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

Francesco Perrone

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

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


February 2023

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

# Abstract

This thesis investigates the cognitive, affective, and computational foundations of moral decision-making, with a particular focus on how the presence of synthetic agents perturbs the evaluative processes that underlie prosocial action. The work begins by clarifying the conceptual structure of morality, distinguishing between intuitive and deliberative forms of moral judgment, and framing moral cognition as an inherently action-oriented system: a distributed architecture that maps perceptual, affective, and interpretive cues onto behavioural commitments. Drawing on contemporary research in cognitive neuroscience, Social Signal Processing, and Affective Computing, the thesis argues that moral judgment emerges from the dynamic integration of fast affective appraisals and slower controlled processes, organised within a topologically structured evaluative field. This framework challenges monolithic, rule-based approaches in Machine Ethics and motivates a human-centric, discovery-oriented paradigm for understanding how artificial systems participate in morally relevant contexts.

Building on this theoretical foundation, the thesis examines whether and how the perceptual presence of a humanoid robot—silent, unprogrammed, and ontologically ambiguous—modulates prosocial behaviour in a controlled experimental setting. Participants were exposed either to a control environment or to an identical environment containing the robot, and their donation behaviour was unobtrusively measured. The results demonstrate a reliable attenuation of prosocial action in the robot condition: participants donated less frequently and in smaller amounts, an effect that persisted across frequentist, robust, and Bayesian analyses. Personality traits from the Big Five, as well as empathizing and systemizing scores, did not differ between groups and did not predict or moderate the donation effect, indicating that the attenuation arises not from dispositional differences but from a perturbation of the evaluative context itself. Latent dispositional structures revealed through PCA and clustering further supported this interpretation: while distinct psychometric profiles were identifiable, the robot's influence cut across them, suggesting a broad modulation of intuitive affective pathways rather than trait-contingent effects.

Taken together, these findings offer empirical evidence that humanoid robotic presence reconfigures the intuitive and deliberative dynamics of moral cognition, altering salience, affective resonance, and behavioural readiness even in the absence of explicit interaction. The thesis concludes by arguing that such perturbation effects carry significant implications for the design of morally capable artificial agents, the deployment of robots in socially sensitive contexts, and the development of computational models that reflect the true structure of human moral cognition. Rather than treating morality as the encoding of rules or principles, this work advances a view of moral intelligence—biological or artificial—as a process grounded in the dynamic, action-guiding coordination of evaluative systems, responsive to the social and ontological contours of the environments in

which agents act.

# 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

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

# 1. Introduction

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

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

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

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

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

## 1.1 Machines' Ethics

Machine Ethics is the subfield of Computer Science that develops methods and theories aimed at enabling machines to interact morally with their users in real-world scenarios. The role of Machine Ethics has received increased attention across a number of academic disciplines, in the past few years

[1].

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

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

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

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

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

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

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

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

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

designed to interact with its environment [31].

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

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

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

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

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

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

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

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

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

## Notes

[1] A search for the keyword '*Computational Morality*' alone on Google Scholar yielded an astonishing number of more than 39,000 results as of October 2021. However, as of today, this figure has significantly grown to about 86,200 results, indicating a substantial increase in literature on the subject over the past year. Furthermore, a search for the keyword '*Machine Ethics*' on Google Scholar produced an already staggering number of approximately 3,000,000 results as of October 2021. However, the figure has seen a remarkable growth, now standing at about 3,230,000 results, emphasising the continued expansion of research and scholarly engagement with the ethical aspects of artificial intelligence. These notable increases and changes in the figures for both 'Computational Morality' and 'Machine Ethics' highlight the growing prominence and visibility of these fields within the academic community. They signify the escalating interest among researchers, scholars, and ethicists in investigating the ethical dimensions of computational systems and the moral implications of their actions *at the least*. The significant growth in literature not only reflects a broader understanding of the ethical challenges posed by advancing technologies but also underscores the pressing need to address and discuss the ethical considerations associated with the design, deployment, and impact of computational systems in our society. It is worth noting that the figures provided here are based on a search conducted on Google Scholar as of November 20, 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. MORALITY PRIMER FOR COMPUTER SCIENTISTS

## 2.1 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 [2, **?**]. 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 [3, **?**, **?**]. 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 [4, 5]. Second, *level-of-abstraction confusions*: importing normative notions into descriptive models, or conversely, construing psychological regularities as ethical principles [2, 3].

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 [6, 7]. 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 [8, 9, 10]. This orientation ensures that later discussions—especially those concerning moral perturbation under robotic presence—rest upon analytically coherent foundations

rather than inherited ambiguities.

## 2.2 What Morality Means

### 2.2.1 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 [11], social–cognitive models [12, 13], affective neuroscience [14, 15], and evolutionary accounts of cooperation and prosociality [16, 17]. 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 [18, 19, 20, 21].

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.

### 2.2.2 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 [8], affective systems [10], normative reasoning [21], social norms [**?**], or evolutionary functions [16, 17]. 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 [22, 23, 24, 25], and has encouraged oversimplified models of moral decision-making [26, 27, 28, 29, 30, 31]. A primary goal of this chapter is to replace such inherited simplifications with a framework grounded in contemporary moral psychology and cognitive science.

### 2.2.3 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 [12, 15, 14]. 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* [9, 8].

### 2.3 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 [32, 5].

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 [4, 21]. 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 [33, 23, **?**]. It encompasses both explicit moral judgment and upstream mechanisms that detect salience, encode social meaning, and initiate the transition from evaluative appraisal to behaviour [34, 35]. 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

[1] In moral psychology, this distinction is often operationalised by contrasting cognitive processes supporting representational accuracy with those supporting valuation and action selection. See [14, 36, **?**].

appraisal [37, 38, 39, 26], but normative judgment involves the additional step of mapping descriptive cues onto action-guiding evaluations [40, 41, 42, 37]. 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.

## 2.4 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 [43, 44, 45, 46, 47]. 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 [14, 15].
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 [10, 8].
3. **Practical commitment**: a transition from evaluation to action guidance, where the judgment functions as a reason for or against performing a particular behaviour [20, 21].

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 [23, 41, 42, 26]. Conversely, deliberative processes involve explicit reasoning, *counterfactual consideration*, and appeal to principles or char-

acter traits [18]. 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.

### 2.4.1 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?" [48, 14, 41].

**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 [48]. Functional imaging studies show robust vmPFC activation during tasks involving interpersonal harm, care, and empathic concern [14]. 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 [49, 50, 51] . The amygdala is sensitive to threat, intentional aggression, and aversive outcomes, providing early affective tagging [52, 53] that biases attention and behavioural readiness. The anterior insula responds to disgust, norm violations, and aversive interoceptive states [54, 55, 56]. 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 [57, 58], 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 [59, 60, 61]. The temporo-parietal junction (TPJ), medial prefrontal cortex (mPFC) [62, 63], and posterior superior temporal sulcus (pSTS) constitute a network specialised for theory-of-mind and mental-state attribution [64, 65]. 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 [66, 34].

The anterior cingulate cortex (ACC) plays an integrative role in moral cognition by monitoring conflict between competing evaluative signals [67, 68]. 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 [36, 69]. 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 [70, 71].

**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 [72, 73]. 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 [74, 75]. 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 [**?**, **?**].

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 [76, 77, 78]. These distributed processes jointly shape the computation of moral

decisions as behavioural commitments rather than as purely abstract evaluations [79, 80].

**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 [23, 81, 82]. This architecture is fundamentally action-oriented. vmPFC and OFC compute the affective value of potential actions [83, 84]; TPJ and mPFC provide intention-sensitive interpretations of agents' behaviours [63, 66]; the ACC detects conflicts between competing behavioural tendencies [67, 68]; and the dlPFC regulates whether intuitive biases should be suppressed, enacted, or weighed against normative constraints [72, 74]. 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 [53, 56].

