Synthetic Presence and the Topology of Moral Evaluation:

How Humanoid Robots Modulate Human Prosocial Action in Experimental Settings.

Francesco Perrone

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

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

November 2025

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

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

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

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

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

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

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

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

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

Francesco Perrone

Al mio compagno d'avventure, Francesco.

Abstract

Moral behaviour emerges not from isolated cognitive modules or explicit reasoning, but from a structured evaluative field shaped by attention, affect, salience, and dispositional architecture. This thesis develops a field-theoretic account of moral cognition grounded in empirical data, formal topology, and the philosophy of information. It argues that artificial systems—particularly humanoid robots—interact with this evaluative field in ways that conventional, rule-based Machine Ethics fails to capture.

A controlled behavioural experiment tested whether the silent presence of a humanoid robot (NAO) modulates prosocial giving under a strong moral cue (the Watching-Eye paradigm). Bayesian estimation and distribution-sensitive regression models reveal a modest but directionally consistent attenuation: participants donated less in the robot's presence despite identical moral affordances. Psychometric measures (EQ, SQ, BFI-10) were used to construct latent dispositional ecologies, which exhibited *heterogeneous* susceptibility to the perturbation. The Prosocial–Empathic cluster showed the strongest attenuation; the Analytical–Structured cluster showed little to none; the Emotionally Reactive cluster showed high variability. These findings confirm that synthetic presence interacts with evaluative topology rather than exerting a uniform behavioural effect.

Interpreted through the mapping $f(\alpha_E, \beta_C, \gamma_R)$ and Floridi's Levels of Abstraction, the results support a structural account in which synthetic presence functions as a perturbation operator. NAO's embodied ambiguity refracts intuitive appraisal: it alters attentional weighting, dilutes affective resonance, and redistributes moral salience before explicit reasoning is engaged. The robot does not supply moral content; it reshapes the perceptual and affective scaffolds through which such content acquires behavioural force.

The thesis concludes that artificial agents, even without agency or intent, act as modifiers of human moral environments. Moral behaviour is shown to be field-dependent, topologically organised, and sensitive to synthetic co-presence. These findings expose a limitation of top-down Machine Ethics and motivate an ecological model of artificial systems—one that prioritises the governance of the evaluative environments machines co-create with humans. The work establishes a methodological foundation for a computational and topological study of moral cognition in the age of artificial presence.

Contents

List of Tables

List of Figures

Acknowledgments

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 unhurried rhythm in which ideas ordinarily choose to unfold, and more by the unyielding cadence of an institution convinced that thought must keep pace with its deadlines. If these pages appear composed at a certain distance from themselves, it is because they bear the traces of that tension: the quiet struggle between what might have matured freely and what the present era insists must be shaped, concluded, and surrendered.

In that unsettled interval I have leaned on those whose presence never depended on the coherence of my arguments. This thesis is dedicated to my son, Francesco, who—like an improbable guardian with a light too bright for the corridors I found myself in—used the simple force of his existence to hold back a darkness I could not have held alone. To my mother, Mirella, whose illness and narrowing sight should have been met with my unhurried company rather than the prolonged absences this work imposed; to my father, Alberto, whose quiet vigilance, wisdom, and unwavering care have sustained us in ways that rarely leave a trace on any page; and to Anna, who has carried more than anyone her age should be asked to bear—not only for reasons that cannot be written here, but because she has returned, again and again, to the limits of my own intellect as though they were a place of rest rather than constraint. And to Dr. Herrera Martín, whose presence is not that of a colleague but of someone woven into the inner architecture of my life: through the most unsettled stretches of this work, he guarded the last spark of curiosity when it flickered low, and, with a steadiness all his own, helped the younger version of myself—long buried under years of urgency—to find a way back.

