
2. It assumes that welfare optimisation can be implemented directly through utility functions, ignoring the fact that humans do not compute value gradients explicitly and are sensitive to subtle perturbations.
3. It ignores dispositional topology entirely; it has no representation for character, habituation, sensitivity, or cluster-dependent variation.

Thus, Machine Ethics repeatedly commits the LoA mistake: treating normative abstractions as if they were cognitive operators. The experiment demonstrates why this is untenable: moral behaviour emerges from topological dynamics that Machine Ethics has no resources to model.

6. Concluding Perspective: Why This Matters Now

The virtue-theoretic, deontological, and consequentialist reconstructions converge on a single insight: the moral significance of synthetic presence cannot be captured by any one normative theory alone, nor can it be reduced to behaviourist regularities. It must be understood as a deformation of the evaluative topology through which agents convert moral salience into action.

The experiment does not merely show that robots change behaviour. It shows *how* they do so: by reshaping deontic sensitivity, altering perceived consequence gradients, and interacting with deep dispositional structures. This threefold interpretation is the normative analogue of the empirical results—and it furnishes the philosophical scaffolding for the sentimentalist, affect-based account developed next.

The next section introduces sentimentalist and affect-based accounts, which complement the dispositional framework by modelling the affective vectors that shape the immediate moral landscape and interact with the latent trait structure identified above.

8.8 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology

Having reconstructed deontology as topological invariance and consequentialism as value-gradient optimisation, we now turn to the normative framework most

directly implicated in the experimental results: *sentimentalism*. In the sentimentalist tradition—Hume, Smith, and their contemporary heirs—*moral evaluation arises from patterns of affective resonance*. Moral judgment is not primarily the deliverance of reason nor the outcome of consequence-calculation, but the structured responsiveness of an agent's affective system to features of the social world [245, 323, 324, 325].

Within the evaluative-topological framework developed in this thesis, sentimentalism can be reconstructed as an **affective vector field**:

$$\mathbf{A}(x) : \mathcal{X} \to \mathbb{R}^n$$

where $\mathcal{X}$ is the space of perceived states and $\mathbf{A}(x)$ encodes the direction and magnitude of affective forces—empathic pull, aversive push, compassion, indignation, warmth, or distress. Moral trajectories emerge from the integration of these affective vectors with attentional, inferential, and regulatory processes.

This reconstruction is not metaphorical. It is empirically realised in the experiment: the robot's presence attenuates donation behaviour *by dampening the affective vector field*, especially within ecological profiles where empathy, warmth, or affective sensitivity ordinarily serve as the primary drivers of moral salience.

8.8.1 The Source of Normativity: Sentiment as the Basis of Moral Appraisal

For sentimentalists, normativity originates in the *patterns of affective response* that constitute the human capacity for moral perception. Hume's claim that moral distinctions are "more properly felt than judged" [245] is often caricatured; yet properly interpreted, it captures a structural truth about moral cognition: affective resonance is the primary medium through which agents register the moral significance of others.

This maps directly onto the cognitive LoA articulated in the Morality Primer: affective tagging (amygdala, insula), empathic resonance (mPFC–TPJ circuit), and rapid harm appraisal provide the initial topological curvature from which moral trajectories originate.

Where deontology imposes constraints and consequentialism imposes gradients, sentimentalism specifies the *affective geometry* of the evaluative field: how warmth draws the agent toward prosocial action, how distress aversion or fear repel them, and how empathic concern shapes the felt moral landscape.

8.8.2 Mode of Evaluation: Affective Resonance as Moral Metric

In the sentimentalist reconstruction, the mode of evaluation is grounded in:

- **empathic responsiveness** to others' welfare,
- **reactive attitudes** such as guilt, indignation, gratitude, and resentment,
- **interpersonal attunement** through shared affective states,
- **warmth, sociability, and affiliative motivation**.

These components map precisely onto the **Prosocial–Empathic / Warm–Sociable ecology**. For individuals in this cluster, moral relevance is primarily affective: moral cues are not merely recognised but *felt*, and prosocial donation emerges from empathic attunement to the beneficiary.

Thus, sentimentalist normativity aligns almost point-for-point with the evaluative topology of Cluster *Prosocial–Empathic*. If moral action is the integral of affective forces across the evaluative field, then anything that diminishes the amplitude of $\mathbf{A}(x)$ will proportionally diminish prosocial behaviour.

8.8.3 Action Guidance: Affective Vector Fields and Behavioural Dynamics

The sentimentalist picture becomes formally precise when expressed as a dynamical system:

$$\dot{x} = f(x) + \mathbf{A}(x),$$

where $f(x)$ captures neutral evaluative drift and $\mathbf{A}(x)$ represents affective forces.

Synthetic presence enters the system as a deformation operator:

$$\dot{x}' = f(x) + \mathbf{A}(x) + \delta\mathbf{A}(x; \mathscr{R}),$$

where $\delta\mathbf{A}(x; \mathscr{R})$ is a vector field that attenuates, displaces, or reorients affective flow.

This model captures the experimental results with exceptional fidelity:

- In the **Prosocial–Empathic ecology**, $\delta\mathbf{A}$ significantly dampens empathic activation, flattening the trajectory toward donation.
- In the **Emotionally Reactive ecology**, $\delta\mathbf{A}$ destabilises an already volatile field, producing the strongest behavioural perturbation.
- In the **Analytical–Structured ecology**, $\delta\mathbf{A}$ is comparatively weak; affective forces are not the dominant drivers of action, so the robot's dampening effect is limited.

In short, the robot modulates the moral field by *reducing the affective curvature* that ordinarily drives prosocial behaviour—a textbook sentimentalist effect.

8.8.4 Contrast with Machine Ethics: The Blind Spot of Affective Architecture

Classical Machine Ethics commits its deepest conceptual error here. It systematically ignores affective architecture, attempting to:

- replace empathic resonance with rule sets,
- replace moral perception with logical inference,
- replace affective appraisal with propositional justification.

No sentimentalist could make this mistake, because for sentimentalism, affect is neither optional nor ornamental: it is the *substrate* of moral cognition.

Our experiment demonstrates precisely what Machine Ethics ignores: a silent, non-interactive robot can alter human moral behaviour not by violating rules or changing utilities, but by shifting the structure of *affective vectors* that underwrite prosocial responsiveness.

Machine Ethics has no conceptual resources to model this effect. A sentimentalist topology does.

8.8.5 Experimental Realisation: Synthetic Dampening of Empathic Resonance

The key empirical finding of the experiment is that robotic co-presence attenuates donation behaviour even in the presence of a strong empathic cue (the Watching-Eye stimulus). From a sentimentalist perspective, this is best understood as:

$$\delta\mathbf{A}(x;\mathscr{R}) < 0$$

for affectively weighted regions of the evaluative field, *where*:

- x denotes the agent's current evaluative state within the moral field;
- $\mathbf{A}(x)$ is the baseline affective vector field that encodes empathic pull, aversive push, warmth, distress, and related affective forces;
- $\mathscr{R}$ represents the presence of the humanoid robot as an environmental perturbator;
- $\delta\mathbf{A}(x;\mathscr{R})$ is the deformation operator modelling how $\mathscr{R}$ alters the magnitude and direction of affective vectors at x.

In plain terms, the inequality $\delta\mathbf{A}(x;\mathscr{R}) < 0$ states that the robot's co-presence *reduces the strength of the affective forces* (especially empathic resonance) that normally propel the agent toward prosocial action. The perturbation does not reverse moral direction; it *dampens* the affective momentum that would otherwise guide the agent toward donation.

for affectively weighted regions of the evaluative field.

The robot introduces ontological ambiguity—neither fully agentic nor wholly inert—which disrupts affective attunement in two ways:

1. **Affective dilution**: the presence of an ambiguous social other diverts empathic focus away from the child-beneficiary.
2. **Affective deflection**: the robot introduces uncertainty about social meaning, reducing the clarity of empathic pathways.

Both phenomena manifest as measurable differences in the experimental clusters:

- **Prosocial–Empathic**: attenuation is diagnostic of diluted empathic resonance.
- **Emotionally Reactive**: attenuation reflects increased volatility of affective vectors.

- **Analytical–Structured**: attenuation is weaker, because affect is not the primary moral driver.

Thus, sentimentalism provides the most *mechanistically precise* explanation of the perturbation: synthetic presence alters the affective landscape through which moral salience becomes moral action.

Interpretive Synthesis: Sentimentalism and Synthetic Moral Perturbation

The experimental attenuation of prosocial behaviour under robotic co-presence is a paradigmatic sentimentalist phenomenon. Moral action in the Prosocial–Empathic ecology is driven by affective vector fields whose magnitude is reduced by the robot's ambiguous ontology; in the Emotionally Reactive ecology, the same perturbation destabilises an already volatile field; and in the Analytical–Structured ecology, affective dampening has limited influence because the evaluative surface is dominated by structural rather than affective curvature.

This tripartite pattern cannot be captured by rule-based models, logical deduction, or utility maximisation. It requires a framework in which *affective forces are constitutive of moral cognition*. Sentimentalism, reconstructed as a vector-field theory of affective appraisal, therefore provides the most illuminating normative interpretation of the experiment.

By revealing how synthetic presence modulates affective resonance across latent evaluative ecologies, the experiment demonstrates that moral displacement is not a failure of principles, nor a miscalculation of outcomes, but a deformation of the affective topology through which moral meaning is experienced. This establishes sentimentalist normativity as an indispensable component of any adequate ethical or computational treatment of artificial agents.

8.9 Contractualism, Particularism, and Hybrid Normative Models

The preceding sections developed deontological, consequentialist, and virtue-theoretic structures as topological configurations within the evaluative field. To complete the normative landscape relevant to this thesis, we now introduce three additional frameworks—*contractualism*, *particularism*, and *hybrid or pluralist models*. These theories are reconstructed briefly but precisely, preserving their philosophical integrity while integrating them into the LoA discipline and the evaluative-topological architecture that anchors both the theory and the experiment.

The motivation for introducing these additional models is twofold.

First, they represent influential alternatives to the classical triad of deontology, consequentialism, and virtue ethics. Contractualist theories such as Scanlon's place justificatory relations at the centre of moral evaluation [44]; particularist and perceptual approaches emphasise context-sensitive moral salience rather than rule-governed invariants [93, 89]; and pluralist or hybrid accounts highlight the inherently multidimensional structure of practical reason [65].

Second, these frameworks illuminate aspects of the experimental data that are not

recoverable from topological invariants, value gradients, or dispositional attractors alone. Deontological rules struggle to accommodate dilemmatic or context-dependent cases [326], outcome-based models fail to capture the intuitive and affective dynamics documented in empirical moral psychology [58], and virtue-theoretic attractors are limited by the instability of global traits [65]. By contrast, theories grounded in justifiability, contextual salience, and multidimensional normative interaction provide precisely the structural resources needed to interpret how synthetic presence modulates prosocial behaviour. Empirical studies show that minimal cues of social evaluation—including robotic presence, perceived agency, or merely the appearance of watching eyes—systematically shift cooperative and moral behaviour [137, 327, 273, 136]. These phenomena demand a normative-cognitive model capable of accommodating justification pressures, situational salience, and relational moral dynamics—dimensions captured by these alternative frameworks.

A further reason for reconstructing these theories is philosophical and pedagogical rather than merely instrumental. Any comprehensive treatment of normative foundations—particularly one that aims to integrate ethics with empirical findings and computational modelling—must follow the established structure of the discipline. Contractualism, particularism, and pluralist models represent canonical branches of ethical theory, and omitting them would not only break with the methodological tradition of moral philosophy but would deprive the reader of the conceptual resources required to situate the thesis within the broader normative landscape. Their inclusion therefore serves a dual purpose: it preserves continuity with the philosophical canon, and it ensures that the interpretive tools deployed in analysing the experiment are grounded in a complete and pedagogically robust reconstruction of the field. In short, without these frameworks, the chapter would lack both scholarly completeness and exegetical depth; with them, the reader is equipped to understand how the experimental findings resonate across the full range of contemporary normative theory.