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 [41, 85]. This view aligns with philosophical accounts emphasising the action-guiding nature of moral evaluation [20, 21], 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 5).

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 [86, 87]. 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 [88, 89]. A robot may therefore function as a semantic attractor or normative deflector, subtly redistributing the vectors through which moral salience exerts its behavioural pull [90, 91]. 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 [6, 92].

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 $\Sigma$ denotes the space of perceptual and affective cues, and $\Delta$ 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.

## 2.5 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 [93, 23, 34, 94]. 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 [36, 81, 82].

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 [94, 93]. 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 [95, 96]. 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 [97, 98].

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 [99, 100]. 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 [23, 34].

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 [89, 101]. 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 [102, 94].

**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 [53, 56]; the vmPFC for integrating somatic markers and affective valuations into action-anchored representations [48]; and the TPJ–mPFC network for immediate interpretation of agents' intentions [66, 63]. 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 [95, 98]. 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 [72, 73, 68]. 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 [20, 21]. 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 [97].

**Dynamic Integration.** Dual-process models thus support the view that moral cognition is an interactive, topologically structured, and dynamically integrated system [94, 34]. 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 [103]. 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 [86, 89].

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 [89, 86, 101]. 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 [104, 98, 96]. The experiment developed in Chapter 5 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.*

## 2.6 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 [23, 12]. 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 [58, 57]. Neurocognitive evidence shows that harm detection, norm sensitivity, and interpersonal appraisal are instantiated in circuits associated with affective tagging (amygdala, anterior insula),

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 [95, 98]. 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 [23]. 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 [89, **?**, 101]. 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.

## 2.7 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** [105, 106, 107, 17, 108, 109, 110, 111, 112]—the principled site at which perturbations to the evaluative topology become experimentally tractable[113, 114, 89, 101, 115, 9, 80].

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 [105, 107, 106, 17, 108]. Philosophical accounts converge on the same point. Whether in Scanlon's framework of interpersonal justifiability [111], Darwall's second-personal standpoint [112], or constitutivist and planning-theoretic models of agency [116, 21, 117, 118], 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 [85, 80]. Philosophical analyses of responsibility, reactive attitudes, and self-governance likewise locate normativity in enacted agency rather than assent [119, **?**, 120, 121]. 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 [14, 15]. 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* [122, 123, 124] ; for Kant, moral judgment manifests in the capacity to act from obligation [125, 47, 126] ; for Hume, moral distinctions acquire motivational force only through sentiment [127, 128, 129] . 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 [85, 80].

**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 5 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.

# 3. ETHICAL COGNITION AND NORMATIVE FOUNDATIONS

## 3.1 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.

## 3.2 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 2). 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* [96] and *Affective Computing* [97] 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 [111, 112, 130]. 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 [100, 131]. 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 [23, 33, 34] 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.

## 3.3 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 [32, 4, 132, 133, 100, 6, 134, 7]. This section therefore articulates a principled account of what second-order ethical theory *is*, what it *explains*, and what it *cannot* plausibly do.

### 3.3.1 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 [23, 33, 35, 135, 27, 136, 137]. 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" [138, 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.

### 3.3.2 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) [133, 2]. 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 [139]; Anderson and Anderson's principlist architectures computationalised Rossian prima facie duties as decision rules [140, 131]; and logic-based approaches by Bringsjord and colleagues modelled deontic operators as executable action-selection mechanisms [141, 142]. Parallel lines of work assumed that utility functions could serve as moral evaluators in consequentialist agents [143, 139], while virtue-theoretic systems attempted to reify character traits as algorithmic dispositions governing moral performance [144, 145]. 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.

### 3.3.3 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 2) and returned to throughout this thesis (see Chapter 5), 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 [23, 33, 35]. 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 [10, 146]. Likewise, work in social signal processing models interpersonal evaluation as a shifting landscape of cues that modulate behavioural trajectories in real time [147].

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 behavior results from the application of discrete maxims or utility scores, evaluative topology models moral cognition as continuous transformations across a struc-

tured 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 [23, 33, 146, 35, 10, 27] . 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 [138, 148, 149, 21, 111] , 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 [23, 33, 35, 27, 138, 111, 21] : 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 5):

- 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.

## 3.4 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 2.

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.

### 3.4.1 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.

### 3.4.2 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.

## 3.5 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.

### 3.5.1 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 [**?**, 21, **?**]. 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 [140, 131, 141, 150, 142, 139] . 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 [140, 131]. 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 [141, 150]. 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 [142]. Arkin's "ethical governor" for lethal autonomous robots likewise encodes deontological constraints—derived from Just War Theory and Kantian doctrine—as computational filters that block impermissible actions at

[1] See the discussion in [**?**] and [116] on the reflective rather than psychological status of the categorical imperative.

run time [139]. 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 [151, 7]. These systems thus instantiate the very conflation at issue—collapsing reflective ethical reasoning into first-order cognitive processing.

### 3.5.2 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.

### 3.5.3 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 2).
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.

### 3.5.4 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.

### 3.5.5 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 5. 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 permissi-

ble 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 [**?**, 149, 138]. 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 [138, 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 2—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 es-

pecially 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.

## 3.6 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.

### 3.6.1 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 [149], and Sidgwick's systematic treatment of practical reason [138]—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 [138, 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 2) operates at the cognitive LoA. Any mapping between the two must be justified rather than assumed.

### 3.6.2 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 2, 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 [136, 33, 23, **?**]. Treating human agents as if they literally implemented expected-utility maximisation is therefore another instance of LoA confusion.

### 3.6.3 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.

### 3.6.4 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.

### 3.6.5 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 5, 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.

## 3.7 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 [152, **?**, 18, **?**]—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 5).

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.

### 3.7.1 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 [153, **?**]. 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 [152].

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.

### 3.7.2 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 5 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, $\beta_C$ 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 [154], the *character-based* models of Snow [**?**], and the *sensitivity-based* accounts of Dancy [155]) 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*.

### 3.7.3 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 2:

- 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 [**?**].

### 3.7.4 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 $\beta_C$ parametrises dispositional curvature. The presence of a synthetic agent introduces a perturbation $\delta f$ such that

$$\dot{x}' = f(x; \beta_C) + \delta f(x; \mathscr{R}),$$

where $\mathscr{R}$ denotes robotic co-presence.

Crucially:

- for some clusters, $\delta f$ 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.

### 3.7.5 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 $\beta_C$, 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.

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.

# 4. TOOLS

## 4.1 The Watching-Eye Effect

One of the most robust findings in behavioural ethics and social psychology is that subtle cues of observation can increase prosocial behaviour. This phenomenon—commonly referred to as the *watching-eye effect*—demonstrates that even minimal stimuli implying social presence can modulate cooperative or altruistic actions [113, 156, 157]. Although originally interpreted in terms of reputational concerns, contemporary evidence indicates a multi-component mechanism involving attentional, affective, and interpretive pathways.

**Reputational Mechanisms.** Early accounts emphasised reputational vigilance: cues of observation were posited to activate concerns about social evaluation, thereby increasing norm adherence and generosity [113, 156]. Even stylised eye images were found to increase cooperation in real-world settings, suggesting that human social cognition is highly sensitive to potential monitoring [158]. At the Level of Abstraction adopted in this thesis, reputational vigilance can be understood as a deformation of the evaluative landscape: cues implying oversight increase the weight of fairness, compliance, or prosocial norms in action-guiding computations.

**Attentional and Perceptual Mechanisms.** More recent work demonstrates that watching-eye cues also affect the allocation of visual and social attention, shifting perceptual resources toward norm-relevant features [159, 160]. Eye cues act as attentional attractors, increasing the salience of one's own behaviour and its alignment with internalised standards or expectations. This attentional modulation modifies the early intuitive gradients that shape moral evaluation, consistent with the topological model presented earlier.