I also acknowledge, with the formality that such matters require, the support of the UK Engineering and Physical Sciences Research Council (EPSRC) through the grant Socially Competent Robots (EP/N035305/1), whose funding partially sustained the research that led to these pages. I thank Alessandro Vinciarelli for having—despite the many difficulties, interruptions, and divergences that marked these years—supported my formal progress through timely extensions and efforts to understand the intentions underlying my work. Whatever distances may have grown around the project, those gestures remain. I further thank Marek Sergot, who first opened this terrain to me during my time at Imperial College and whose early guidance set in motion curiosities that continue, in altered forms, to resonate here.

If this work lacks the calm of a true gestation, it rests nonetheless on the grace with which all of them, in different ways, have borne its cost.

Declaration

I confirm that I am the sole author of this thesis. With the exception of background sections that review existing literature and established theoretical frameworks (chapters 2, 4, and the non-original contextual material in chapters 3 and 8), all research reported here is my own independent work.

This includes the conception and design of the experiment, all data collection and preprocessing, the statistical and clustering analyses, and the development of the evaluative–topological framework used throughout the thesis, except where explicit acknowledgements are provided.

All sources have been appropriately cited, and no part of this thesis has been submitted for any other degree or qualification.

1. Introduction

Think of moral decision-making as the full mental sequence we go through when we're choosing between competing ideas of what the 'right thing' might be. It starts with what we notice: certain details stand out, others fade into the background. Those initial impressions shape what we care about, which in turn shapes what we treat as relevant. Only then does our reasoning step in to organise all of that into a sense of, 'This is what I should do.' In a way, it's the process that turns a handful of moral impressions into a genuine commitment to act.

And most of the time, this isn't a slow, deliberate calculation. It's closer to an immediate sense of something feeling right or wrong, which we then test against the situation and the social world around us. We respond to small cues—a shift in tone, a facial expression, the atmosphere of a room—and they quietly push us toward one reaction rather than another long before we begin to articulate reasons.

After that early, intuitive pull, we start to refine it. We call to mind similar situations. We notice details we missed at first glance. We talk it through, sometimes out loud, sometimes just internally. And we develop reasons that make sense of the direction we're already leaning toward. The decision is still real, but it grows out of these quick, socially shaped impressions that guide us well before any careful reflection begins.

This is precisely why the idea of creating a 'moral' machine by embedding a single ethical theory—utilitarianism, deontology, or any other framework—is so misguided. Those theories are helpful tools for analysing moral arguments after they've happened, but they're not the engines we rely on when we actually navigate a situation. They're abstractions, not working models of human judgment.

Yet in the technology world, you still encounter the view that if you program a system to follow a specific theory, you've solved the moral problem. That assumption is, at best, overly optimistic. A machine following a tidy rulebook bears little resemblance to what humans do when we sense tension in a room, register someone's discomfort, or feel the pull of how our actions will land with others. Real moral life is textured, social, emotional, and deeply dependent on context. There isn't a clean set of instructions that captures all that.

SUONA TROPPO PROCESSO ALLE INTENZIONI

So when somebody claims to have built an algorithm that 'acts ethically,' it often reflects an academic game of who can produce the most polished theoretical model rather than a meaningful engagement with how moral decisions actually work.

The theory may look elegant on paper, but it doesn't map onto the realities of human moral experience.

And this, is exactly the space where our work begins. We know that our moral reactions are shaped by tiny cues—someone's expression, the tension in their posture, the energy in a room, even things as subtle as the smell of someone who's had a long day. These details don't just colour the moment; they steer our judgment before we're even aware of it.

So the real question for us is this: what happens when the agent in front of you isn't a person at all, but a humanoid robot? How do we respond when the timing of a gaze is algorithmic, and the emotional tone is produced by design rather than by experience?

We still react. We can't help it. Our perceptual systems are tuned to pick up anything that looks or behaves like a person. But the meaning of those reactions becomes murkier. Are we responding to genuine social cues, or to clever mimicry? And if a robot can reliably trigger the same moral intuitions that another human does, what does that say about the foundations of our own judgments?