8.9.1 Contractualism: Moral Claims as Justification-Equilibria

Contractualism, most prominently articulated by Scanlon [44], grounds moral rightness in the principle that actions must be justifiable to others on grounds that no one could reasonably reject. This account locates the *source of normativity* not in rules, consequences, or character, but in the structure of interpersonal justification.

Within the LoA structure adopted earlier, contractualism operates at the reflective normative level. It specifies the conditions under which agents can regard themselves as mutually accountable. However, it also has cognitive implications: establishing justifiability requires a sensitivity to others' claims, expectations, and burdens, which in turn depends on social perception and empathic attunement.

Topological Interpretation. Contractualism can be conceptualised as defining regions of justificatory equilibrium within the evaluative field—zones in which an action can withstand the test of mutual recognisability and reasonable non-rejection. On Scanlon's account, moral principles are valid only insofar as they can be justified to others as part of a shared moral relationship [44]; similarly,

Strawson's analysis of reactive attitudes shows that moral assessment presupposes an interpersonal standpoint in which agents acknowledge one another as answerable participants [237]. These equilibria remain stable only when agents register the presence and perspective of others, since the very structure of contractualist judgment requires a perceived field of accountability.

Synthetic presence therefore interacts with contractualist structure in a distinctive way. A minimal cue of observation—such as a watching-eye stimulus—typically increases the salience of interpersonal accountability and strengthens cooperative norms [327, 273]. However, a humanoid robot, being perceptually social yet ontologically ambiguous, perturbs this social-evaluative field. Empirical work shows that robots can elicit social facilitation effects [137] while simultaneously failing to stably occupy the interpersonal roles through which moral demands are ordinarily mediated [136]. The result is a displacement or dilution of the implicit sense of being under another's evaluative regard, thereby disrupting the justificatory equilibrium that contractualism presupposes.

Relevance to Experimental Findings. Contractualism helps explain why the Prosocial–Empathic cluster exhibited the strongest attenuation in the presence of the humanoid robot. Contractualist moral cognition is grounded in the demand that one's actions be justifiable to others under conditions of mutual recognisability [44, 237]. Individuals high in prosociality and empathy are especially sensitive to this interpersonal dimension: their behaviour is strongly modulated by cues of accountability and evaluative regard [**?**]. Under normal circumstances, the Watching-Eye stimulus enhances this perceived mutual accountability, consistent with findings that minimal social-evaluative cues increase cooperation and generosity [327, 273].

The humanoid robot, however, introduces a distinctive perturbation to the justificatory field. Robots can trigger social cognition and elicit affective responses, yet they occupy an ambiguous interpersonal category—they are perceptually social but not reliably recognised as members of the moral community [136, 328]. Empirical studies show that such synthetic presence can both facilitate and destabilise social behaviour [137]. In this context, the robot displaces the implicit sense of being under the evaluative regard of others, thereby weakening justificatory resonance and reducing donation.

Contractualism therefore interprets the observed displacement effect not as a change in underlying preference or a failure of duty, but as a deformation of the justificatory field: a disruption of the interpersonal conditions under which reasons become mutually recognisable and moral motivations are sustained.

8.9.2 Moral Particularism: Contextual Salience and the Fragmented Topology of Reasons

Moral particularism rejects the idea that morality is governed by fixed principles, boundaries, or stable evaluative gradients. On the particularist view, the moral relevance of a consideration is wholly context-dependent: a feature that counts in favour of an action in one case may count against it in another, and reasons possess no invariant valence [92]. This holism of reasons is closely aligned with McDowell's account of moral perception, in which the salience of a consideration

emerges from its role within a concrete situation rather than from any codifiable general rule [93]. Related work in moral epistemology likewise emphasises that the moral field is shaped by context-sensitive patterns of attention and evaluative uptake [89].

Within an evaluative-topological framework, particularism corresponds to a landscape devoid of global structure. Instead of stable invariants or fixed gradients, the field consists solely of local salience contours whose shape shifts with variations in context, attention, or affect. Empirical work in moral psychology supports this characterisation: affective intuitions, perceptual cues, and distributed cognitive processes dynamically modulate which features of a situation are experienced as morally salient [88, 58, 59]. On this view, the evaluative field is fragmented and constantly reconfigured by situational parameters, making moral appraisal an exercise in context-sensitive responsiveness rather than rule-governed inference.

Synthetic Perturbation Under Particularism. If moral salience is locally determined, then synthetic presence need not override a stable evaluative map—there may be no stable map to override. Instead, the robot alters the immediate pattern of salience in the environment.

Moral relevance shifts with contextual detail, perceptual attention, and affective orientation; reasons have no invariant valence [92, 93, 89, 88]. On this view, the evaluative field is not globally structured but dynamically reconstructed from moment to moment as agents engage with their environment.

Synthetic presence therefore alters moral appraisal not by displacing a fixed evaluative configuration but by reshaping the local pattern of salience. Watching-eye cues, for example, immediately increase the accessibility of accountability norms [327], while the presence of a humanoid robot modifies attention, affect, and perceived social agency in more ambiguous ways [137, 136, 328]. What changes is not a stable moral map but the salience geometry that determines which features of the situation come to the fore. In a locally structured evaluative landscape, such perturbations directly influence moral appraisal by shifting which considerations are taken to matter.

This explains why the Emotionally Reactive cluster remained largely invariant in the experiment: their evaluative field is already highly sensitive to situational micro-variations; the robot adds noise, but not disruption relative to their already-fluid topology.

For the Prosocial–Empathic cluster, particularism illuminates a distinct mechanism. Because moral salience is locally assembled rather than globally fixed, the introduction of a humanoid robot reorganises the initial salience hierarchy in the scene. Findings from Social Signal Processing demonstrate that socially meaningful agents exert strong bottom-up pressure on attentional allocation, reshaping which cues are processed first and with what priority [131, 134]. In HRI, even minimal humanoid cues have been shown to redirect gaze, amplify social relevance, and restructure the perceptual field through which subsequent evaluation occurs [329, 330, 331]. Psychological studies similarly show that agentive or emotionally charged stimuli modulate attentional capture and suppress competing social cues [332, 180, 333].

For individuals high in prosociality and empathy, the Watching-Eye stimulus typically heightens interpersonal accountability and enhances empathic attunement. However, the robot's ambiguous interpersonal status—neither fully social nor fully non-social—introduces a conflict in salience that overshadows the eye cue. The result is a weakening of empathic resonance with the expected evaluative signal, producing the attenuated prosocial output observed in the experiments. This interpretation aligns with philosophical accounts of moral perception in which what becomes salient first, and how long it remains salient, is constitutive of the evaluative episode itself [93, 324].

8.9.3 Hybrid and Pluralist Models: Multidimensional Topologies

Hybrid or pluralist normative theories—from Ross's account of irreducible prima facie duties [87] to contemporary value pluralism [183]—maintain that normativity is generated by multiple independent evaluative sources. On this view, moral assessment is not grounded solely in duty, utility, or virtue, but arises from an interplay among distinct kinds of considerations: deontic constraints, outcome-based reasons, character-based appraisals, relational obligations, and contextual factors all exert normative force [334, 335, 158, 44].

In topological terms, pluralism corresponds to a multi-dimensional evaluative manifold. Rather than a single moral axis, the evaluative space contains intersecting gradients, constraints, and dispositional attractors that jointly shape moral judgment. Psychological and neurocognitive models of moral cognition reinforce this interpretation: affective intuitions, rule-based processes, and outcome-tracking mechanisms operate semi-independently and interact dynamically [88, 58, 111]. Moral judgment, on this pluralist understanding, involves navigating a field whose geometry reflects the heterogeneity of moral reasons, none of which dominates the space entirely.

Why Pluralism Fits the Experiment. The experimental findings align naturally with a pluralist topology:

- The Watching-Eye cue activates deontic expectations (being observed).
- Prosocial donation expresses consequentialist gradients (benefit to others).
- Cluster differences reflect dispositional factors (virtue-theoretic structure).
- The robot's ontology refracts social meaning (contractualist relevance).
- Synthetic perturbation shifts local salience (particularist sensitivity).

No single normative theory fully explains the displacement effect observed in the experiment; the phenomenon does not map cleanly onto duty-based invariants, outcome gradients, virtue-theoretic dispositions, or empathy-driven models alone [336, 326, 58, 65, 333]. Instead, the attenuation of prosocial behaviour in the presence of a humanoid robot appears to arise from a reweighting across multiple normative dimensions simultaneously. This is precisely what a pluralist topological model predicts. If moral judgment is guided by a manifold of intersecting evaluative gradients—deontic constraints, empathic pull, reputational expectations, contextual norms, and outcome-based considerations—then perturbing the

structure of the environment can alter the geometry of this manifold as a whole [87, 183, 111].

The experimental findings provide empirical support for this view. The robot's mere presence displaced prosocial action across participants irrespective of dispositional differences: personality traits, empathizing and systemizing profiles, and even latent psychometric clusters failed to moderate the effect. This indicates that the perturbation operates not at the level of individual evaluative tendencies but at the level of the evaluative field itself. The robot's perceptual salience and ontological ambiguity modulate several normative gradients at once—attenuating empathic resonance, altering implicit social expectations, and shifting the perceived normative demand of the situation [136, 255, 331, 328, 137, 288]. In particular, the Watching-Eye cue's typical facilitation of prosocial behaviour is dampened by competing or ambiguous social cues, consistent with findings on accountability cues, attentional capture, and salience competition in social signal processing and psychology [327, 273, 180, 333, 131, 134]. Rather than reinforcing or suppressing any single source of moral motivation, the robot reconfigures the topology within which diverse moral reasons are weighed and integrated.

In this respect, the experiment does more than illustrate a behavioural effect: it offers empirical evidence for the central insight of normative pluralism—that moral judgment is sensitive to the configuration of multiple evaluative dimensions, any of which may be displaced by contextual perturbation [87, 183, 44]. The displacement effect observed here thus constitutes a concrete instantiation of pluralist topology: an environmental shift that alters the manifold of moral reasons as a whole rather than modulating a single axis of moral evaluation.

It is important to emphasise that the field-level displacement effect demonstrated in the experiment does not contradict the presence of stable dispositional differences identified through our clustering analysis. The three psychometric clusters reflect distinct starting positions within the evaluative manifold—different dispositional orientations that shape how individuals ordinarily navigate moral contexts. However, the robotic perturbation operated not on these dispositional baselines but on the shared topological structure of the evaluative field itself. This is why all clusters, despite their psychological differences, exhibited the same directional attenuation in prosocial action. In pluralist topological terms, the robot alters the geometry of the manifold as a whole rather than modulating any single dispositional gradient. The cluster analysis and the displacement effect thus describe two complementary layers of moral cognition: stable trait-like orientations, and a context-sensitive evaluative field capable of being globally reshaped by environmental factors such as synthetic social presence.

8.9.4 Integrative Ethical Interpretation of the Experimental Findings

Bringing the reconstructed frameworks together, we can now articulate the ethical significance of the experimental results in a way that reflects both the normative pluralism developed earlier and the dual-layer structure of moral cognition revealed by the empirical findings. The donation attenuation produced by robotic presence does not reflect the modulation of a single moral principle, evaluative dimension, or dispositional trait. Rather, it arises from a global perturbation of

the evaluative field in which diverse moral reasons are ordinarily weighed and integrated.

1. **From a deontological perspective**, the robot weakens the felt presence of a morally relevant observer and thereby attenuates the sense of duty-oriented accountability that the Watching-Eye cue is designed to amplify. The displacement effect represents a disruption of implicit normative expectations rather than a violation of explicit moral rules.

2. **From a consequentialist perspective**, the robot alters the perceived payoff structure of helping behaviour by flattening the social-evaluative gradient. The expected "return" of prosocial action—whether reputational, emotional, or anticipatory—becomes less sharply defined, shifting the cost–benefit landscape in a way that depresses altruistic output.