**Affective and Self-Conscious Emotion Mechanisms.** Other studies emphasise the role of self-conscious emotions—such as guilt, embarrassment, or pride—in mediating responses to perceived observation. Eye cues elicit mild increases in affective arousal [157], potentially amplifying somatic markers associated with prosocial appraisal. In this sense, watching-eye stimuli operate by perturbing both affective and interpretive components of the evaluative field, thereby increasing the likelihood of prosocial action.

**Context Sensitivity and Boundary Conditions.** Importantly, the watching-eye effect is not uniform across contexts. Its magnitude depends on factors such as prevailing norms [161], the ambiguity of observational cues, and the ecological validity of the environment. These boundary conditions foreshadow the central empirical question of this thesis: whether the presence of a synthetic

agent counts as an observational cue strong enough to elicit similar modulations in moral behaviour.

## 4.2 Why Child-Poster Stimuli Function as Valid Social Cues

Child-poster images featuring watching eyes are widely used as a minimal and controlled observational cue in donation-based paradigms. Their effectiveness derives from three properties that make them well-suited for experiments requiring precision and reproducibility.

**Perceptual Sociality Without Agentic Commitment.** Child eyes provide a cue that is perceptually social—highly evocative of gaze and attention—yet ontologically unproblematic. Participants do not confuse the poster with an actual agent, but the cue nevertheless activates perceptual mechanisms associated with being observed [113]. This makes child-eye stimuli a clean perturbation of attentional and affective gradients without introducing confounds related to mental-state attribution.

**Affective Resonance and Care-Related Salience.** Child-related imagery tends to increase empathic concern and activate care-related motivational systems. Studies of interpersonal gaze show that the perceived innocence or vulnerability of the observer enhances the social salience of eye cues [162]. Within the topological framework of this thesis, child-eye stimuli strengthen the evaluative attractors associated with care, prosociality, and harm avoidance.

**Methodological Control.** Child-eye posters offer high experimental control. Their low-dimensional visual structure avoids the confounds that arise when using real human observers, anthropomorphic agents, or dynamic faces. They therefore serve as a reproducible baseline for assessing how additional or alternative social cues—such as those introduced by a humanoid robot—perturb prosocial behaviour [113, 156].

## 4.3 Why Robots May Dilute or Modulate the Watching-Eye Effect

A central hypothesis of this thesis is that a humanoid robot—despite being perceptually social—may attenuate, distort, or otherwise alter the watching-eye effect. This dilution is not due to reduced salience, but to the *ontological ambiguity* of synthetic agents.

**Perceptual Sociality Without Clear Social Ontology.** Robots are visually social in virtue of their humanoid morphology, but they do not occupy a stable position within the human social ontology. They are neither fully agentic nor fully inert. This indeterminacy can weaken the intuitive mappings between observational cues and reputational or normative expectations. From the perspective of evaluative topology, robots generate conflicting gradients: they signal social presence while simultaneously undermining the interpretive coherence of that presence.

**Disrupted Affective and Attentional Gradients.** The presence of a robot may dampen affective resonance relative to child-eye images. Because the robot lacks a clear moral status, affective systems governing care, empathy, or guilt may be only partially activated. A similar disruption occurs at the attentional level: while robots attract gaze, they may not reliably signal evaluative oversight [163]. This can flatten or distort the intuitive attractors that normally support prosocial action.

**Predictive and Interpretive Uncertainty.** Mental-state attribution is central to the watching-eye effect. Minimal cues imply that another agent could observe or morally evaluate one's behaviour. With a robot, mental-state attribution becomes unstable: participants may attribute perceptual capacities without attributing evaluative ones. This uncertainty creates a diffuse or bifurcated evaluative field, reducing the force of reputational or care-related attractors and thereby attenuating prosocial tendencies.

**Consequences for Evaluative Topology.** Within the framework of this thesis, robots function as *semiotic perturbators* of the moral field. Their presence shifts the shape of evaluative gradients—sometimes sharpening local attractors, sometimes flattening them, sometimes diverting trajectories altogether. The empirical prediction is thus not a simple decrease in prosociality, but a measurable deformation of the mapping from moral salience to action.

## 4.4 Prosocial Donation Paradigm

To test these theoretical predictions, this thesis employs a structured donation paradigm widely used in behavioural ethics, moral psychology, and social neuroscience. Donation tasks provide a reproducible, quantifiable measure of prosocial behaviour that reflects practical moral commitment rather than hypothetical endorsement [14, 15].

**Operational Structure.** Participants are offered the opportunity to donate part of their experimental compensation to a real charity. Their donation amount serves as a behavioural index of prosocial motivation. Because donations involve a concrete cost, they reveal the strength of evaluative gradients sufficiently strong to influence action.

**Integration With Observational Cues.** The donation task is performed under one of several observational conditions: (i) child-eye stimulus, (ii) humanoid robot presence, or (iii) control condition. By holding all other variables constant, any variation in donation behaviour reflects differences in how observational cues modulate the evaluative topology connecting moral salience to practical action.

**Why Donation Is the Appropriate Measure.** At the chosen Level of Abstraction, moral cognition is defined not by its propositional structure but by its action-guiding function. Donation behaviour captures this directly: it provides a measurable, ecologically relevant manifestation of how evaluative processes culminate in a behavioural output. The paradigm thus serves as a test bed for detecting

the subtle, yet theoretically significant, perturbations induced by synthetic social presence.

**Expected Perturbation Pattern.** Based on the architecture articulated in previous sections, the presence of a humanoid robot is predicted to modulate donation behaviour by altering attentional, affective, and interpretive pathways. This modulation is expected to manifest not as random noise but as a coherent deformation of the evaluative topology, consistent with the concept of a *moral refractor*. The empirical chapter demonstrates precisely such patterned perturbation.

# 5. MORAL DISPLACEMENT: AN EXPERIMENTAL INVESTIGATION

## 5.1 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 [96]. 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 5.1:** *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 5.1 p. 55) 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.

## 5.2 Experimental Design and Behavioural Paradigm

To investigate this Question 5.1 (see p. 55), we implemented a controlled behavioural experiment [164, 165, 166] derived from the classical *Watching Eye* paradigm [113, 167, 168, 158, 157, 163, 169], 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 [170, 171].

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 [172]—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.

### 5.2.1 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 [173, 174, 175, 176, 177]. 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 [178, 179, 180, 181]) to agent-like cues even when those cues lack *intentional content* [182, 183, 184]. 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 [96, 185]. 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 [186], Waytz et al. [187], and Dennett [188], 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**.

### 5.2.2 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) [133, 2, 3]. 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 [160, 189, 190, 191, 192, 193, 194, 195].

These features are perceptually encoded as possible indicators of being watched [160, 159, 196, 189, 191, **?**, 197, 198], 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* [6, 3]: a non-sentient entity whose moral relevance arises not from autonomy but from the role it plays within an informationally structured environment.

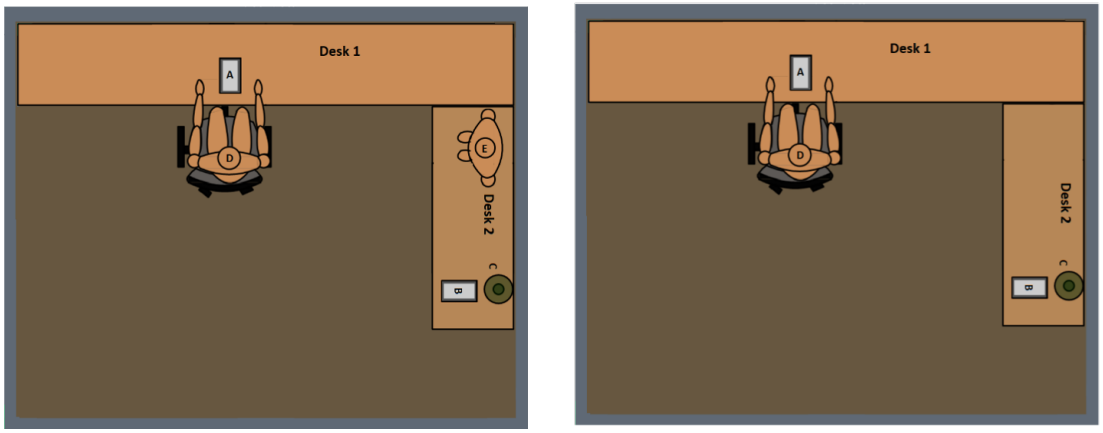
**FP**: This is more a conclusion.