For us, that's the critical challenge. Not whether a machine can follow a rule-book, but *but how our deeply human, automatic moral instincts adapt—or fail to adapt—when something built rather than born is standing in front of us. And not just in a lab, but in our rooms, in our kitchens, woven into the background of daily life.*

Moral decision-making is:

> *The cognitive process through which agents select between competing moral judgments—mutually exclusive evaluations of what is right or wrong, good or bad—that provide the motive, direction, and justificatory structure of their practical behaviour. It is a composite operation: perceptual encoding, affective appraisal, memory, attentional orientation, and interpretive reasoning jointly determine how morally salient cues are registered, weighted, and transformed into a behavioural commitment.*

The work we present here developes withing this framework, which we applied to a concrete and experimentally tractable setting within Human–Robot Interaction and Social Signal Processing.

We conducted a study in which participants enter a small room and encounter a simple but meaningful moral choice: they may donate part of their participation payment to a real charity, or keep the full amount for themselves. This setting does not claim to capture moral cognition in its entirety; instead, it offers a minimal, controlled environment in which the elements of its definition become empirically observable.

Upon entering the room, participants first engage in **perceptual encoding**: they register the coins on the table, the charity materials, and the child-poster overhead with its large, expressive eyes. These elements constitute the *morally salient cues* structuring the situation, consistent with work showing that minimal observational cues and child-like eyes heighten perceived social relevance and implicit monitoring [1, 2, 3, 4, 5, 6].

Almost immediately, **affective appraisal** is recruited. The charitable context elicits a mild empathic pull in line with established findings on affective resonance and empathetic sensitivity [7, 8, 9]. Simultaneously, the watching-eye cue introduces an implicit sense of being observed, activating reputational and attentional systems documented in observational-cue research [1, 2, 5]. The prospect of giving up one's own money further evokes the familiar tension between prosocial motivation and self-interest captured in dual-process and motivational models of moral decision-making [10, 11, 12].

Alongside these immediate appraisals, **memory and normative expectations** shape interpretation: past experiences with charitable giving, internalised cultural norms of generosity, and well-established associations between being watched and acting prosocially influence how the evaluative field is instantiated in the moment [4, 2, 9].

At the same time, **attentional orientation** determines which elements dominate the evaluative landscape: is the participant more attuned to the need expressed by the charity? to the coins that could be kept?

To describe moral decision-making in this sense is to recognise its fundamentally *teleological* character, a view rooted in classical action-centred accounts of ethics [13, 14, 15]. Moral cognition unfolds toward action: it organises the evaluative conditions under which an agent adopts one course rather than another, consistent with empirical models linking appraisal to action selection [16, 10, 17]. The transition from moral judgment to behaviour is not an optional addendum to the process—it is its natural terminus. A moral evaluation that does not shape the field of possible actions has not yet completed its function; a moral action, conversely, is the crystallised endpoint of evaluative dynamics that have been unfolding long before reflection makes them explicit [18, 19, 16].

The participant's eventual choice to donate or not is the behavioural crystallisation of this entire evaluative process. This thesis examines how the silent co-presence of a humanoid robot modulates that transformation. The robot does not request, instruct, or communicate, yet its ambiguous social ontology—perceptually agentic, normatively indeterminate—reshapes the conditions under which moral judgments are formed and resolved. In this way, the experiment offers a precise instantiation of the definition of moral decision-making introduced above: a setting in which perceptual cues, affective resonance, attentional dynamics, and implicit social meaning combine to produce a practical moral commitment, and in which that process can be systematically perturbed.

Moral cognition thus operates within a social environment dense with cues—gaze, posture, interpersonal distance, implicit accountability signals—that modulate the affective and attentional components of evaluation. These modulations occur upstream of explicit reasoning: they determine *what becomes salient* well before agents deliberate on what *ought* to be done.