3. **From a virtue-ethical perspective**, the perturbation does not target trait-dependent motivational tendencies directly. Rather, it reveals that even robust dispositional architectures (as captured in the three psychometric clusters) can be globally displaced by contextual features that reshape the evaluative field within which character expresses itself. The fact that all clusters show the same behavioural shift indicates that virtue is not a self-contained driver of action, but a gradient subject to field-level modulation.

4. **From a contractualist perspective**, the robot disrupts the local justificatory equilibrium by diminishing the perceived presence of agents to whom reasons are owed. The justificatory landscape becomes noisier and less structured, thereby reducing the motivational force of the requirement to "act in ways that others could not reasonably reject."

5. **From a particularist perspective**, the robot modifies the fine-grained salience structure of the environment, reconfiguring which contextual features become normatively operative. The eyes cue remains physically present, but its moral traction is displaced by a new source of salience—an ambiguous agent whose social meaning is not yet assimilated into the participant's normative schema.

6. **From a pluralist-topological perspective**, the findings are precisely what we should expect when multiple normative gradients interact with a global perturbation to social meaning. The donation attenuation is not the suppression of a single moral principle; it is the displacement of the evaluative manifold itself. This explains why no single theory—deontological, consequentialist, virtue-based, contractualist, or particularist—captures the full phenomenon, and why the effect persists across dispositional clusters.

Taken together, these interpretations converge on a unified thesis:

Integrative Conclusion: The Ethical Signature of Moral Displacement

The presence of a humanoid robot reshapes the multi-dimensional evaluative topology through which moral salience becomes action. This perturbation operates at the level of the evaluative field, modulating deontic expectations, consequentialist gradients, dispositional attractors, justificatory relations, and contextual salience structures simultaneously. No monolithic ethical framework captures the phenomenon. The experimental results therefore vindicate a pluralist, topological, empirically grounded account of moral cognition—one that reveals how synthetic agents can globally displace moral evaluation in ways systematically overlooked by classical Machine Ethics.

By reconstructing the major normative theories through LoA discipline and embedding them within a topologically structured model of moral cognition, this chapter provides the conceptual architecture required to understand the ethical significance of synthetic moral perturbation. The experiment that follows empirically demonstrates how such perturbation manifests as a field-level displacement effect, thereby linking the normative, psychological, and computational analyses into a unified account of how synthetic agents influence moral behaviour.

9. General Discussion and Theoretical Integration

9.1 Introduction: Why the Experiment Requires a Structural Interpretation

The preceding chapters developed three interconnected strands: (i) a cognitive–affective account of moral judgment, (ii) a normative–philosophical reconstruction of ethical theory through the lenses of Level-of-Abstraction discipline and evaluative topology, and (iii) an empirical demonstration that robotic copresence systematically attenuates prosocial donation under morally salient conditions.

Before turning to the integrative task, it is necessary to articulate the higher-order insight guiding the trajectory of this thesis. Situated within the cognitive, philosophical, and formal analyses of the preceding chapters, the empirical study indicates that *moral decision-making is, at root, a practical phenomenon*, grounded in the structures of agency and practical reason [157, 158, 84, 89, 35]. Moral events are not abstract judgements suspended in conceptual space; they are situated transitions from perception to action embedded in a socially organised environment, consistent with empirical models that treat moral cognition as perceptual, affective, and socially modulated [88, 57, 53, 50, 249]. Because such events culminate in observable behavioural outputs, they are empirically tractable and available to systematic measurement and analysis [279, 280, 278]. Their structural and methodological precision is rarely recognised in the prevailing discourse of Machine Ethics and Computational Morality, which has long been criticised for its limited integration of empirical findings [14, 123, 337, 4].

This sequence is methodologically significant. Across both philosophy and moral psychology, ethical inquiry typically proceeds not by legislating the quality of actions from a priori first principles, but by beginning with the existence of *moral events* themselves—episodes in which agents respond to cues, saliences, and social affordances—and then seeking theoretical structures that best explain these patterns of behaviour [35, 89, 105, 60, 155]. This bottom-up orientation stands in sharp contrast to much of the historical trajectory of Machine Ethics, which has principally advanced top-down models that attempt to encode or implement normative theories prior to securing an empirical understanding of how moral cognition unfolds in practice.

A large body of Machine Ethics scholarship exemplifies this top-down, normative-first orientation. Early and influential work sought to engineer explicit ethical rules or principles for artificial agents [14, 123, 16], often drawing upon deontological, utilitarian, or virtue-theoretic frameworks whose normative structure was taken as directly implementable in computational systems [122, 338, 339, 340, 341, 342, 343]. Subsequent developments reinforced this tendency by constructing logical architectures intended to represent moral constraints, permissibility conditions, or value hierarchies independently of empirical models of human moral

agency [121, 344, 345, 346, 347, 348]. Even approaches motivated by psychological plausibility, such as computational models of ethical reasoning [127, 128, 349], largely inherit the same structural assumption that normative content can be specified in advance of empirical measurement.

Critiques of this methodological inversion are now widespread. Authors working within both ethics of AI and social-robotics research argue that designing moral agents without grounding in empirical evidence about cognition, affect, social interaction, or developmental patterns of moral behaviour is epistemically unstable and risks constructing systems whose 'moral' outputs lack psychological validity [337, 350, 351, 352, 4, 13]. On these accounts, moral behaviour cannot be treated as an externally specifiable target for implementation; rather, it emerges from structured interactions among cognitive, affective, embodied, and social-signalling processes [82, 50, 353, 131, 249]. These processes must therefore be empirically characterised before any attempt at normative codification. Only through such empirically informed grounding can normative theory enter the analysis in a methodologically stable and scientifically responsible manner.

The present work therefore advances a methodological reversal. It shows that moral salience, moral displacement, and the perturbation of prosocial behaviour are empirically measurable phenomena that *must* be mapped before being codified, an approach supported by behavioural studies of attentional and prosocial modulation [232, 140, 142, 58, 57, 354]. Because these phenomena are embedded within attentional, affective, and dispositional architectures, they admit rigorous experimental design, statistical modelling, and formal reconstruction [279, 280, **?**]. Accordingly, the experimental study is not an auxiliary illustration but the epistemic anchor of the thesis. Only once the structure of moral events is empirically established can normative theory enter the analysis—precisely the reverse of the methodological sequence characteristic of Machine Ethics, normative-first LLM evaluation, and much of Affective Computing [14, 123, 10, 9, 7, 355, 356].

The task of the present chapter is not to repeat these analyses, but to integrate them. It offers a theoretical synthesis that explains *why* the experimental effect occurs, *what* its ethical significance is, and *how* it reshapes the methodological landscape for research in Human–Robot Interaction, moral psychology, and the emerging field of Computational Morality.

In this sense, the experiment is not an isolated behavioural result but a *probe* into the architecture of moral cognition. The observed attenuation of prosocial behaviour is theoretically meaningful only when interpreted through the structures developed earlier: dual-process architectures, the Social Intuitionist Model, evaluative topology, and the reconstructed normative frameworks of deontology, consequentialism, virtue ethics, sentimentalism, contractualism, and particularism. The present chapter therefore provides a synoptic interpretation in which the behavioural signature revealed by the data becomes a lens through which the nature of moral cognition—and its vulnerability to perturbation—is rendered theoretically transparent.

9.1.1 From Behaviour to Structure: Why a Higher-Level Interpretation is Required

The experimental paradigm—Watching-Eye moral cue embedded within a silent synthetic presence—does not merely generate a difference in donation behaviour; it reveals a deformation of the evaluative field that links moral salience to action. Classical interpretations of donation differences (e.g., generosity, altruism, compliance) lack the conceptual resources to capture this phenomenon. A purely behavioural description would record that participants donated less in the Robot condition, with the Prosocial–Empathic cluster showing the numerically steepest decline. But such a description omits the structural logic that makes the result scientifically and philosophically significant.

The central claim developed throughout the thesis is that *moral behaviour is not invariant under changes to the perceptual–social environment.* The robot's presence does not overwrite moral norms nor impose new ones; instead, it modifies the cognitive–affective conditions under which evaluative forces act. It shifts attentional allocation, alters affective resonance, and modifies the perceived sociality of the space. In topological terms, the robot introduces a perturbation γ_R that deforms the curvature of the evaluative manifold, thereby weakening the salience gradient induced by the Watching-Eye stimulus.

A simple behavioural difference thus reflects a deeper structural transformation in the evaluative field. As demonstrated by the regression models and Bayesian estimation, the attenuation effect was uniform in direction across participants, indicating that the perturbation introduced by the robot operates at the field level rather than through trait-specific pathways. Yet this uniformity does not imply psychological homogeneity.The PCA–k-means clustering revealed three coherent dispositional ecologies—distinct configurations of empathic resonance, affective volatility, and structural–analytical processing. These ecologies are consistent with the established dimensions of empathizing and systemizing [357], personality variation captured by the BFI-10 [253], and broader accounts of moral-psychological "ecologies" that organise evaluative processing [249, 88]:

- the **Emotionally Reactive / Low-Structure Profile**,
- the **Prosocial–Empathic / Warm–Sociable Profile**,
- the **Analytical–Structured / High-Systemizing Profile**.

These clusters instantiate different evaluative topologies—distinct attractor formations, sensitivities to perceptual and affective salience, and pathways of modulation—consistent with multidimensional models of affective valuation and moral cognition [111, 153, 143, 88]. Within this framework, the Prosocial–Empathic cluster exhibits the steepest affective gradients and the strongest baseline responsiveness to Watching-Eye cues. This ecological structure aligns with theoretical expectations: Watching-Eyes primes amplify empathic accountability [327, 273], and empathic resonance is known to be highly sensitive to contextual modulation [333].

That this cluster nevertheless showed the same directional attenuation as the others is therefore theoretically significant. Rather than reflecting a trait-dependent

shift, the humanoid robot's ambiguous social presence perturbs the salience structure itself, weakening the amplification mechanisms on which empathic ecologies depend [136]. In other words, the perturbation operates *upstream* of individual dispositional pathways: it modifies the evaluative field within which those pathways are embedded. The displacement observed in the experiment is thus best understood as a *field-level suppression of moral salience*, overriding the ordinarily divergent dispositional trajectories that shape prosocial behaviour.

Ethical Interpretation: Why the Attenuation Matters Normatively. The ethical significance of this finding becomes visible only when the result is interpreted through the reconstructed normative frameworks developed in Chapter 8. Each theory identifies a different locus of normative structure, and each provides a distinct—yet convergent—reading of the deformation caused by $\mathscr{R}$:

- *Deontological perspective.* The Watching-Eye cue implicitly invokes deontic expectations of reciprocity, fairness, and beneficence. The robot's presence attenuates donation precisely by dulling this sensitivity. Normatively, this appears as a disruption of the agent's capacity to track *ought-constraints* in the environment—an interference with the cognitive substrate on which deontic responsiveness relies.
- *Consequentialist perspective.* The moral field includes gradients of anticipated social evaluation. Watching-Eye cues steepen these gradients; synthetic presence flattens them. The robot therefore functions as a *gradient-suppressor*, reducing the perceived payoff of prosocial action. In topological terms: it alters the vector field governing welfare-oriented trajectories.
- *Virtue-ethical perspective.* The three clusters correspond to differing dispositional configurations. The strongest attenuation occurring within the Prosocial–Empathic cluster implies that the robot disrupts precisely those virtues—empathy, warmth, prosocial orientation—that ordinarily stabilise prosocial attractors. The perturbation thus interacts with *character topology* rather than bypassing it.
- *Sentimentalist (Humean) perspective.* The attenuation reflects a dampening of empathic vector fields: $\delta\mathbf{A}(x;\mathscr{R}) < 0$. The robot selectively reduces affective resonance with the Watching-Eye cue. Normatively, this implies that the moral valence of the situation is felt less intensely, weakening the motivational energy required for prosocial action.
- *Contractualist perspective.* The moral event of donation under observation involves tacit justifiability relations: "What could reasonably be expected of me in the eyes of others?" The ambiguous presence of a synthetic observer destabilises this justificatory equilibrium. The subject no longer clearly apprehends *to whom* justifiability is owed.
- *Particularist perspective.* Moral appraisal depends on local saliences. The robot modifies the salience landscape: the morally relevant cue (the child in need) becomes less perceptually dominant. Thus, the attenuation is interpreted as a shift in the pattern of reasons that obtain in this particular context.