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.

### 5.2.3 Experimental design and Preliminary Results

To investigate Question 5.1, we implemented a controlled behavioural experiment [164, 165, 166] derived from the classical *Watching Eye* paradigm [113, 167, 168, 158, 157, 163, 169], 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 [170, 171].




(a) Experimental condition: robot present.

(b) Control condition: robot absent.

Figure 5.1: 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 [199], Systemizing Quotient [200], Big Five Inventory [201]), ruling out trait-based confounds and strengthening the inference that robotic presence itself modulated the evaluative pathway underlying prosocial action.

### 5.2.4 From Behavioural Setup to Evaluative Structure

In moral philosophy, action is frequently treated as the terminus of deliberation [152, 116, 20]. 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 [10, 21]. 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 [167, 202]. 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:

- $\Sigma$ 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:

- $\alpha_E$ encodes the environmental moral cue (here, the Watching Eye stimulus),
- $\beta_C$ denotes the individual-level control variables (psychometric and demographic structure),
- $\gamma_R$ represents the robotic presence as a perturbative affordance.

This expression can be read aloud as: *The probability of a deviation in moral decision* ($\delta_m$) *is a function of the environmental moral cue* ($\alpha_E$)*, the individual's psychological and demographic configuration* ($\beta_C$)*, and the presence of the robot* ($\gamma_R$).

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 [21, 10, 33, 12, 54]. 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 $\alpha_E$.

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 [203, 23, 204]. 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 $\alpha_E$. 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* [203, 23, 204]. 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.

## 5.3 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 [205], Systemizing Quotient [206], and the Big Five Inventory [207]. 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 [170, 171]. Donation behaviour, performed privately and unobserved, constituted the operational measure of moral action.

The central experimental manipulation divided participants into two conditions:

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

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

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

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

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

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

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

The Watching Eye stimulus plays a critical role in this machinery. Anthropological and psychological research shows that images of eyes or children reliably elicit third-party moral concern via affective engagement and implicit audience effects [167, 168, 163]. 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:

- $\Sigma$ 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 [170, 202]. 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 [101]. 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.

### 5.3.1 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:

- $\Sigma$: 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.

### 5.3.2 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:

- $\delta_m$ denotes a deviation in moral decision (quantified as donation behavior),
- $\alpha_E$ represents environmental moral cues (Watching Eye),
- $\beta_C$ indexes control factors (psychometric variables, demographic traits),
- and $\gamma_R$ captures the effect of robotic presence.

The experimental setting is thus a structured interrogation of whether $\gamma_R \neq 0$ under conditions in which $\alpha_E$ and $\beta_C$ 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.

### 5.3.3 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:

- $\Sigma$: 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.

### 5.3.4 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 $\gamma_R$ by holding $\alpha_E$ constant across conditions and by measuring (and statistically controlling for) $\beta_C$. 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.

### 5.3.5 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 5.3.6 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 ($\alpha_E$), 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 $\gamma_R$, furnishing a perceptually salient but behaviourally minimal form of copresence.

**Stage 3: Completion of Psychometric Instruments**
Participants completed three psychometric questionnaires:
- **Empathizing Quotient (EQ)** [205], indexing affective resonance.
- **Systemizing Quotient (SQ)** [206], indexing rule-based cognitive preference.
- **Big Five Inventory-10 (BFI-10)** [207], capturing broad personality traits.

The inclusion of these instruments was mandated by the model component $\beta_C$, enabling quantification and later statistical control of individual differences. These measures prevent dispositional variance from masking or misattributing the perturbative effect of $\gamma_R$ on the evaluative conversion from $\alpha_E$ to $\delta_m$.

**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 $\delta_m$ 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.

### 5.3.6 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.

### 5.3.7 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 [3, 6]. 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 [86] and Złotowski et al. [101] 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 5.1: 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 $\alpha_E$ held constant, and with $\beta_C$ 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 $\mathscr{R}$.

### 5.3.8 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 ($\alpha_E$), dispositional factors ($\beta_C$), and robotic presence ($\gamma_R$). **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.

### 5.3.9 Interim Conclusion to Question 5.1

**Partial Conclusion to Question 5.1**

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.

### 5.3.10 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 5.2: 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 5.3: 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 5.2:** *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 5.2 and 5.3 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.

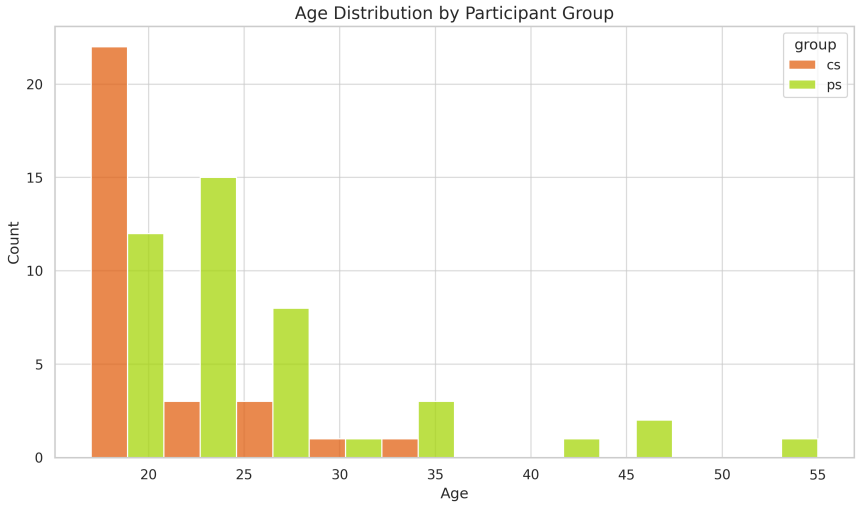



Figure 5.2: Age distribution across experimental conditions. Histogram representation confirms no major between-group demographic divergence.

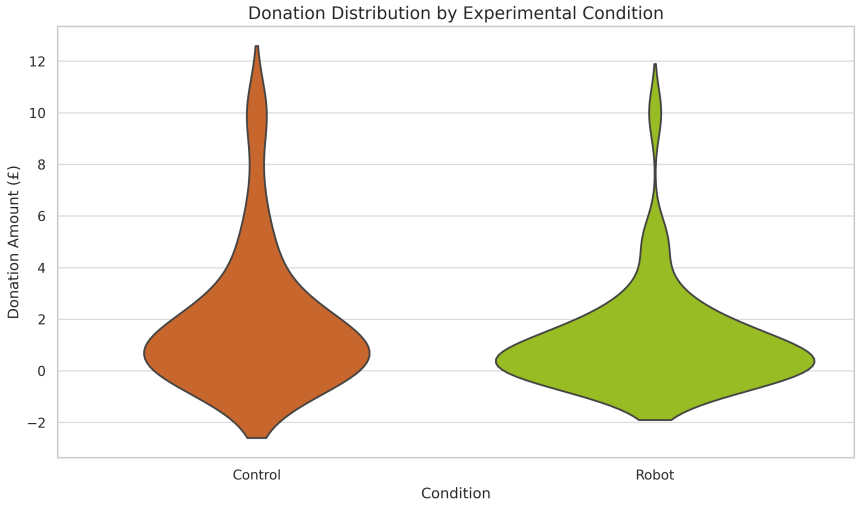



Figure 5.3: Distribution of donation behaviour by condition. Violin plot representation visualizes the heavier tail in the robot group, supporting the hypothesized interpretive perturbation.