The introduction of synthetic agents into this environment raises a conceptual and empirical challenge. Humanoid robots occupy a liminal ontological space: perceptually social yet not persons, agent-shaped yet not agents. Their presence recruits perceptual and affective systems that evolved for human–human interaction, while simultaneously withholding the ordinary resources through which social meaning stabilises. This thesis examines the possibility that *such entities* *reshape the evaluative conditions of moral cognition not by acting, but simply by being present.*

One may picture the problem in concrete terms of our example above.Imagine the participant in the experimental room. On a table: the charity box, a few pound coins, and a simple instruction inviting a donation. The child in need, with big expressive eyes—an established prime of perceived accountability—looks down from a poster. Alone, the participant might experience a mild empathic pull, a subtle sense of being expected to act prosocially.

Now place a NAO robot on the same table. It does nothing. It does not speak, gesture, or request. Yet its humanoid shape, its forward posture, its apparent capacity for attention, reframes the scene. The participant hesitates: the social field has changed. Something in the evaluative machinery has shifted—an attenuation of empathic pull, a dilution of accountability, a re-weighting of salience.

We started by looking at something very simple: what happens when a humanoid robot is present in the room while someone is making a moral decision. The robot doesn't talk, it doesn't give instructions, it doesn't ask for anything. It just shares the space—quietly, almost like another person waiting their turn.

But that quiet presence turns out to matter. A robot like that sits in an odd position: it looks and moves in ways that make us treat it as an agent, yet we don't quite know what kind of 'being' it is or what norms apply to it. That ambiguity changes the atmosphere. It shifts how people interpret the situation, what they take to be appropriate, and how comfortable they feel committing to one judgment over another.

So even without speaking, the robot reshapes the background against which moral choices are made. It nudges the whole process—not by argument or instruction, but simply by being there, hovering between the familiar category of a person and the familiar category of a machine. That's where we see the transformation beginning.

This modest behavioural moment is the phenomenon under investigation.What has changed? And why?

The central question that follows from this observation frames the entire research programme:

> *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?*

This question is motivated by the theoretical claim that synthetic agents may function as *operators on the evaluative field* in which moral decisions are formed. If their perceptual salience or ambiguous social ontology alters the distribution of attention, empathy, or accountability, then the evaluative trajectory that links perception to action may shift accordingly. In such a case, moral behaviour would not be changed by explicit influence but by modulation of the cognitive–affective machinery upstream of conscious judgment.

In that case, moral behaviour wouldn't be shifting because the robot told anyone what to do. It would be shifting because the upstream machinery—the mix of perception, emotion, and expectation that feeds into conscious judgment—has been quietly modulated. The influence is silent, indirect, and deeply embedded in the way we make sense of the world. That's why this moment, small as it looks, matters.

1.1 From Research Question to Hypotheses: Framing the Investigative Architecture

Our question comes from a broader theoretical idea: that synthetic agents might operate on the moral landscape in which our decisions take shape. Not by persuasion, not by argument, but by subtly altering the conditions under which those judgments form. If a robot's visual presence, or the uncertainty about what kind of 'being' it is, changes where people direct their attention, or how much empathy they feel, or who they think is accountable, then the whole path from perception to action can start to bend.

If the simple presence of a synthetic agent shifts that chain of inferences,then the traditional approach in machine ethics—starting with abstract principles and trying to code them directly into a system [20, 21, 22, 23, 24]—can't explain what's going on. Those models operate at the reflective level, the level where we articulate reasons and moral rules. But the effects we're observing happen earlier, in the pre-reflective machinery that sets the stage for those reasons.

So we need a different way of thinking about moral behaviour. A framework that treats it as the outcome of a field shaped by attention, emotion, and the way certain cues stand out or fade away. In that view, moral action isn't just a conclusion drawn from a principle; it's the end point of a landscape structured by what feels salient, what draws concern, and what seems to matter in the moment. That's the level at which synthetic presence exerts its influence—and the level we have to model if we want to understand it.