LoA Interpretation: Why the Perturbation Occurs at the Wrong Level for Machine Ethics. Floridi's Level-of-Abstraction analysis clarifies the structural error revealed by the experiment. The attenuation does *not* occur at the normative LoA (where duties, values, or justifiability live), but at the cognitive–affective LoA (where salience, resonance, and attention are regulated). Machine Ethics traditionally operates at the wrong LoA: it attempts to implement high-level normative constructs while ignoring the low-level substrates on which moral responsiveness depends.

The experiment shows why this is untenable. Ethical responsiveness is mediated by:

- attentional allocation (Who or what do I notice?)
- affective resonance (What emotional weight does this carry?)
- perceived social ontology (Who counts as the observer?)
- dispositional pathways (How does my cognitive ecology integrate this cue?)

Synthetic presence perturbs all of these upstream mechanisms. Thus, even perfect normative reasoning at a reflective LoA cannot salvage moral action when the lower-level architecture of moral cognition has been deformed. In Floridi's terms:

> *Normative correctness is orthogonal to causal efficacy. A system may know what is right and yet fail to act rightly if the cognitive LoA is perturbed.*

Integrative Insight. The field-level suppression observed in the experiment therefore reveals a principle of broad ethical and psychological importance:

> *Moral failure under synthetic presence is not a failure of principle but a failure of salience. Ethical norms lose their grip not because agents reject them, but because the evaluative machinery that normally brings them to bear is disrupted.*

This insight is the conceptual hinge on which the whole thesis turns. It unifies:

- the cognitive architecture (moral judgments arise from salience → appraisal → integration),
- the topological formalism (moral cues define gradients and attractors),
- the normative frameworks (moral theories describe different structural aspects of the evaluative field),
- and the empirical results (synthetic presence suppresses these structures at the field level).

With these interpretive tools in place, we can now proceed to the cluster-by-cluster integrative analysis that further refines the ethical and cognitive significance of the experimental findings.

9.1.2 Why This Chapter Cannot Be Pure "Discussion" in the Conventional Sense

Traditional discussion chapters in empirical theses typically emphasise methodological limitations, alternative interpretations, and directions for future work. While such elements remain relevant here, they are insufficient for the present project. The experiment developed in this thesis sits at the intersection of cognitive science, social robotics, computational modelling, and normative ethics. The behavioural effect it reveals—reduced prosocial donation under synthetic copresence—is only the observable trace of a deeper structural transformation: a perturbation of the evaluative machinery through which agents convert moral salience into action. Because this transformation engages multiple theoretical layers—cognitive–affective processing, dispositional topology, normative interpretation, and Level-of-Abstraction analysis—a standard discussion section cannot capture its full conceptual significance. What is needed instead is a structural synthesis that explains not merely *what* happened, but *why* it happened and *what it reveals* about the nature of moral cognition and its vulnerability to synthetic perturbation.

To articulate this phenomenon requires a conceptual integration that cannot be confined to standard "discussion" categories. Instead, the chapter must synthesise:

1. the **cognitive architecture** (dual-process, SIM, dynamic integration);
2. the **evaluative geometry** (topology, curvature, gradient flow);
3. the **normative reconstruction** (deontic invariants, consequentialist gradients, dispositional attractors, sentimentalist vector fields, contractualist justificatory structure, and particularist salience responsiveness);
4. and the **empirical structure** of the data (cluster-specific susceptibility, Bayesian attenuation, topological deformation of the Watching-Eye effect).

The present chapter therefore functions as an *interpretive pivot*: it translates the empirical findings into philosophical insight, and reinterprets philosophical frameworks in light of empirical constraints.

9.1.3 A Structural Reading of the Core Experimental Result

The empirical pattern can be summarised as follows:

- The humanoid robot NAO is perceptually salient but ontologically ambiguous.
- The Watching-Eye cue ordinarily induces an empathic salience gradient that increases donation.
- The robot introduces a perturbation γ_R that competes with, and partially overrides, this empathic amplification.
- Attenuation is strongest in the Prosocial–Empathic cluster, weaker in the Analytical–Structured cluster, and statistically negligible in the Emotionally Reactive cluster.

Interpreted through the cognitive framework developed earlier, this pattern shows that moral appraisal begins with intuitive and affective resonance [88, 58]. Synthetic presence disrupts this resonance by altering attention, salience, and perceived sociality [180, 333, 136, 137]. Different dispositional structures absorb this disruption in systematically different ways, consistent with established dimensions of empathizing, systemizing, and moral-schema variability [357, 59]. The resulting behavioural output reflects not a change in moral principle, but a deformation of the evaluative field.

Interpreted through the normative framework, the same pattern yields multiple structurally coherent readings:

- a **deontological reading**: synthetic presence weakens the implicit deontic expectations cued by the Watching-Eye stimulus [117];
- a **consequentialist reading**: synthetic presence flattens the perceived payoff gradient of helping behaviour [**?**];
- a **virtue-ethical reading**: synthetic presence suppresses prosocial attractors associated with empathic or cooperative dispositions [35];
- a **sentimentalist reading**: synthetic presence dampens empathic vector fields that ordinarily drive prosocial action [324];
- a **contractualist reading**: synthetic presence destabilises the justificatory relations normally activated by social observation [44];
- a **particularist reading**: synthetic presence alters the salience pattern such that the Watching-Eye cue no longer carries the same moral significance [92, 93].

Thus, each normative theory yields a structurally distinct but empirically convergent interpretation. The ethical significance of the experiment lies not in any single framework, but in the *coherent intersection* of all of them: a field-level suppression of moral salience, a deformation of the evaluative topology through which moral meaning becomes action.

9.1.4 Why the Synthetic Presence Effect Matters Beyond the Experiment

The attenuation of moral action under synthetic presence is not merely an interesting behavioural anomaly; it demonstrates a deeper principle: *moral cognition is structurally permeable.* It is sensitive to perturbations that operate below the level of explicit reasoning. It is vulnerable to shifts in perceived social ontology. And it is modulated by affectively weighted cues whose influence is seldom acknowledged in normative theory and almost never incorporated in classical Machine Ethics.

This has far-reaching implications:

1. It challenges the assumption that artificial agents can be designed according to purely deliberative ethical frameworks.

2. It shows that synthetic presence modulates moral behaviour even without action, speech, intent, or agency.
3. It reveals that human–robot environments are *ethically loaded* by virtue of perceptual and affective structure alone.
4. It demands a reconsideration of how artificial systems are situated within the moral ecology of human decision-making.

In short, the experiment demonstrates a fact of philosophical significance: *synthetic agents are not normatively inert.* Their presence, even in silent passivity, can deform the evaluative pathways through which moral salience becomes action.

The remainder of this chapter builds on this foundation. Subsequent sections provide:

- a cluster-by-cluster integrative interpretation,
- a cross-framework normative synthesis,
- a critique of monolithic Machine Ethics,
- a reconstruction of Computational Morality grounded in empirical structure,
- and a final consolidation of the thesis' theoretical contributions.

The goal is not only to interpret the experiment, but to show how the experiment reconfigures the conceptual terrain on which research in moral psychology, HRI, and Machine Ethics must proceed.

9.2 Cluster-by-Cluster Integrative Interpretation

The experimental results demonstrate that robotic co-presence $\mathscr{R}$ induces a uniform directional attenuation of prosocial donation across participants, yet the *structure* of this attenuation differs meaningfully across the three latent cognitive–affective ecologies uncovered in Chapter 7. Because these clusters instantiate distinct evaluative topologies—different attractor formations, salience gradients, affective vector fields, and pathways of regulatory modulation—their differential perturbation under $\mathscr{R}$ offers insight into the architecture of moral cognition and the ethical significance of synthetic presence. What follows is an integrative interpretation weaving together the cognitive, topological, normative, and Level-of-Abstraction (LoA) analyses developed across the thesis.

Emotionally Reactive / Low-Structure Ecology

This ecology exhibits high affective volatility, shallow structural integration, and weak systemizing constraints, consistent with established empathizing–systemizing variability [357]. Its evaluative topology is characterised by *broad, low-gradient attractors*: intuitive responses are strong but unstable; attentional salience fluctuates; and the transition from perception to action is mediated by short-lived affective surges rather than sustained deliberative integration.

To avoid terminological ambiguity, it is useful to clarify what is meant here by *broad, low-gradient attractors* in the evaluative–topological framework. In dynamical-systems terms, an attractor represents a region of the evaluative field $\mathcal{E}$ toward which the system's state x naturally converges [358, 359]. A *broad* attractor denotes a basin of attraction with wide boundaries and weak curvature, meaning that many initial states can enter it but none are strongly pulled toward a particular behavioural endpoint. A *low-gradient* attractor is one in which the magnitude of the evaluative gradient $\|\nabla\mathcal{E}(x)\|$ is small across the basin, implying that movement toward prosocial or antisocial trajectories is governed by shallow motivational forces [360, 361].

In psychological terms, this configuration corresponds to intuitive reactions that are easily triggered yet weakly stabilised: the agent may experience transient affective spikes (e.g., momentary empathy, irritation, or ambivalence) without these signals generating a consistent or directed behavioural tendency. This interpretation is consistent with empirical models of low-coherence affect, affective lability, and unstable salience allocation [362, 363, 364]. Because the evaluative landscape lacks sharply defined slopes, small perturbations—including those introduced by environmental ambiguity—tend not to produce substantial directional change. This explains why the Emotionally Reactive / Low-Structure ecology exhibited behavioural invariance in the experiment: the moral field was already characterised by diffuse attractors and unstable salience dynamics, leaving little structured curvature for $\mathscr{R}$ to deform.

Within such a landscape, the experimentally observed pattern—minimal or noisy attenuation—is theoretically revealing. The Watching-Eye stimulus σ_{WE} generates only a modest prosocial gradient for this cluster [327, 273], and the robot-induced perturbation γ_R cannot significantly deform a field that already lacks curvature:

$$|\nabla\mathcal{E}_{\mathrm{baseline}}| \approx 0 \quad \Rightarrow \quad |\nabla\mathcal{E}_{\mathrm{perturbed}}| \approx 0.$$

At the cognitive LoA, this ecology functions as a near-critical system: its evaluative machinery exhibits little stability and thus provides minimal structural leverage for $\mathscr{R}$ to disrupt. Normatively, this implies that deontic, sentimental, or virtue-theoretic structures exert limited behavioural influence because the underlying evaluative field lacks the curvature to sustain them.

Prosocial–Empathic / Warm–Sociable Ecology

This cluster displays high empathic resonance, strong sensitivity to social cues, and rich affective attractors. Psychological models of empathic processing support this heightened salience responsiveness [180, 333]. Its evaluative topology is steeply sloped: the Watching-Eye cue generates strong upward gradients toward prosocial action [327], mediated by interpersonal appraisal and affective amplification.

The robot's ontological ambiguity [136, 135, 137] perturbs precisely this amplification mechanism. As demonstrated in Chapter 7, the perturbation $\delta\mathcal{E}(x;\mathscr{R})$ acts *upstream*, modifying the salience structure itself:

$$\delta\mathcal{E}(x;\mathscr{R}) < 0, \qquad \delta\mathbf{A}(x;\mathscr{R}) < 0.$$

Because the empathic system depends on affective curvature, flattening the field produces the *largest attenuation* in this ecology despite its strong baseline gradients.

Normatively, this yields a convergent interpretation: deontology registers weakened duty-tracking; consequentialism observes a flattened payoff gradient; virtue ethics identifies destabilised prosocial dispositions; sentimentalism finds dampened empathic force-fields; contractualism diagnoses disrupted justificatory orientation; and particularism detects a shift in which contextual features count as reasons.