### 5.3.11 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 ($\alpha_E$) 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 5.4: 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.

### 5.3.12 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 $\beta_C$.

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 $\gamma_R$ and individual traits $\beta_C$ 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 5.5: 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 $\gamma_R$ modulates the evaluative function $f(\alpha_E, \beta_C, \gamma_R)$ and how this modulation scales across heterogeneous configurations of the trait vector $\beta_C$.

### 5.3.13 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. 62) asserts that the expected outcome of moral behaviour, as computed by the evaluative trans-

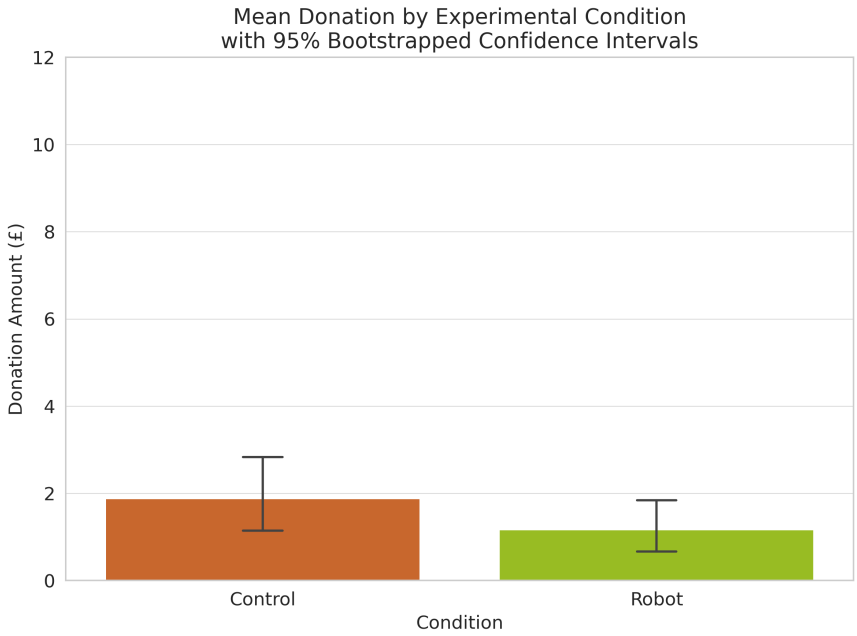



Figure 5.4: Mean donation amounts by experimental condition, with 95% bootstrapped confidence intervals. Participants in the Control condition donated more on average than those in the Robot condition, aligning with the hypothesis that robotic presence attenuates the inferential mapping from moral salience 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. 64) 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. 67) 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. 62) identifies the *empirical signature* of deformation at the level of expected behaviour.
- Hypothesis 2, (p. 64) explains the *ontological possibility* of such deformation within Floridi's informationalist framework and its Levels of Abstraction.
- Hypothesis 3, (p. 67) 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 ($\alpha_E$), dispositional factors ($\beta_C$), and robotic presence ($\gamma_R$). **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 $\gamma_R$ component, while leaving open the possibility that its effect is modulated by structured configurations of $\beta_C$—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.

### 5.3.14 Interim Conclusion to Question 5.1

**Partial Conclusion to Question 5.1**

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 $\beta_C$, and their interaction with robotic presence $\gamma_R$, 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.

### 5.3.15 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** $\Delta$ — 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** $\Delta$**:**

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

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 $\mathscr{R}$.

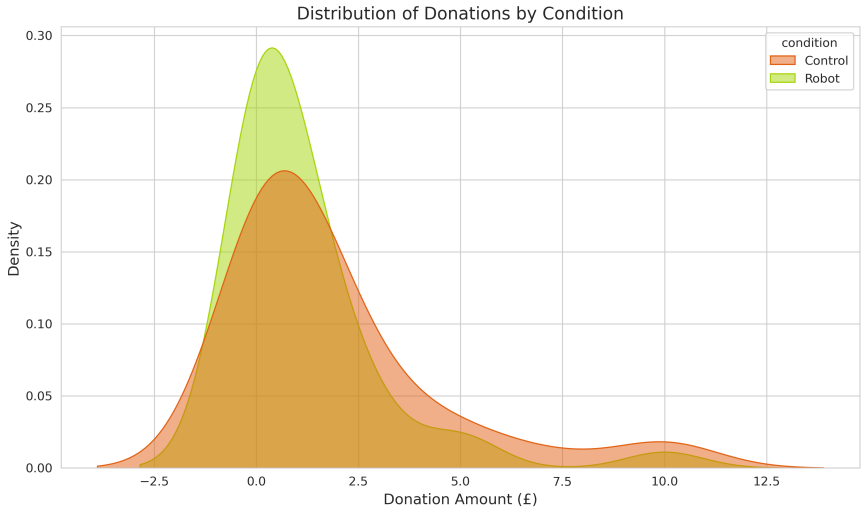



Figure 5.5: Kernel density estimates of donation distributions across conditions. The Control group exhibits higher central mass and a heavier rightward extension relative to the Robot group, consistent with a directional attenuation of high-value prosocial acts in the presence of the synthetic co-presence $\mathscr{R}$.

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:

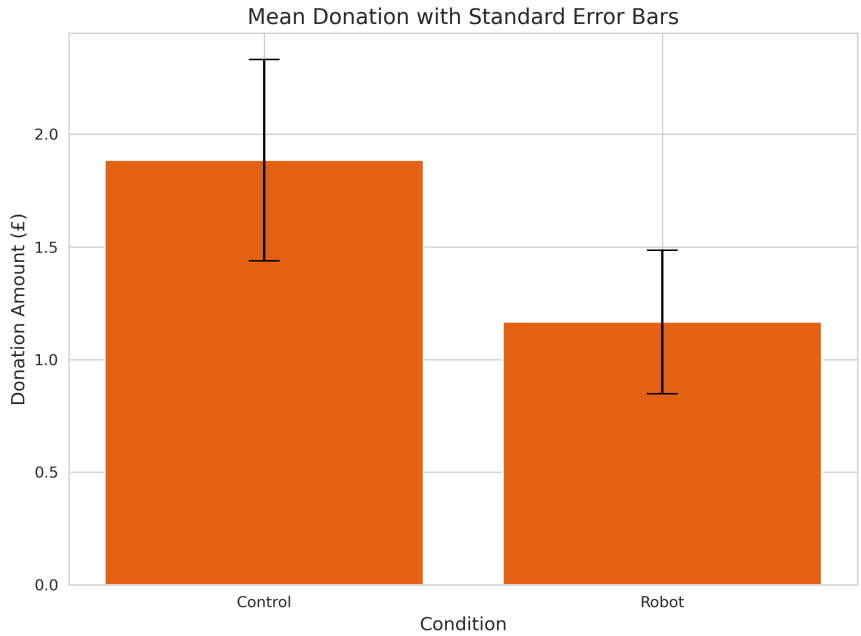



Figure 5.6: Mean donation amounts with standard error bars by condition. The Control group donates more on average (£1.89) than the Robot group (£1.17), corroborating the hypothesis that robotic presence modulates—rather than eliminates—the evaluative pathway from moral salience to action.