One way to make sense of this is by borrowing a notion from Luciano Floridi: the Level of Abstraction [25, 26]. It's a simple idea with a lot of power behind it. Whenever we study a system—whether it's a computer, a person, a society—*we have to decide the level at which we're describing it.* Are we talking about the underlying code? The behaviour? The motivations? The social context? Each level reveals some things and hides others.

Most classical work in machine ethics starts at a very high, reflective level of abstraction. It focuses on principles—rules about what the system should or shouldn't do—and tries to formalise those rules so they can be implemented [27, 28, 29, 30]. That's useful if your goal is to build a system that behaves consistently with a particular ethical theory. But it tells you almost nothing about what happens at the cognitive level, where perception and emotion begin shaping the decision long before anyone appeals to a principle.

Our work sits at a different level of abstraction. We're looking at the machinery that turns raw perception into a sense of what matters, and then into action. At that level, the presence of a humanoid robot isn't a question about the robot's rights or intentions; it's a question about how its appearance and behaviour reshape the informational landscape the human is navigating.

Once we fix the Level of Abstraction—the cognitive level where perception, concern, and action are linked—we can be precise about what we're testing. The thesis proposes three hypotheses, each tied to a different kind of perturbation at that level. They're not rivals. They're three structurally distinct ways in which the presence of a synthetic agent might reshape the evaluative process itself. Each one captures a different mechanism through which the perceptual and affective landscape can shift before conscious judgment begins. The thesis therefore develops three hypotheses, each mapped onto a different kind of perturbation within the cognitive–affective system that generates moral judgment. They're not competing explanations; each one isolates a distinct structural route through which the simple presence of a synthetic agent might influence the transformation from perception to action.

Taken together, these hypotheses define the theoretical space of the project. They mark out the possibilities that become visible once we commit to the correct Level of Abstraction—the level where shifts in salience, attention, and affect reorganise the evaluative field long before a person arrives at a conscious moral conclusion.

The first hypothesis says that the robot changes the function that maps what you perceive to how you evaluate it.

Hypothesis 1: Evaluative Deformation

Synthetic presence alters the evaluative function $f : \mathcal{X} \to \mathcal{A}$ by reshaping salience gradients, affective weights, or attentional trajectories. In this model, the robot acts as a *field operator*: its perceptual salience deforms the topology through which moral cues acquire behavioural force.

The mathematical notation—$f : \mathcal{X} \to \mathcal{A}$—just means: given some input from the world, how do you turn it into a sense of what matters? What we test here is very simple: does having a humanoid robot in the room subtly shift what stands out to the subject, what feels important, or what pulls their attention?

If the robot is visually or socially salient—even without speaking—it might 'bend' the landscape you're navigating. Think of it like a small gravitational field: it doesn't tell you what to do, but it changes the shape of the space you're moving through. This hypothesis asks:

> *does the robot's presence deforms that evaluative landscape just enough to change how moral cues gain their force?*

The second hypothesis is about how people interpret responsibility and expectations in the presence of a humanoid robot. Here the claim is not that the robot has moral status or intentions. It's that its human-like appearance gives it certain practical effects in how people interpret the situation.

Hypothesis 2: Synthetic Normativity of Moral Displacement

A humanoid robot acquires *normative affordances* through its ambiguous social ontology. Without communicating or expressing intention, it may refract perceived accountability relations, modifying how agents interpret morally salient cues within the situation.

People may unconsciously treat it as if it participates in the moral scene, even though it hasn't said or done anything. So this hypothesis asks:

> *Does the robot shift who people feel accountable to, or who they think is paying attention, or what they think 'counts' in that moment?*

The robot's ambiguous status—something between a person and a tool—may subtly redirect moral attention. It's not giving orders; it's reframing the situation just by being there.

The third hypothesis looks at what happens in the transition from noticing something morally important to actually doing something about it.