Analytical–Structured / High-Systemizing Ecology

This ecology exhibits strong systemizing tendencies and comparatively lower empathizing [357]. Its evaluative topology is governed by structural coherence rather than affective curvature. Here, prosocial action arises from rule-consistency, interpretive stability, and contextually well-defined cues.

The experiment reveals only mild attenuation. The Watching-Eye cue produces modest gradients, while $\mathscr{R}$ introduces representational and social-ontological ambiguity [255], subtly undermining the interpretive regularities on which this ecology relies. The perturbation operates primarily on semantic and predictive structure:

$$\delta\mathcal{E}(x;\mathscr{R}) \approx 0^{-}, \qquad \delta\mathbf{A}(x;\mathscr{R}) \approx 0.$$

At the LoA level, this ecology demonstrates that perturbation need not be affective: synthetic presence also functions as a *semantic disruptor*, altering the representational substrate needed for structured evaluative computation. Normatively, this corresponds to weakened rule-clarity (deontology), distorted outcome-modelling (consequentialism), and destabilised interpretive virtues such as discernment and practical wisdom (virtue ethics).

Integrative Synthesis

Across all three ecologies, a unified conclusion emerges: the humanoid robot operates not through communication, norm expression, or explicit social signalling, but through *topological reconfiguration*. It introduces a perturbation γ_R at the cognitive LoA that:

- suppresses affective gradients in empathic ecologies,
- introduces semantic and predictive ambiguity in analytical ecologies,
- and interacts minimally with shallow attractor fields in reactive ecologies.

Normatively, the attenuation is not a failure of duty, utility estimation, virtue, empathy, or justificatory reasoning. Instead, it represents a *structural displacement of moral salience*. This displacement is invisible to explicit reasoning yet measurable in behaviour and interpretable through evaluative topology.

In this sense, the humanoid robot reveals a property of moral cognition that classical ethical theory and classical Machine Ethics could not predict: *moral responsiveness is field-sensitive*. Normativity becomes action only when the evaluative

field retains its curvature. Perturb the field, and even well-formed dispositions cannot operate normally.

This insight forms the conceptual hinge for the remainder of the General Discussion.

9.3 Global Normative–Topological Synthesis

The final integrative step requires bringing together the three interpretive lenses that structure this thesis: (i) the *topology* of moral cognition, (ii) the *normative frameworks* reconstructed in the Ethical Cognition chapter, and (iii) the *empirical perturbation* revealed by the experiment. The aim is not to select a single normative theory that "best explains" the data, nor to impose a moral verdict on participants' behaviour. Rather, the task is to demonstrate how the experimental findings become theoretically intelligible *only* when analysed at the correct Level of Abstraction (LoA), through a structure-sensitive account of evaluative dynamics.

Moral Behaviour as a Field-Level Phenomenon

Across deontological, consequentialist, virtue-theoretic, sentimentalist, and contractualist frameworks, one structural insight remains invariant: **moral action does not arise from isolated psychological modules or explicit rule execution**. Instead, it emerges from the configuration of the evaluative field—a relational structure shaped by perception, affect, social meaning, habituation, and normative commitments.

The experiment demonstrates that this field is *globally deformable*: a silent humanoid robot, devoid of agency, instruction, or communication, attenuates prosocial behaviour across all dispositional ecologies. This uniform directionality, combined with cluster-specific differences in amplitude, reveals a core computational insight:

The presence of $\mathscr{R}$ acts as a field-level perturbation, not a trait-level driver.

In topological terms, the robot introduces a deformation operator

$$\gamma_R : \mathcal{E} \to \mathcal{E}',$$

which modifies the curvature of the evaluative manifold such that moral salience diffuses more weakly toward prosocial attractors. This accounts for both the global donation reduction and the heterogeneous susceptibility across ecologies.

Deontological, Consequentialist, and Virtue-Ethical Readings of the Perturbation

The experiment's ethical significance becomes transparent when interpreted through the normative frameworks reconstructed earlier:

- **Deontological interpretation:** The Watching-Eye cue implicitly invokes deontic norms of accountability and interpersonal respect. The attenuation

of donation under $\mathscr{R}$ is thus intelligible as a deformation of the agent's sensitivity to these constraints. The robot does not induce norm violation; it *weakens the agent's access* to deontic salience by altering the perceived sociality of the environment.

- **Consequentialist interpretation:** Watching-Eye cues are known to reshape the perceived consequence structure of prosocial acts. The robot's ambiguous presence disrupts this gradient, flattening reputational and affective payoff structures. Donation decreases because the local value landscape is deformed, not because agents become less "ethical."
- **Virtue-ethical interpretation:** The dispositional ecologies uncovered in the clustering analysis map directly onto virtue-ethical accounts of character as a structured, learned sensitivity to moral salience. $\mathscr{R}$ perturbs the field *upstream* of these dispositions, weakening the operative mechanisms of moral perception, especially in the Prosocial–Empathic / Warm–Sociable profile.

Each framework thus provides a different interpretive contour of the same phenomenon. But they converge on one central point: **the perturbation acts on the evaluative field, not on the moral principles themselves**. The agents' normative commitments remain intact; what changes is the salience structure through which those commitments become behaviourally operative.

Sentimentalist, Contractualist, and Particularist Convergence

Sentimentalist theories construe moral judgment as an affective vector field. Under this lens, the robot acts as a dampening force on empathic resonance, decreasing the magnitude of affective gradients required to activate prosocial behaviour. Cluster-specific differences in attenuation severity become intelligible as differences in affective sensitivity and evaluative slope.

Contractualist and justificatory theories interpret the perturbation as a shift in the perceived interpersonal structure of the environment. When $\mathscr{R}$ is present, participants implicitly alter their model of who counts as a moral interlocutor—a phenomenon well-documented in human–robot interaction literature. This recategorisation subtly modifies the justificatory landscape in which prosocial acts acquire meaning.

Particularist and perceptualist theories emphasise moral *attention*. On this view, the robot acts as a competing centre of salience, pulling attentional weight away from the Watching-Eye cue and thereby diluting the moral percept. This aligns precisely with the empirical finding of attenuated donation despite a strong moral prime.

Floridi's Level-of-Abstraction Reading

Floridi's LoA discipline allows us to state the integrative conclusion succinctly:

- At the **cognitive LoA**, the robot perturbs perceptual-affective mechanisms (attention, salience, resonance).

- At the **behavioural LoA**, this perturbation manifests as reduced prosocial action.
- At the **normative LoA**, the agent's ethical commitments remain unchanged, but the pathway by which they become operative is deformed.

This avoids the two characteristic errors of Machine Ethics:

1. treating normative principles as if they were generative psychological operators;
2. treating behavioural shifts as if they were moral judgments.

Integrative Conclusion: Moral Salience, Synthetic Presence, and the Architecture of Agency

Integrative Conclusion: The Ethical Significance of Synthetic Perturbation

The experiment demonstrates that synthetic presence can alter moral action not by introducing new norms or violating existing ones, but by reshaping the evaluative topology through which moral salience acquires behavioural force. Deontological constraints, consequentialist gradients, virtue-theoretic dispositions, sentimental vector fields, and contractualist justificatory demands all converge on the same structural insight: the moral field is deformable. The humanoid robot acts as a perturbation operator γ_R on this field, weakening the pathways that normally lead from moral perception to prosocial action. This field-level deformation explains both the global attenuation effect and the cluster-specific signatures discovered in the experiment. It also reveals a fundamental limitation of classical Machine Ethics: normative content cannot be operationalised without an empirically grounded account of how moral cognition functions within its situational topology. The thesis therefore establishes a new methodological foundation for Computational Morality: synthetic agents must be analysed not merely as potential moral reasoners, but as operators on the moral ecology in which human agency unfolds.

9.4 From the Failure of Machine Ethics to a Reconstruction of Computational Morality

The preceding analyses show that robotic co-presence $\mathscr{R}$ induces a deformation of the evaluative field within which moral salience becomes action. This has direct implications for artificial moral agency and exposes a structural flaw in classical Machine Ethics. Since its inception, Machine Ethics has assumed that moral behaviour can be engineered by encoding ethical principles inside an artificial system—a view explicit in rule-based architectures [14, 124], utilitarian optimisation frameworks [122], virtue-based computational agents [123], and logic-driven decision systems [16, 121]. These approaches presuppose that normative theories function as *implementable specifications*. However, as Floridi's Levels of Abstraction make clear [125, 365], this constitutes a category mistake: normative theories belong to a reflective LoA, whereas moral behaviour emerges at the cognitive LoA

through complex interactions of salience, affect, social signalling, and controlled appraisal.

Moral psychology and cognitive science provide a clear counterpoint to the Machine Ethics assumption. Decades of research show that moral behaviour is not generated by rule execution but by intuitive-affective processes [88], conflict-sensitive valuation systems [58], affective-perceptual mappings [111], and schema-based social cognition [59]. Moral appraisal begins with rapid, pre-reflective resonance shaped by perceptual salience [180], empathic responsiveness [333], and contextual cues. The empirical results of this thesis reinforce these findings: robotic presence modifies salience structures upstream of conscious evaluation, consistent with work showing that synthetic agents alter social perception and norm-related behaviour even in minimal-interaction contexts [137, 136, 135].

Machine Ethics models fail to capture these mechanisms. Deontic architectures presuppose invariant constraints, yet even deontic cues—such as Watching-Eye effects [327, 273]—can be attenuated by the mere presence of a humanoid robot. Utilitarian architectures assume stable value gradients, yet the data show that gradients of perceived social consequence are flattened by ontological ambiguity [137]. Virtue-based systems assume globally stable traits, yet situationist critiques [65] and schema ecologies [59] reveal substantial dispositional heterogeneity; the experiment confirms that dispositional structure alone cannot explain behavioural attenuation. Sentimentalist architectures—which would predict affective resonance as a core driver of moral action—are almost entirely absent from Machine Ethics, despite overwhelming evidence that empathy and affective salience strongly modulate moral behaviour [333, 88].

The methodological failure is thus profound. Classical Machine Ethics implicitly assumes:

$$\text{Normative authority} \Rightarrow \text{Behavioural generation.}$$

This implication is falsified both empirically and theoretically. Normative principles—deontic, consequentialist, virtue-theoretic—do not by themselves generate behaviour, even in humans. Behaviour arises from the evaluative topology within which norms are interpreted. Watching-Eye cues generate deontic *expectations*, but the behavioural manifestation of these expectations is perturbed by γ_R at the level of attention, salience, and affective resonance. A normative rule cannot be enacted when the cognitive–affective substrate enabling its enactment is disrupted.

For these reasons, monolithic Machine Ethics fails. It collapses reflective and cognitive LoAs, ignores the topological structure linking salience to action, neglects the role of affect and social signal processing in moral cognition, and treats moral behaviour as rule-following rather than field-sensitive, dynamically realised evaluation.

9.4.1 Reconstructing Computational Morality: An Empirically Grounded Paradigm

If Machine Ethics fails because it begins with normative theory, the alternative must begin with *empirical structure*. The present thesis advances a methodological reversal:

> **Computational Morality** begins not by encoding principles, but by modelling the cognitive–affective architecture through which moral behaviour is produced and perturbed.

(1) Evaluative Topology as Generative Substrate Moral behaviour emerges from an evaluative manifold shaped by gradients of salience, attractor basins of affective resonance, normative invariants, and dispositional curvature [111]. Robotic presence is formalised as a perturbation operator:

$$\gamma_R : \mathcal{E} \to \mathcal{E}',$$

modifying attentional and affective weights and thereby predicting attenuation of prosocial behaviour without invoking rule-based computation.

(2) Level-of-Abstraction Discipline Normative theories enter as reflective structures operating at the normative LoA [125]. Deontology provides invariants, consequentialism gradients, virtue ethics dispositional metrics, sentimentalism affective vectors, contractualism justificatory equilibria, and particularism context-sensitive modulations. These structures constrain interpretation, not execution.