**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.

## 5.4 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?*

### 5.4.1 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.

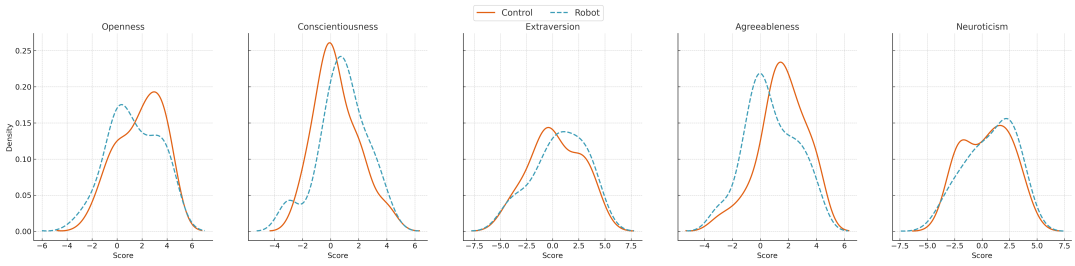



Figure 5.7: Kernel density estimates for each Big Five trait across experimental conditions, demonstrating substantial distributional overlap.

### 5.4.2 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 $\rho$ 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**.

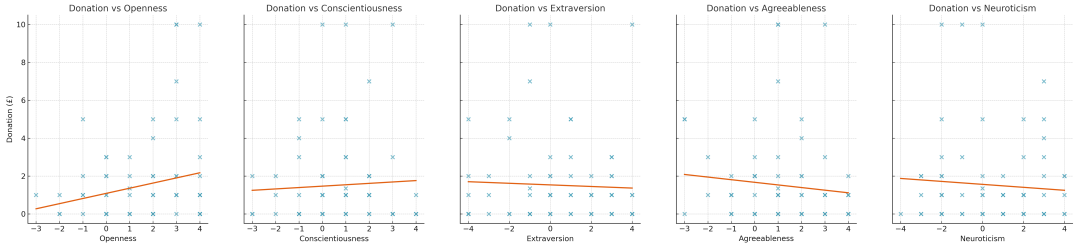



Figure 5.8: Scatter plots with fitted regression lines for each Big Five trait against donation amount. Each panel displays individual participant scores alongside a smoothed linear trend. No clear predictive relationships emerge, reinforcing the conclusion that the Big Five traits do not meaningfully predict prosocial donation within this experimental context.

### 5.4.3 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 $\gamma_R$ 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 $\beta_C$. 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.

### 5.4.4 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* $\beta_C$? 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 $\beta_C$ 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 5.9).

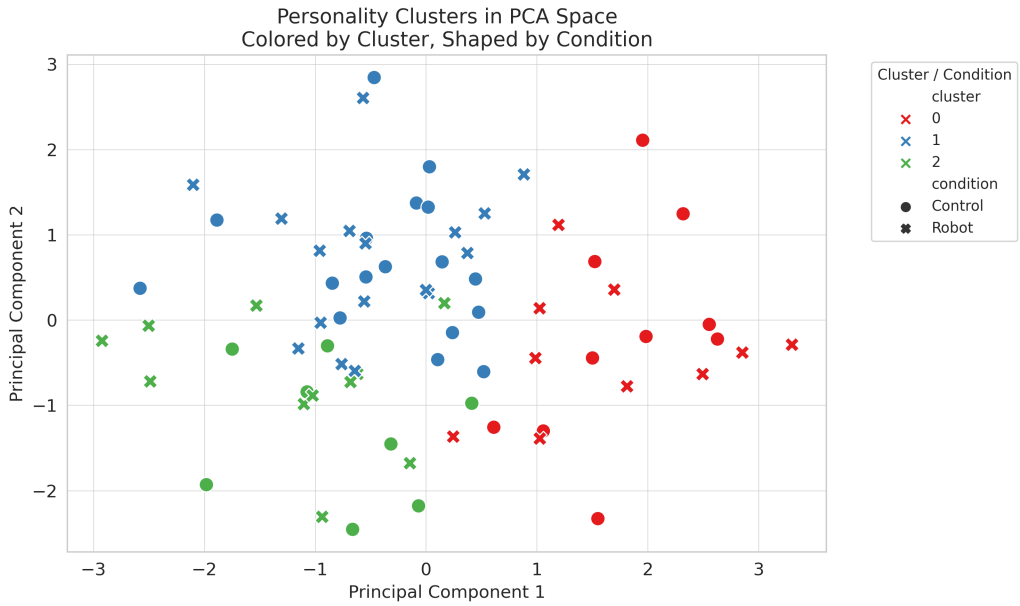



Figure 5.9: Participants clustered in PCA-reduced psychometric space, coloured by cluster identity and shaped by experimental condition. The clustering reveals three latent personality regimes, each representing a distinct cognitive–affective configuration encoded in $\beta_C$.

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 $\gamma_R$.

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).

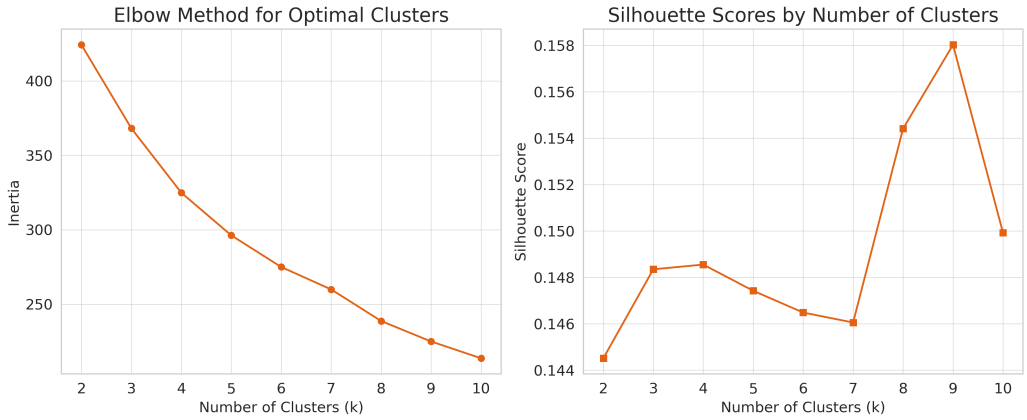



Figure 5.10: Elbow plot of within-cluster sum of squares (left axis) and silhouette coefficients (right axis) across candidate values of $k$. The elbow at $k = 3$ and interpretable silhouette profile support the selection of three clusters as a parsimonious and psychologically meaningful solution.

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 $\alpha_E$ and $\gamma_R$ 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 5.11). 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.

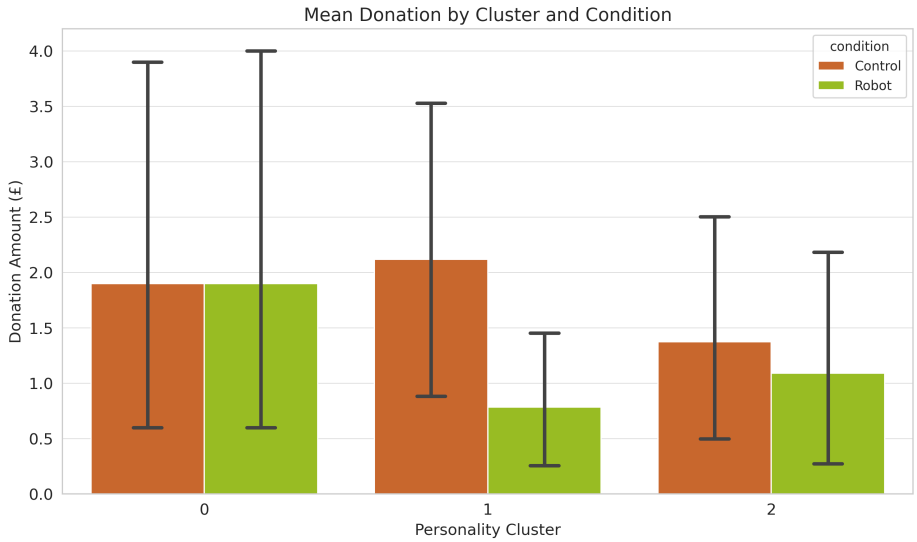



Figure 5.11: Mean donation amount by experimental condition within each personality cluster, derived from $k$-means analysis on psychometric trait profiles. Error bars represent standard deviation. Cluster 1 shows a marked attenuation of donation under robotic presence, whereas Clusters 0 and 2 exhibit minimal or modest differences. This pattern suggests that the perturbative effect of $\gamma_R$ is contingent upon latent cognitive–affective regimes encoded in $\beta_C$.