Humans don't move straight from perception to action. There's a whole middle layer: empathy, emotional resonance, a sense of alignment with others. This hypothesis asks whether the robot interferes with that middle layer.

Does its presence dampen empathy? Does it redirect attention? Does it change how strongly certain cues 'tag' the situation as requiring action?

Hypothesis 3: Synthetic Perturbation of Moral Inference

Synthetic presence interferes with the transition from moral salience to prosocial action by modulating empathic resonance, affective tagging, or attentional alignment. This mechanism predicts differential perturbation across dispositional ecologies, precisely as observed in the experimental results.

So this final hypothesis says:

> *the robot doesn't change the rule you apply—it changes the internal bridge that links your moral perception to your moral behaviour*

And importantly, this hypothesis predicts that people with different dispositions—different personalities, sensitivities, backgrounds—will be affected differently. That's exactly what the experiments showed: the effect isn't uniform; it varies depending on the person.

These hypotheses structure the theoretical and empirical work that follows. They operationalise the core research question—whether synthetic presence can perturb the inferential machinery that links moral perception to moral action—and provide the conceptual scaffolding through which the experiment in Chapter ?? is interpreted.

BROKEN LINK

Together, these three hypotheses outline the whole space in which synthetic presence might influence moral judgment. Each captures a different mechanism, and all of them operate at the cognitive level—the level where perception and affect set the stage for what we later call 'a moral decision.'

1.2 The Need for a New Theoretical Orientation

All three hypotheses point to the same structural insight: the traditional tools of Machine Ethics operate at the wrong level of explanation. Classic Machine Ethics starts with high-level principles—rules, utilities, virtues—and then tries to engineer machines that follow them. That's perfectly coherent if your aim is to design a system that behaves consistently with an ethical theory.

But that framework doesn't touch the kind of phenomenon our research question is targeting. We're asking whether the presence of a synthetic agent reshapes the process by which humans move from perception to moral action. And that process unfolds long before anyone appeals to principles or reasons.

In other words, the phenomenon we're investigating doesn't live at the reflective Level of Abstraction. It shows up upstream, in the cognitive–affective machinery that makes moral reasoning possible in the first place. When a humanoid robot is in the room, it can alter what draws attention, how empathy is allocated, and what feels socially significant. That isn't a change in moral reasoning—it's a change in the conditions under which moral reasoning forms.

And if that's where the modulation happens, then a principle-first approach to moral AI can't explain it. We cannot start with abstract theories and work downward. You have to start with the architecture of moral cognition and work upward. Moral behaviour isn't just the outcome of applying a rule; it is the emergent trajectory of a system sculpted by perceptual salience, affective appraisal, and socially mediated cues—processes that moral psychology has shown to precede and shape explicit judgment [16, 31, 18, 32]. These evaluative dynamics are deeply sensitive to contextual modulation: shifts in attention, affective resonance, or perceived social presence can reconfigure the very pathway through which an agent moves from appraisal to action [6]. Artificial agents, even without agency or intention, participate in this structure by perturbing the field of salience and social meaning [33, 34, 35].

So the shift we're proposing isn't just methodological; it's conceptual. It reframes the core task of moral AI. Instead of asking, How can machines apply moral principles? we have to ask:

> *How do artificial agents alter the environment in which humans experience, interpret, and act on moral cues?*

That's the question that anchors the thesis. And later, when we look at the experimental results, we'll see why a principle-driven account simply can't capture the effects we observe.

The argument developed so far brings us to a decisive shift. The question that will guide the remainder of the thesis is no longer whether artificial agents can execute or approximate moral principles, but how their presence reshapes the very field in which humans perceive, interpret, and respond to moral cues. This reframing closes the introduction and opens the path to the theoretical and empirical work that follows.

1.3 Structure of the Thesis

The chapters that follow are arranged to make the implications of this shift increasingly explicit. The progression is cumulative. Each chapter establishes the conditions under which the next can be understood, and together they build a unified account of machine–mediated, machine-detactable moral cognition.