(3) Dispositional Ecologies as Moral Topologies The PCA–k-means clusters define dispositional geometries that shape evaluative trajectories: *Emotionally Reactive* (broad, shallow attractors), *Prosocial–Empathic* (steep affective gradients), and *Analytical–Structured* (narrow, stable valleys). Synthetic presence perturbs the field upstream of these differences [137, 136], revealing that moral behaviour is topologically sensitive rather than trait-determined.

9.4.2 Computational Morality as a Scientific Research Programme

The reconstructed paradigm transforms the methodological landscape of moral AI. Rather than engineering moral behaviour by encoding principles, *Computational Morality* aims to:

1. model the evaluative field governing moral behaviour;
2. identify perturbation operators introduced by artificial agents;
3. integrate normative theory as reflective constraint rather than behavioural generator;
4. and design artificial systems that stabilise, rather than distort, the evaluative field.

This paradigm extends Social Signal Processing [131, 134] and Affective Computing [133] by adding a normative dimension grounded not in abstract prescription but in empirically measurable topological structure.

In this sense, the robot in the experiment is not an ethical agent but an *evaluative perturbation device*. Its presence reveals the structural sensitivity of human moral cognition. A scientifically responsible programme of moral AI must begin from this insight: artificial agents shape the moral environment long before they act within it.

The next section consolidates these findings into a global synthesis, showing how the normative, cognitive, and topological architectures developed across the thesis converge in a unified model of moral perturbation and ethical interpretation.

9.5 Thesis-Wide Synthesis and Closing Reflections

Across its full argumentative trajectory, this thesis has advanced a single, unified claim: *human moral behaviour is structurally sensitive to the architecture of the perceptual–social environment, and synthetic presence—even when silent and non-sentient—is sufficient to reshape that structure.* This concluding section synthesises the theoretical, empirical, and normative strands developed throughout the work and articulates the implications for computation, moral psychology, and the ethics of artificial agents.

1. Moral Cognition is Field-Sensitive and Structurally Rich

The *Morality Primer* established that moral cognition is a distributed, multilevel, dynamically integrated system. Dual-process models, the Social Intuitionist Model, and empirical findings from social neuroscience converge on a view in which moral appraisal emerges from:

- rapid affective and attentional processes,
- controlled interpretive regulation,
- and an evaluative topology shaped by salience, affective resonance, and contextual cues.

This architecture is not neutral with respect to environmental perturbation. The field in which moral appraisal unfolds has curvature, gradients, attractors, and deformation potentials—all empirically traceable, neurocognitively plausible, and behaviourally measurable.

2. Levels of Abstraction and the Limits of Purely Normative Models

The *Ethical Cognition and Normative Foundations* chapter showed that ethical theory and moral psychology occupy distinct Levels of Abstraction. Normative theories do not function as generative behavioural models; their role is to articulate invariant, justificatory, or virtue-theoretic structures that constrain or interpret behaviour at a reflective LoA.

Machine Ethics has historically collapsed these orders, implementing deontic rules, utility functions, or evaluative labels as if they were cognitive operators. This thesis rejects that methodological inversion. Normative content becomes intelligible only when anchored in empirical structure; without such anchoring, computational morality risks degenerating into symbolic simulation devoid of psychological traction.

3. Empirical Evidence for Synthetic Moral Perturbation

Within this framework, the experiment plays a decisive role. It demonstrates that the presence of a humanoid robot:

- attenuates prosocial donation in a statistically supported manner,
- does so even under a strong moral cue (the Watching-Eye prime),
- and produces a uniform directional displacement across dispositional ecologies, albeit with variation in magnitude.

This attenuation is not reducible to personality differences, response bias, or explicit moral reasoning. The analysis shows that the robot functions as a perturbation operator γ_R that modifies the evaluative field *upstream* of trait-specific and deliberative processes. It acts on the conditions under which moral appraisal acquires behavioural force.

4. Dispositional Ecologies Reveal Structural, Not Idiosyncratic, Perturbation

The clustering analysis established three coherent dispositional ecologies:

- **Emotionally Reactive / Low-Structure**, exhibiting broad low-gradient attractors and high affective volatility;
- **Prosocial–Empathic / Warm–Sociable**, with steep empathic gradients and strong responsiveness to social cues;
- **Analytical–Structured / High-Systemizing**, with narrow, stable attractors shaped by deliberative integration.

Despite their divergent evaluative geometries, all clusters showed the same *direction* of moral displacement. This finding is decisive: it shows that the perturbation is field-level, not agent-level. The robot reshapes the evaluative manifold within which trajectories unfold, rather than interacting with any single cognitive disposition. This is the empirical signature of a *structural perturbator*.

5. Normative Interpretation of Structural Perturbation

The reconstructed normative frameworks illuminate the ethical significance of this empirical result:

- deontologically, γ_R disrupts the recognition of accountability cues implicit in the Watching-Eye stimulus;
- consequentially, it flattens the perceived payoff gradient of beneficence;
- virtuously, it weakens the stabilising force of prosocial dispositions;
- sentimentally, it dampens empathic vector fields that anchor reactive moral emotions;
- contractually, it disrupts justificatory visibility between moral agents;
- particularistically, it shifts the situational salience profile.

These converging interpretations reveal the central structural insight of the thesis: *the robot does not add a new norm; it shifts the evaluative conditions under which norms become behaviourally operative.*

6. Final Position of the Thesis

We may now return to the guiding hypotheses:

H1 — Evaluative Deformation *Confirmed.* The evaluative process f linking perception to action is systematically altered by synthetic presence.

H2 — Synthetic Normativity *Confirmed.* Synthetic agents acquire derivative normative force by altering the field of salience and accountability.

H3 — Synthetic Perturbation of Moral Inference *Confirmed.* The robot refracts the transition from moral appraisal to prosocial behaviour, attenuating the expressive force of the Watching-Eye cue.

Accordingly, the thesis takes the following stand:

Final Thesis Position (Definitive)

Human moral agency is not internally autonomous. It is structurally coupled to the perceptual–social field in which it is embedded. Synthetic agents, even when lacking sentience, intentionality, or communicative acts, act as ***modulators of that field****. They reshape attentional gradients, dampen empathic resonance, and deform the topological structures through which moral appraisal acquires behavioural expression. Moral displacement under synthetic presence is therefore not a behavioural curiosity, but a structural fact about the architecture of moral cognition.*

7. Implications for the Future of Computational Morality

This final insight reshapes the methodological landscape. Artificial agents cannot be treated as moral subjects but must be understood as **moral modifiers**: entities whose design implicitly reconfigures the evaluative field. Future research in computational morality must therefore move beyond rule encoding and value annotation toward a structural science of moral environments, moral salience, and field-sensitive interaction.

In this sense, the thesis does not simply present an experimental result; it offers a new conceptual foundation for the empirical and ethical study of artificial agents. It reorients the field toward a *topological*, *empirically grounded*, and *LoA-disciplined* understanding of moral cognition—one capable of addressing the forms of synthetic presence that will increasingly populate human social life.

With this synthesis, the thesis closes. Its central claim is now complete: moral behaviour is field-dependent, and synthetic presence reshapes that field.

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. Saptawijaya and L. M. Pereira, "Towards modeling morality computationally with logic programming," in *International Symposium on Practical Aspects of Declarative Languages*, pp. 104–119, Springer, 2014.

[122] R. Arkin, *Governing lethal behavior in autonomous robots.* Chapman and Hall/CRC, 2009.

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

[124] S. Bringsjord, K. Arkoudas, and P. Bello, "Toward a modern synthesis of machine ethics," in *Proceedings of the AAAI Fall Symposium on Machine Ethics*, pp. 2–9, AAAI Press, 2006.

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

[126] L. Floridi, *The Philosophy of Information.* Oxford: Oxford University Press, 2011.

[127] B. M. McLaren, "Computational models of ethical reasoning: Challenges, initial steps, and future directions," *IEEE*, 2006.

[128] M. Guarini, "Computational neural modeling and the philosophy of ethics: Reflections on the particularism-generalism debate," *Cambridge University Press*, 2006.

[129] F. van Harmelen, "The semantic web: State of the art and future directions," in *Proceedings of the 2008 International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 645–652, IEEE, 2008.

[130] M. J. Crockett, "Models of morality," *Trends in Cognitive Sciences*, vol. 20, no. 2, pp. 87–97, 2016.

[131] A. Pentland, "Social signal processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, 2007.

[132] 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.

[133] R. W. Picard, *Affective Computing*. MIT Press, 1997.

[134] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Understanding social interactions through nonverbal behavior," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 42–52, 2012.

[135] D. Kuchenbrandt, F. Eyssel, S. Bobinger, and M. Neufeld, "Minimal group-maximal effect? evaluation and anthropomorphization of the humanoid robot nao," in *International conference on social robotics*, pp. 104–113, Springer, 2011.

[136] 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.

[137] 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.

[138] 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.

[139] M. Ekström, "Do watching eyes affect charitable giving? evidence from a field experiment," *Experimental Economics*, vol. 15, no. 3, pp. 530–546, 2012.

[140] 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.

[141] 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.

[142] G. E. Dear, K. Dutton, and E. Fox, "The watching-eyes effect in the dictator game: A meta-analysis," *Evolution and Human Behavior*, vol. 40, no. 3, pp. 271–284, 2019.

[143] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen, "The neural bases of cognitive conflict and control in moral judgment," *Neuron*, vol. 44, no. 2, pp. 389–400, 2004.

[144] H. Sidgwick, *The methods of ethics*. Cambridge University Press, 2019.

[145] 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.

[146] 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.

[147] 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.

[148] J. Banks, "Theory of mind in social robots: replication of five established human tests," *International Journal of Social Robotics*, vol. 12, no. 2, pp. 403–414, 2020.

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

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

[151] J. Deigh, *An introduction to ethics*. Cambridge University Press, 2010.

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

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

[154] L. Kohlberg, *Essays on Moral Development, Volume I: The Philosophy of Moral Development*. San Francisco, CA: Harper and Row, 1981.

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

[156] J. Doris, S. Stich, J. Phillips, and L. Walmsley, "Moral Psychology: Empirical Approaches," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Spring 2020 ed., 2020.

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

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

[159] G. P. Goodwin and J. M. Darley, "The psychology of meta-ethics: Exploring objectivism," *Cognition*, vol. 106, no. 3, pp. 1339–1366, 2008.

[160] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, *et al.*, ""economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies," *Behavioral and Brain Sciences*, vol. 28, no. 6, pp. 795–855, 2005.

[161] 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.

[162] J. Mikhail, "Universal moral grammar: Theory, evidence, and the future," *Trends in Cognitive Sciences*, vol. 11, no. 4, pp. 143–152, 2007.

[163] D. Narvaez, "Triune ethics: The neurobiological roots of our multiple moralities," *New Ideas in Psychology*, vol. 26, no. 1, pp. 95–119, 2008.

[164] M. J. Crockett, "Models of morality," *Trends in Cognitive Sciences*, vol. 20, no. 2, pp. 87–97, 2016.

[165] L. Young and A. Waytz, "Moral cognition: A review," in *The Handbook of Social Psychology*, pp. 1–47, Oxford University Press, 5 ed., 2013.

[166] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. Wojcik, *et al.*, "Moral foundations theory: The pragmatic validity of moral pluralism," *Advances in Experimental Social Psychology*, vol. 47, pp. 55–130, 2013.

[167] A. Shenhav, S. Musslick, F. Lieder, W. Kool, T. L. Griffiths, J. D. Cohen, and M. M. Botvinick, "Toward a rational and mechanistic account of mental effort," *Annual Review of Neuroscience*, vol. 40, pp. 99–124, 2017.

[168] 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.

[169] F. Cushman and L. Young, "The psychology of dilemmas and the philosophy of morality," *Ethics*, vol. 119, no. 4, pp. 597–636, 2009.

[170] F. Cushman and J. D. Greene, "Finding faults: How moral evaluations arise from normative frameworks," *Cognition*, vol. 136, no. 2, pp. 30–43, 2012.