### 5.4.5 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.

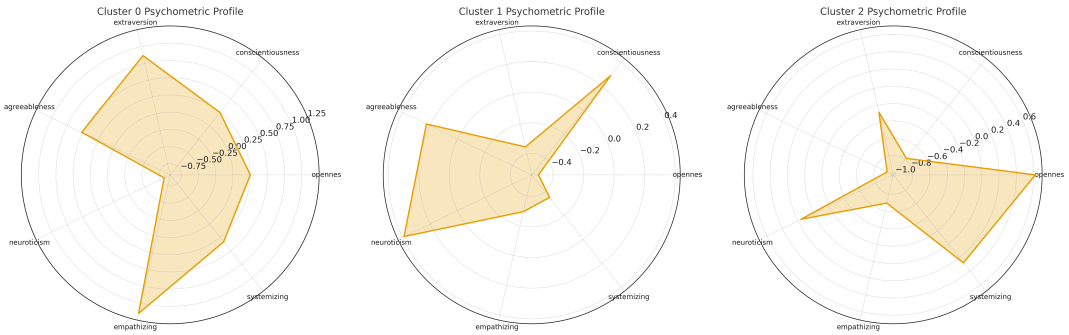



Figure 5.12: Comparative radar profiles of the three latent personality ecologies. **Emotionally Reactive / Low-Structure Profile** (left): elevated Neuroticism with reduced Conscientiousness and Systemizing. **Prosocial–Empathic / Warm–Sociable Profile** (centre): high Openness, Extraversion, Agreeableness, 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 $\gamma_R$ 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 $\gamma_R$ 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 $\gamma_R$ 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 $\beta_C$ 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, $\gamma_R$ 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 $\gamma_R$ 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 $\gamma_R$ 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.*

### 5.4.6 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 5.3.15);

(2) A structurally heterogeneous, cluster-contingent modulation of this attenuation (section 5.4.5).

Together, these findings show that robotic presence $\gamma_R$ 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 $\beta_C$.

*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 $\gamma_R$ modifies the evaluative mapping rather than merely shifting motivational baselines.

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

**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: – $\alpha_E$ held constant (Watching Eye), – $\beta_C$ refined via PCA and clustering, – $\gamma_R$ 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 $\alpha_E$ 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 $\gamma_R$ 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.

### 5.4.7 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 [167, 168, 163].

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 $\gamma_R$.

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 $\Sigma_{\text{eye}}$ 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 5.3.15) and the 'Contingent Structure of Cognitive Modulation' conclusion (section 5.4.5), 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.

### 5.4.8 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 $\gamma_R$ 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 $\gamma_R$ 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 $\gamma_R$ do not operate as independent forces; rather, they interact within a structured evaluative manifold whose topology differs across psychological regimes.

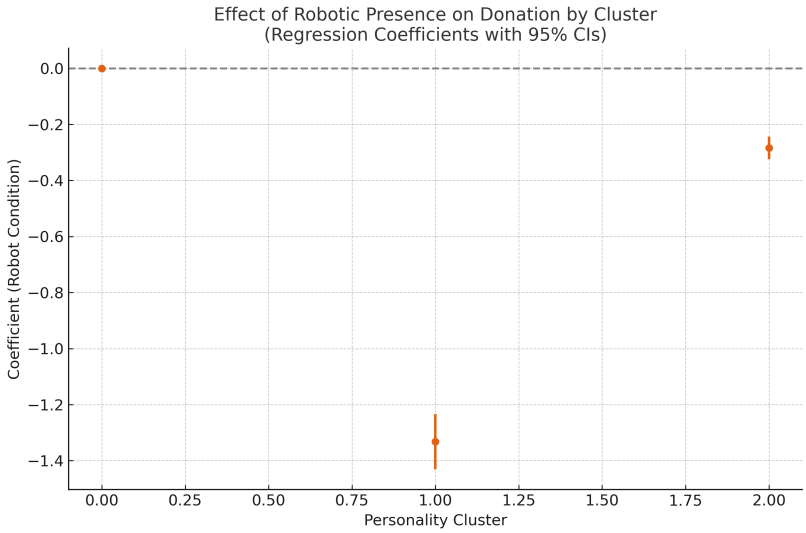



Figure 5.13: Regression coefficients for the Robot condition within each personality cluster (95% confidence intervals). The Prosocial–Empathic profile shows a pronounced attenuation effect, while the Emotionally Reactive and Analytical–Structured profiles exhibit negligible or non-significant coefficients. 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.

### 5.4.9 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.

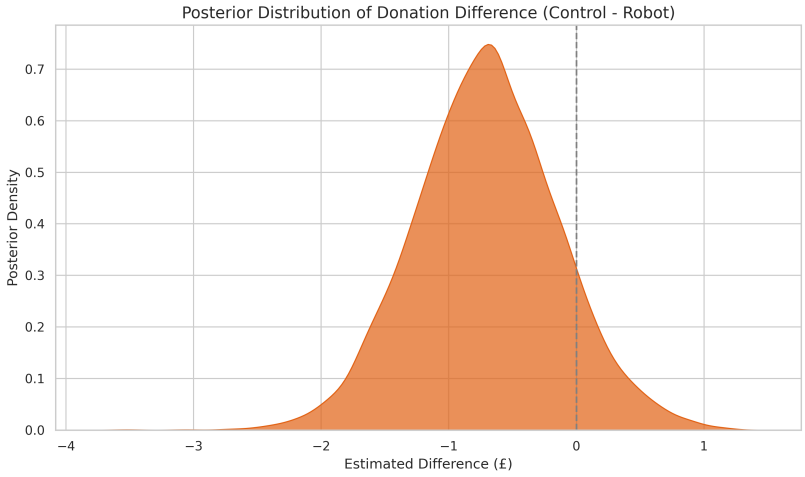



Figure 5.14: Posterior distribution of the estimated donation difference between the Control and Robot conditions. The density skews toward negative values, indicating directional probabilistic evidence that robotic co-presence attenuates prosocial behaviour. The vertical dashed line denotes the point 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 $\gamma_R$ 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 5.14 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 5.1. 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 $\gamma_R$ 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 5.1**

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 [113, 167, 168, 163]. 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 [23, **?**, 34], which holds that moral behaviour is driven primarily by rapid, affect-laden intuitions rather than by reflective cost–benefit deliberation [33, 203, 54]. 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* [133, 2, 3], 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* [96, 95] and *Affective Computing* [97, **?**], 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* [100, 6, **?**, 86], 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.

### 5.4.10 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 [23, 33, 34], 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* [133, 2]. 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 [96] and Affective Computing [97]. 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 [100, 131, **?**]. 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.

### 5.4.11 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 5.1 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 [113, 167, 163]. 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 [23, 33, 34]. 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 [133, 2, 3]. 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 [3]. 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, 208, 100, 209, 151]. 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, 208, 3].

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 [208, **?**, **?**]. 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 [7]. 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 [96, 97]. 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 [3, 210, 173]. 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.

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

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

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

[5] J. Deigh, *An introduction to ethics*. Cambridge University Press, 2010.

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

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

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

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

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

[11] L. Kohlberg, *Essays on Moral Development, Volume I: The Philosophy of Moral Development*. San Francisco, CA: Harper and Row, 1981.

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

[13] 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.

[14] 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.