Chapter 2 establishes the philosophical and methodological ground of the thesis. It disentangles the two projects often grouped under Machine Ethics—Human–Machine Ethics and Computational Machine Ethics—and shows why neither operates at the cognitive Level of Abstraction required to explain synthetic moral perturbation. Drawing on normative ethics, moral psychology, and Social Signal Processing, the chapter argues that moral behaviour arises from a salience-weighted evaluative process rather than from the application of encoded principles. Its central conclusion introduces the core tension that motivates the thesis:

NON SONO SICURO
L'ACRONIMO SIA STATO
INTRODOTTO PRIMA. CHECK

> *Classical Machine Ethics works at the reflective LoA, while the phenomenon under investigation unfolds at the cognitive LoA, upstream of explicit moral reasoning.*

Chapter 3 provides the conceptual architecture needed to understand moral cognition empirically. It introduces dual-process theories, the Social Intuitionist Model, affective tagging, attentional capture, and accountability structures, illustrating how these mechanisms shape the path from moral perception to action. The chapter identifies the inferential gap: *the transformation from moral appraisal to moral behaviour.* This gap motivates the thesis's central question—whether synthetic presence can perturb that transformation—and prepares the reader for a systematic account of the evaluative processes at stake.

Chapter 4 specifies the methodological infrastructure through which the thesis renders evaluative cognition empirically tractable. Whereas the previous chapters developed the theoretical topology of moral appraisal, the present chapter introduces the instruments—psychometric, dispositional, and perturbational—that operationalise that topology in experimental form. It clarifies how established constructs from moral psychology, cognitive science, social signal processing, and

FORSE TROPPO PRESTO PER USARE FORMALISMI.

HRI serve not as neutral measurement devices but as theoretically motivated probes into the latent dispositional manifold modelled as β_C.

NON USARE PAROLE COME "DIMOSTRARE". MEGLIO "SHOWS".

By situating the Empathizing Quotient, the Systemizing Quotient, the Big Five Inventory, and the Watching–Eye paradigm within the evaluative–topological framework, the chapter ~~demonstrates~~ SHOWS that each tool targets a distinct dimension of the architecture through which moral salience is encoded, transformed, and expressed in behaviour. Their role is therefore conceptual rather than merely procedural: these instruments define the coordinate system in which the perturbation introduced by synthetic presence becomes detectable as a deformation of the evaluative field rather than as a trait-driven behavioural fluctuation.

> *The tools introduced here provide the empirical interface between theoretical topology and behavioural data: they operationalise the dispositional term β_C and supply the salience baselines against which synthetic perturbation can be identified.*

This chapter therefore establishes the measurement logic of the thesis. It shows why these specific instruments are required to distinguish dispositional variation from field-level modulation, and how they allow the experiment to test whether humanoid robotic presence alters not who participants are, but the evaluative topology within which their moral trajectories unfold.

Chapter 5 constitutes the empirical core of the thesis. It operationalises the evaluative–topological model developed in the earlier chapters into a full experimental framework, integrating design, measurement, and statistical inference into a single methodological architecture. The chapter introduces the controlled observational conditions, reconstructs the Watching–Eye paradigm, and justifies the use of the NAO platform as a parametrically stable source of synthetic presence. It specifies all behavioural measures, psychometric instruments, and salience manipulations, and it details the complete analytical pipeline—from preprocessing and cluster formation to non-parametric tests, regression modelling, and Bayesian estimation.

Its function is foundational: this is the chapter in which the three central hypotheses of the thesis—Evaluative Deformation, Synthetic Normativity, and Synthetic Perturbation of Moral Inference—are formally operationalised and subjected to empirical test. By consolidating the full experimental architecture with the statistical logic required to evaluate deformation in the evaluative field, the chapter provides the decisive evidence for the thesis' central claim: that synthetic copresence induces a measurable, structured alteration in the mapping from moral salience to action that cannot be reduced to trait-level variation or noise.