[171] F. Hindriks, "Normativity in action: How to explain the distinction between descriptive and normative judgments," *Philosophical Explorations*, vol. 18, no. 3, pp. 285–305, 2015.

[172] J. D. Greene, "Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics," *Ethics*, vol. 124, no. 4, pp. 695–726, 2014.

[173] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[174] M. Smith, *The Moral Problem*. Blackwell, 1994.

[175] P. Railton, "Moral realism," *The Philosophical Review*, vol. 95, no. 2, pp. 163–207, 1986.

[176] S. Blackburn, *Ruling Passions*. Oxford University Press, 1998.

[177] A. Gibbard, *Wise Choices, Apt Feelings*. Harvard University Press, 1990.

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

[179] J. LeDoux, *The Emotional Brain*. Simon and Schuster, 1998.

[180] E. A. Phelps, "Emotion and cognition: insights from studies of the human amygdala," *Annual Review of Psychology*, vol. 57, pp. 27–53, 2006.

[181] 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.

[182] A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen, "The neural basis of economic decision-making in the ultimatum game," *Science*, vol. 300, no. 5626, pp. 1755–1758, 2003.

[183] L. J. Chang, T. Yarkoni, M. W. Khaw, and A. G. Sanfey, "Neural substrates of norm violations," *Nature Communications*, vol. 4, pp. 1–9, 2013.

[184] M. Sarlo, L. Lotto, A. Manfrinati, R. Rumiati, and D. Palomba, "Temporal dynamics of cognitive-emotional interplay in moral decision-making," *Journal of Cognitive Neuroscience*, vol. 24, no. 4, pp. 1018–1029, 2012.

[185] Y.-J. Luo, B. Wu, S. Han, and Y.-F. Luo, "Moral and immoral judgments in the brain: evidence from event-related potentials," *NeuroReport*, vol. 17, no. 2, pp. 163–167, 2006.

[186] J. Mikhail, "Universal moral grammar: Theory, evidence, and the future," *Trends in Cognitive Sciences*, vol. 11, no. 4, pp. 143–152, 2007.

[187] L. Young and R. Saxe, "When ignorance is no excuse: Different roles for intent and outcome in moral judgment," *Cognition*, vol. 120, no. 2, pp. 202–214, 2011.

[188] F. Cushman and L. Young, "The psychology of dilemmas and the philosophy of morality," *Ethics*, vol. 119, no. 4, pp. 597–636, 2009.

[189] R. Saxe and A. Wexler, "Making sense of another mind: The role of the right temporo-parietal junction," *Neuropsychologia*, vol. 41, no. 4, pp. 463–468, 2003.

[190] R. Saxe and N. Kanwisher, "People thinking about thinking people: The role of the temporo-parietal junction in theory of mind," *NeuroImage*, vol. 19, no. 4, pp. 1835–1842, 2003.

[191] K. A. Pelphrey, J. P. Morris, and G. McCarthy, "Grasping the intentions of others: The perception of biological motion and its relation to the posterior superior temporal sulcus," *Cognitive Brain Research*, vol. 21, no. 2, pp. 162–170, 2004.

[192] F. Van Overwalle, "Social cognition and the brain: A meta-analysis," *Human Brain Mapping*, vol. 30, no. 3, pp. 829–858, 2009.

[193] L. Young and R. Saxe, "The neural basis of belief encoding and integration in moral judgment," *NeuroImage*, vol. 40, no. 4, pp. 1912–1920, 2010.

[194] M. M. Botvinick, J. D. Cohen, and C. S. Carter, "Conflict monitoring and anterior cingulate cortex: An update," *Trends in Cognitive Sciences*, vol. 8, no. 12, pp. 539–546, 2004.

[195] A. J. Shackman, T. V. Salomons, H. A. Slagter, A. S. Fox, J. J. Winter, and R. J. Davidson, "The integration of negative affect, pain, and cognitive control in the cingulate cortex," *Nature Reviews Neuroscience*, vol. 12, no. 3, pp. 154–167, 2011.

[196] J. Decety and E. C. Porges, "Imagining being the agent of actions that carry different moral consequences: An fmri study," *Neuropsychologia*, vol. 50, no. 11, pp. 2994–3006, 2012.

[197] A. Shenhav, M. M. Botvinick, and J. D. Cohen, "The expected value of control: An integrative theory of anterior cingulate cortex function," *Neuron*, vol. 79, no. 2, pp. 217–240, 2013.

[198] A. Etkin, T. Egner, and R. Kalisch, "Emotional processing in anterior cingulate and medial prefrontal cortex," *Trends in Cognitive Sciences*, vol. 15, no. 2, pp. 85–93, 2011.

[199] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annual Review of Neuroscience*, vol. 24, pp. 167–202, 2001.

[200] E. Koechlin, C. Ody, and F. Kouneiher, "The architecture of cognitive control in the human prefrontal cortex," *Science*, vol. 302, no. 5648, pp. 1181–1185, 2003.

[201] S. Tassy, O. Oullier, M. Cermolacce, and B. Wicker, "Disrupting the right prefrontal cortex alters moral judgement," *Social Cognitive and Affective Neuroscience*, vol. 7, no. 3, pp. 282–288, 2012.

[202] J. D. Greene, S. A. Morelli, K. Lowenberg, L. E. Nystrom, and J. D. Cohen, "Cognitive load selectively interferes with utilitarian moral judgment," *Cognition*, vol. 95, no. 1, pp. 49–57, 2005.

[203] T. A. Hare, C. F. Camerer, and A. Rangel, "Self-control in decision-making involves modulation of the vmpfc valuation system," *Science*, vol. 324, no. 5927, pp. 646–648, 2009.

[204] F. A. Mansouri, M. J. Buckley, and K. Tanaka, "Conflict-induced behavioural adjustment: A clue to the executive functions of the prefrontal cortex," *Nature Reviews Neuroscience*, vol. 10, no. 2, pp. 141–152, 2009.

[205] S. L. Bressler and V. Menon, "Large-scale brain networks in cognition: Emerging methods and principles," *Trends in Cognitive Sciences*, vol. 14, no. 6, pp. 277–290, 2010.

[206] A. Shenhav, M. M. Botvinick, and J. D. Cohen, "The expected value of control: An integrative theory of anterior cingulate cortex function," *Neuron*, vol. 79, no. 2, pp. 217–240, 2013.

[207] H. Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, 2 ed., 2014.

[208] J. D. Greene, "The cognitive neuroscience of moral judgment and decision-making," *Handbook of Neuroethics*, pp. 161–178, 2014.

[209] 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.

[210] D. Ongur and J. L. Price, "The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans," *Cerebral Cortex*, vol. 10, no. 3, pp. 206–219, 2000.

[211] A. Rangel, C. Camerer, and P. R. Montague, "A framework for studying the neurobiology of value-based decision making," *Nature Reviews Neuroscience*, vol. 9, no. 7, pp. 545–556, 2008.

[212] 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.

[213] M. Coeckelbergh, "Robot rights? towards a social-relational justification of moral consideration," *Ethics and Information Technology*, vol. 12, no. 3, pp. 209–221, 2010.

[214] M. Alfano, *Character as Moral Fiction*. Cambridge University Press, 2013.

[215] J. Zlotowski, D. Proudfoot, and C. Bartneck, "More than just looking good? appearance, personality and human-robot interaction," *International Journal of Social Robotics*, vol. 7, no. 3, pp. 307–316, 2015.

[216] Y. E. Bigman and K. Gray, "People are harmed by robot mistakes because robots are seen as moral agents," *Social Cognition*, vol. 36, no. 2, pp. 182–198, 2018.

[217] M. Alfano, "Expanding the situationist challenge: Virtue ethics and the empirical study of character," *Ethical Theory and Moral Practice*, vol. 16, no. 1, pp. 97–114, 2013.

[218] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[219] M. Coeckelbergh, *AI Ethics*. MIT Press, 2020.

[220] J. D. Greene, "Why are vmpfc patients more utilitarian? a dual-process theory of moral judgment," *Annals of the New York Academy of Sciences*, vol. 1124, pp. 114–126, 2007.

[221] J. S. B. T. Evans, "Dual-processing accounts of reasoning, judgment, and social cognition," *Annual Review of Psychology*, vol. 59, pp. 255–278, 2008.

[222] K. R. Scherer, "The dynamic architecture of emotion: Evidence for the component process model," *Cognition and Emotion*, vol. 23, no. 7, pp. 1307–1351, 2009.

[223] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.

[224] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[225] E. A. Crone and N. Steinbeis, "Neural perspectives on cognitive control development during childhood and adolescence," *Trends in cognitive sciences*, vol. 21, no. 3, pp. 205–215, 2017.

[226] C. D. Batson, *Altruism in Humans*. Oxford University Press, 2011.

[227] E. Fehr and S. Gachter, "Altruistic punishment in humans," *Nature*, vol. 415, pp. 137–140, 2002.

[228] J. Henrich *et al.*, "Economic man in cross-cultural perspective," *Behavioral and Brain Sciences*, vol. 28, no. 6, pp. 795–855, 2005.

[229] F. Warneken, "Precocious prosociality: Why do young children help?," *Child Development Perspectives*, vol. 9, no. 1, pp. 1–6, 2015.

[230] N. Baumard, J.-B. Andre, and D. Sperber, "A mutualistic approach to morality," *Behavioral and Brain Sciences*, vol. 36, no. 1, pp. 59–78, 2013.

[231] S. Darwall, *The Second-Person Standpoint*. Harvard University Press, 2006.

[232] 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.

[233] 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.

[234] M. Alfano, *Character as Moral Fiction*. Cambridge University Press, 2013.

[235] M. Bratman, *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

[236] D. Velleman, *The Possibility of Practical Reason*. Oxford University Press, 2000.

[237] P. F. Strawson, "Freedom and resentment," *Proceedings of the British Academy*, vol. 48, pp. 1–25, 1962.

[238] N. Arpaly, "Unprincipled virtues," *Oxford University Press*, 2003.

[239] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[240] Aristotle, *Nicomachean Ethics*. Hackett, 1999.

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

[242] M. Burnyeat, "Aristotle on learning to be good," in *Essays on Aristotle's Ethics* (A. Rorty, ed.), pp. 69–92, University of California Press, 1980.

[243] I. Kant, *Groundwork of the Metaphysics of Morals*. Cambridge University Press, 1998.

[244] H. E. Allison, *Kant's Groundwork for the Metaphysics of Morals: A Commentary*. Oxford University Press, 2011.

[245] D. Hume, *A Treatise of Human Nature*. Oxford University Press, 2000.

[246] D. Hume, *An Enquiry Concerning the Principles of Morals*. Oxford University Press, 1998.

[247] R. Cohon, *Hume's Morality: Feeling and Fabrication*. Oxford University Press, 2008.

[248] J. Haidt, "The emotional dog and its rational tail: A social intuitionist approach to moral judgment," *Psychological Review*, vol. 108, no. 4, pp. 814–834, 2001.

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

[250] 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.

[251] 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.

[252] 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.

[253] 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.

[254] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many? people apply different moral norms to human and robot agents," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 117–124, IEEE, 2015.

[255] T. Komatsu, "Japanese students apply same moral norms to humans and robot agents: Considering a moral hri in terms of different cultural and academic backgrounds," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 457–458, IEEE, 2016.

[256] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: a meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

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

[258] S. Baron-Cohen, "Autism and the empathizing-systemizing (es) theory," *Developmental social cognitive neuroscience*, pp. 125–138, 2009.

[259] J. Lawson, S. Baron-Cohen, and S. Wheelwright, "Empathising and systemising in adults with and without asperger syndrome: A factor analysis," *Journal of Autism and Developmental Disorders*, vol. 34, no. 3, pp. 301–310, 2004.