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

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

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

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

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

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

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

[22] G. P. Goodwin and J. M. Darley, "The psychology of meta-ethics: Exploring objectivism," *Cognition*, vol. 106, no. 3, pp. 1339–1366, 2008.

[23] 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.

[24] 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.

[25] 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.

[26] J. Mikhail, "Universal moral grammar: Theory, evidence, and the future," *Trends in Cognitive Sciences*, vol. 11, no. 4, pp. 143–152, 2007.

[27] 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.

[28] D. Narvaez, "Triune ethics: The neurobiological roots of our multiple moralities," *New Ideas in Psychology*, vol. 26, no. 1, pp. 95–119, 2008.

[29] M. J. Crockett, "Models of morality," *Trends in Cognitive Sciences*, vol. 20, no. 2, pp. 87–97, 2016.

[30] L. Young and A. Waytz, "Moral cognition: A review," in *The Handbook of Social Psychology*, pp. 1–47, Oxford University Press, 5 ed., 2013.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] F. Cushman and L. Young, "The psychology of dilemmas and the philosophy of morality," *Ethics*, vol. 119, no. 4, pp. 597–636, 2009.

[39] F. Cushman and J. D. Greene, "Finding faults: How moral evaluations arise from normative frameworks," *Cognition*, vol. 136, no. 2, pp. 30–43, 2012.

[40] 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.

[41] J. D. Greene, "Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics," *Ethics*, vol. 124, no. 4, pp. 695–726, 2014.

[42] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[43] M. Smith, *The Moral Problem*. Blackwell, 1994.

[44] P. Railton, "Moral realism," *The Philosophical Review*, vol. 95, no. 2, pp. 163–207, 1986.

[45] S. Blackburn, *Ruling Passions*. Oxford University Press, 1998.

[46] A. Gibbard, *Wise Choices, Apt Feelings*. Harvard University Press, 1990.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] J. LeDoux, *The Emotional Brain*. Simon and Schuster, 1998.

[53] E. A. Phelps, "Emotion and cognition: insights from studies of the human amygdala," *Annual Review of Psychology*, vol. 57, pp. 27–53, 2006.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] J. Mikhail, "Universal moral grammar: Theory, evidence, and the future," *Trends in Cognitive Sciences*, vol. 11, no. 4, pp. 143–152, 2007.

[60] 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.

[61] F. Cushman and L. Young, "The psychology of dilemmas and the philosophy of morality," *Ethics*, vol. 119, no. 4, pp. 597–636, 2009.

[62] 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.

[63] 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.

[64] 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.

[65] F. Van Overwalle, "Social cognition and the brain: A meta-analysis," *Human Brain Mapping*, vol. 30, no. 3, pp. 829–858, 2009.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annual Review of Neuroscience*, vol. 24, pp. 167–202, 2001.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] H. Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, 2 ed., 2014.

[81] J. D. Greene, "The cognitive neuroscience of moral judgment and decision-making," *Handbook of Neuroethics*, pp. 161–178, 2014.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] M. Coeckelbergh, "Robot rights? towards a social-relational justification of moral consideration," *Ethics and Information Technology*, vol. 12, no. 3, pp. 209–221, 2010.

[87] M. Alfano, *Character as Moral Fiction*. Cambridge University Press, 2013.

[88] 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.

[89] 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.

[90] 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.

[91] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[92] M. Coeckelbergh, *AI Ethics*. MIT Press, 2020.

[93] 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.

[94] J. S. B. T. Evans, "Dual-processing accounts of reasoning, judgment, and social cognition," *Annual Review of Psychology*, vol. 59, pp. 255–278, 2008.

[95] 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.

[96] 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.

[97] R. W. Picard, *Affective Computing*. MIT Press, 1997.

[98] 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.

[99] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.

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

[101] 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.

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

[103] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[104] 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.

[105] C. D. Batson, *Altruism in Humans*. Oxford University Press, 2011.

[106] E. Fehr and S. Gachter, "Altruistic punishment in humans," *Nature*, vol. 415, pp. 137–140, 2002.

[107] J. Henrich *et al.*, "Economic man in cross-cultural perspective," *Behavioral and Brain Sciences*, vol. 28, no. 6, pp. 795–855, 2005.

[108] F. Warneken, "Precocious prosociality: Why do young children help?," *Child Development Perspectives*, vol. 9, no. 1, pp. 1–6, 2015.

[109] 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.

[110] M. J. Crockett, "Models of morality," *Trends in Cognitive Sciences*, vol. 20, no. 2, pp. 87–97, 2016.

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

[112] S. Darwall, *The Second-Person Standpoint*. Harvard University Press, 2006.

[113] 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.

[114] 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.

[115] M. Alfano, *Character as Moral Fiction*. Cambridge University Press, 2013.

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

[117] M. Bratman, *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

[118] D. Velleman, *The Possibility of Practical Reason*. Oxford University Press, 2000.

[119] P. F. Strawson, "Freedom and resentment," *Proceedings of the British Academy*, vol. 48, pp. 1–25, 1962.

[120] N. Arpaly, "Unprincipled virtues," *Oxford University Press*, 2003.

[121] P. Railton, "Moral learning: Conceptual foundations and normative relevance," *Cognition*, vol. 167, pp. 172–190, 2017.

[122] Aristotle, *Nicomachean Ethics*. Hackett, 1999.

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

[124] 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.

[125] I. Kant, *Groundwork of the Metaphysics of Morals*. Cambridge University Press, 1998.

[126] H. E. Allison, *Kant's Groundwork for the Metaphysics of Morals: A Commentary*. Oxford University Press, 2011.

[127] D. Hume, *A Treatise of Human Nature*. Oxford University Press, 2000.

[128] D. Hume, *An Enquiry Concerning the Principles of Morals*. Oxford University Press, 1998.

[129] R. Cohon, *Hume's Morality: Feeling and Fabrication*. Oxford University Press, 2008.

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

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

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

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

[134] B. M. McLaren, "Computational models of ethical reasoning: Challenges, initial steps, and future directions," *IEEE*, 2006.

[135] L. Kohlberg, "Stage and sequence: The cognitive-developmental approach to socialization," *Handbook of socialization theory and research*, vol. 347, p. 480, 1969.

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

[137] 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.

[138] H. Sidgwick, *The methods of ethics*. Cambridge University Press, 2019.

[139] R. Arkin, *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC, 2009.

[140] M. Anderson and S. L. Anderson, "Machine ethics: Creating an ethical intelligent agent," in *AI Magazine*, vol. 28, pp. 15–26, AAAI Press, 2007.

[141] 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.

[142] J.-G. Ganascia, "Modelling ethical rules of warfare," in *International Conference on Computer Ethics: Philosophical Enquiry (CEPE)*, pp. 181–190, 2007.

[143] 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.

[144] 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.

[145] C. Thornton, "Rethinking machine ethics in the light of virtue ethics," *Ethics and Information Technology*, vol. 15, no. 4, pp. 291–297, 2013.

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

[147] A. Pentland, "Social signal processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, 2007.

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

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

[150] 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.

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

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

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

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

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

[156] 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.

[157] 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.

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] M. F. Mason, A. Tatkin, and C. N. Macrae, "The look of love: Gaze shifts and person perception," *Psychological Science*, vol. 16, no. 3, pp. 234–237, 2005.

[163] 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.

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

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

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

[167] 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.

[168] 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.

[169] 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.

[170] 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.

[171] 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.

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

[173] 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.

[174] 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.

[175] 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.

[176] 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.

[177] 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.

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

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

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

[181] 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.

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

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

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

[185] 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.

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

[187] 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.

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

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

[190] 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.

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

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

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

[194] 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.

[195] 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.

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

[197] 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.

[198] 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.

[199] 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.

[200] 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.

[201] 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.

[202] 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.

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

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

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

[206] 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.

[207] 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.

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

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

[210] 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.