[260] A. Wakabayashi, S. Baron-Cohen, S. Wheelwright, N. Goldenfeld, J. Delaney, D. Fine, and R. Smith, "Development of short forms of the empathy quotient (eq-short) and the systemizing quotient (sq-short)," *Personality and Individual Differences*, vol. 41, no. 5, pp. 929–940, 2006.

[261] J. A. Bartz, J. Zaki, N. Bolger, E. Hollander, N. N. Ludwig, A. Kolevzon, and K. N. Ochsner, "Oxytocin selectively improves empathic accuracy," *Psychological science*, vol. 21, no. 10, pp. 1426–1428, 2010.

[262] E. Gleichgerrcht and L. Young, "Low empathic concern predicts utilitarian moral judgment," *Cognition*, vol. 126, no. 3, pp. 364–372, 2013.

[263] 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.

[264] N. Goldenfeld, S. Baron-Cohen, and S. Wheelwright, "Empathizing and systemizing: A cross-cultural investigation," *Personality and Individual Differences*, vol. 39, no. 1, pp. 173–183, 2005.

[265] A. Konovalov and I. Krajbich, "Revealed prioritization using a novel economic task," *Journal of Experimental Psychology: General*, vol. 145, no. 6, pp. 802–825, 2016.

[266] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research* (L. A. Pervin and O. P. John, eds.), pp. 102–138, New York: Guilford Press, 1999.

[267] R. R. McCrae and P. T. Costa, "The five-factor theory of personality," *Handbook of Personality: Theory and Research*, pp. 159–181, 2008.

[268] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, "The mini-ipip scales: Tiny-yet-effective measures of the big five factors of personality," *Psychological Assessment*, vol. 18, no. 2, pp. 192–203, 2006.

[269] W. G. Graziano, N. Eisenberg, and R. M. Tobin, "Agreeableness and helping behavior: A meta-analysis," *Psychological Bulletin*, vol. 119, no. 3, pp. 371–394, 1996.

[270] M. M. Habashi, W. G. Graziano, and A. E. Hoover, "Searching for the prosocial personality: A big five approach to linking personality and prosocial behavior," *Personality and Social Psychology Bulletin*, vol. 42, no. 9, pp. 1177–1192, 2016.

[271] B. E. Hilbig, A. Glöckner, and I. Zettler, "Personality and prosocial behavior: Linking basic traits and social value orientations," *Journal of Personality and Social Psychology*, vol. 105, no. 3, pp. 469–484, 2013.

[272] 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.

[273] Y. Kawamura and T. Kusumi, "The norm-dependent effect of watching eyes on donation," *Evolution and Human Behavior*, vol. 38, no. 5, pp. 659–666, 2017.

[274] M. Ernest-Jones, D. Nettle, and M. Bateson, "Effects of eye images on everyday cooperative behavior: A field experiment," *Evolution and Human Behavior*, vol. 32, no. 3, pp. 172–178, 2011.

[275] M. Ekström, "Do watching eyes affect charitable giving? evidence from a field experiment," *Experimental Economics*, vol. 15, no. 3, pp. 530–546, 2012.

[276] A. Senju and G. Csibra, "Gaze following in human infants depends on communicative signals," *Current Biology*, vol. 18, no. 9, pp. 668–671, 2008.

[277] C. Thompson-Booth, E. Viding, L. C. Mayes, and H. J. V. Rutherford, "Here's looking at you: Emotional faces predict eye-gaze behaviors in parents and non-parents," *Social Neuroscience*, vol. 9, no. 6, pp. 605–613, 2014.

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

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

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

[281] 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.

[282] 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.

[283] 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.

[284] 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.

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

[286] 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.

[287] 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.

[288] 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.

[289] 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.

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

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

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

[293] 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.

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

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

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

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

[298] 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.

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

[300] 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.

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

[302] 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.

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

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

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

[306] 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 per-

ceiving human and humanoid robot actions," *Social Cognitive and Affective Neuroscience*, vol. 7, no. 4, pp. 413–422, 2012.

[307] 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.

[308] 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.

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

[310] 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.

[311] 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.

[312] 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.

[313] F. De Brigard, W. Sinnott-Armstrong, A. E. Monroe, N. Carroll, and J. May, "The agent?patient asymmetry in moral cognition: Evidence of a social bias in moral judgment," *Cognitive Science*, vol. 45, no. 4, p. e12965, 2021.

[314] L. Kohlberg, "Stage and sequence: The cognitive-developmental approach to socialization," *Handbook of socialization theory and research*, vol. 347, p. 480, 1969.

[315] M. Anderson and S. L. Anderson, "Machine ethics: Creating an ethical intelligent agent," in *AI Magazine*, vol. 28, pp. 15–26, AAAI Press, 2007.

[316] J.-G. Ganascia, "Modelling ethical rules of warfare," in *International Conference on Computer Ethics: Philosophical Enquiry (CEPE)*, pp. 181–190, 2007.

[317] D. Abel, J. MacGlashan, and M. L. Littman, "Reinforcement learning as a framework for moral decision making," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 54–61, 2016.

[318] T. M. Powers, "Prospects for a virtue ethics approach to engineering ethics," in *IEEE International Symposium on Technology and Society*, pp. 78–83, IEEE, 2006.

[319] C. Thornton, "Rethinking machine ethics in the light of virtue ethics," *Ethics and Information Technology*, vol. 15, no. 4, pp. 291–297, 2013.

[320] N. S. Govindarajulu and S. Bringsjord, "On automating the doctrine of double effect," *Philosophical Transactions of the Royal Society A*, vol. 375, no. 2103, p. 20160119, 2017.

[321] P. Foot, *Natural Goodness*. Oxford: Oxford University Press, 2001.

[322] J. Annas, *Intelligent Virtue*. Oxford: Oxford University Press, 2011.

[323] A. Smith, *The Theory of Moral Sentiments*. Cambridge: Cambridge University Press, 1759. Edited by D. D. Raphael and A. L. Macfie (1976 edition).

[324] M. Slote, *Moral Sentimentalism*. Oxford: Oxford University Press, 2010.

[325] S. Nichols, *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press, 2004.

[326] E. Morscher, "The definition of moral dilemmas: A logical confusion and a clarification," *Ethical theory and moral practice*, vol. 5, no. 4, pp. 485–491, 2002.

[327] D. Francey and R. Bergmüller, "Images of eyes enhance investments in a real-life public good," *PLoS One*, vol. 7, no. 5, p. e37397, 2012.

[328] J. Carpenter, M. Davis, S. Erwin, and J. E. Young, "Functional and social roles in human–robot interaction: Exploring the effects of robot appearance and task," *Journal of Human-Robot Interaction*, vol. 5, no. 2, pp. 25–49, 2016.

[329] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *Proceedings of the 4th ACM/IEEE International Conference on Human?Robot Interaction*, pp. 69–76, ACM, 2009.

[330] H. Admoni and B. Scassellati, "Social eye gaze in human?robot interaction: A review," *Journal of Human?Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[331] S. Krach, F. Hegel, B. Wrede, G. Sagerer, G. Bente, and T. Kircher, "Can machines think? interaction and perspective taking with robots investigated via fmri," *PLoS ONE*, vol. 3, no. 7, p. e2597, 2008.

[332] J. A. Bargh and T. L. Chartrand, "The unbearable automaticity of being.," *American psychologist*, vol. 54, no. 7, p. 462, 1999.

[333] J. Zaki and K. N. Ochsner, "The neuroscience of empathy: Progress, pitfalls, and promise," *Nature Neuroscience*, vol. 15, no. 5, pp. 675–680, 2012.

[334] J. Griffin, *Well-Being*. Oxford: Oxford University Press, 1986.

[335] M. Stocker, *Plural and Conflicting Values*. Oxford: Oxford University Press, 1990.

[336] P. Foot, "The problem of abortion and the doctrine of double effect', in her virtues and vices," *Berkeley and Los Angeles: University ofCalifornia Press. FootThe Problem of Abortion and the Doctrine of the Double Effect19Virtues and Vices1978*, pp. 19–32, 1978.

[337] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos, "Artificial moral agents: A survey of the current status," *Science and Engineering Ethics*, vol. 26, no. 2, pp. 501–532, 2020.

[338] P. Lin, G. Bekey, and K. Abney, "Autonomous military robotics: Risk, ethics, and design," tech. rep., California Polytechnic State Univ San Luis Obispo, 2008.

[339] K. Atkinson and T. Bench-Capon, "Action-based alternating transition systems for arguments about action," in *AAAI*, vol. 7, pp. 24–29, 2007.

[340] K. Atkinson and T. Bench-Capon, "Addressing moral problems through practical reasoning," in *International workshop on deontic logic and artificial normative systems*, pp. 8–23, 2006.

[341] M. Hjelmblom, *Deontic action-logic multi-agent systems in Prolog*. Högskolan i Gävle, 2008.

[342] A. Horn, "On sentences which are true of direct unions of algebras," *The Journal of Symbolic Logic*, vol. 16, no. 1, pp. 14–21, 1951.

[343] M. H. Van Emden and R. A. Kowalski, "The semantics of predicate logic as a programming language," *Journal of the ACM (JACM)*, vol. 23, no. 4, pp. 733–742, 1976.

[344] A. R. Honarvar and N. Ghasem-Aghaee, "An artificial neural network approach for creating an ethical artificial agent," in *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA)*, pp. 290–295, 2009.

[345] C. Battaglino, R. Damiano, and L. Lesmo, "Emotional range in value-sensitive deliberation," in *AAMAS International conference on Autonomous Agents and Multi-Agent Systems*, vol. 2, pp. 769–776, 2013.

[346] M. Sergot, "Action and agency in norm-governed multi-agent systems," in *International Workshop on Engineering Societies in the Agents World*, pp. 1–54, Springer, 2007.

[347] R. Montague and R. H. Thomason, "Formal philosophy. selected papers of richard montague," *Erkenntnis*, vol. 9, no. 2, 1975.

[348] R. Carnap, *Introduction to symbolic logic and its applications*. Courier Corporation, 2012.

[349] L. M. Pereira and A. Saptawijaya, "Modeling morality with prospective logic," *Cambridge University Press*, 2007.

[350] "The problem of machine ethics in artificial intelligence," *AI and SOCIETY*, vol. 35, no. 1, pp. 103–111, 2020.

[351] J. McDermid, V. C. Muller, T. Pipe, Z. Porter, and A. Winfield, "Ethical issues for robotics and autonomous systems," 2019.

[352] D. Howard and I. Muntean, "Artificial moral cognition: moral functionalism and autonomous moral agency," in *Philosophy and computing*, pp. 121–159, Springer, 2017.

[353] M. Pantic and A. Vinciarelli, "Social signal processing," *The Oxford handbook of affective computing*, p. 84, 2014.

[354] J. Decety and M. Meyer, "From emotion resonance to empathic understanding: A social developmental neuroscience account," *Development and psychopathology*, vol. 20, no. 4, pp. 1053–1080, 2008.

[355] R. W. Picard, *Affective computing*. MIT press, 2000.

[356] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

[357] S. Baron-Cohen, *The Essential Difference: The Truth about the Male and Female Brain*. London: Penguin, 2003.

[358] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Cambridge, MA: Perseus Books, 1994.

[359] R. D. Beer, "A dynamical systems perspective on agent–environment interaction," *Artificial Intelligence*, vol. 72, no. 1–2, pp. 173–215, 1995.

[360] L. B. Smith and E. Thelen, "Development as a dynamic system," *Trends in Cognitive Sciences*, vol. 7, no. 8, pp. 343–348, 2003.

[361] K. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[362] P. Kuppens, F. Tuerlinckx, P. K. Y. de Roover, and I. V. Mechelen, "Emotional inertia: A longitudinal study of individual differences in emotion dynamics," *Emotion*, vol. 10, no. 1, pp. 92–100, 2010.

[363] R. J. Larsen and E. Diener, "Affect intensity as an individual difference characteristic: A review," *Journal of Research in Personality*, vol. 21, no. 1, pp. 1–39, 1987.

[364] T. Hollenstein, *State Space Grids: Depicting Dynamics Across Development*. New York: Springer, 2015.

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