Chapter 7 reconstructs the major normative traditions—deontology, consequentialism, virtue ethics, sentimentalism, contractualism, particularism, and hybrid views—at the appropriate Level of Abstraction for the thesis. Instead of treating them as implementable rule systems, the chapter interprets their normative structures as patterns that constrain or guide evaluation within human moral cognition. Floridi's Level-of-Abstraction discipline is introduced here as a methodological tool for locating where an explanation must live. The chapter concludes by synthesising these perspectives into a coherent view of moral behaviour as a field-sensitive process shaped by both normative expectations and cognitive–affective dynamics. This synthesis provides the philosophical infrastructure that makes the subsequent hypotheses meaningful.

Chapter 8 provides the structural integration of the thesis. It unifies the cognitive–affective architecture, the normative analyses, and the experimental findings into a single theoretical account of how synthetic presence perturbs moral cognition. Building on the experimental result—uniform attenuation of prosocial donation under humanoid co-presence—the chapter shows that the effect cannot be understood as a trait-level phenomenon, a local behavioural anomaly, or a deficit of explicit reasoning. Instead, it requires a field-level interpretation: synthetic presence deforms the evaluative topology that ordinarily carries moral salience into action. By bringing together the three dispositional ecologies, the topological formalism, the reconstructed normative frameworks, and Floridi's Level-of-Abstraction analysis, the chapter argues that the humanoid robot operates as a perturbation operator on the moral field, not as an ethical agent. Its role is therefore decisive: it offers a general theoretical synthesis through which the empirical signature revealed by the data becomes a window into the structure of moral cognition and the methodological limits of Machine Ethics.

CLAIM MOLTO FORTE. RAC-COMANDO PRUDENZA.

Taken together, these chapters form a cumulative argumentative trajectory. Each chapter establishes the conditions of intelligibility for the next, guiding the reader from conceptual reframing to cognitive mechanism, from mechanism to experimental design, from empirical outcome to theoretical explanation. The result is a systematic account of how synthetic presence perturbs human moral cognition and what this means for the future of moral AI.

2. Literature Review

2.1 Introduction: Scope, Objectives, and Theoretical Commitments

This chapter establishes the conceptual and methodological terrain on which the remainder of the thesis proceeds. This review isn't just a background filler; but ratgehjr a the first test looking for the the assumptions which the experimental results depend on, the levels of abstraction they operate at, the mechanisms they take for granted, and the gaps they leave unexplained. It lets us ask:

> *Wheather the synthetic presence really does modulate the path from perception to action, which existing frameworks can even see that phenomenon? And which ones are blind to it by design?*

By examining the published work through that lens, we start to see an emerging pattern: almost all of classical Machine Ethics operates at the reflective level—principles, rules, deliberation—while the phenomenon we are studying unfolds at the cognitive level, upstream of reasoning. That mismatch isn't an opinion; it's a structural finding that the literature itself reveals.

The aim here is therefore to reposition the study of moral behaviour under artificial co-presence—and the design of artificial moral systems more broadly—within a theoretically unified space at the intersection of *Machine Ethics*, *Computational Morality*, and *Social Signal Processing* (SSP). Although these fields emerged from distinct disciplinary lineages, the experimental results presented in Chapter 5 show that they now converge around a single problem: artificial agents, even when silent, passive, and non-interactive, *modulate the evaluative conditions under which moral judgment and action unfold.* Understanding this phenomenon requires an integration of normative philosophy, moral psychology, computational modelling, and HRI.

Hence, the project takes root here. The literature review is the first piece of evidence. It shows that if we stay at the reflective level, we can't even formulate the right kind of question, let alone explain the modulation we later observe experimentally (Chapter 5). That's why the review matters so much—it's the tool that tells us where the explanation has to live before you collect single data points.

One of the core findings of the literature is that classical Machine Ethics starts from the wrong end of the problem. The whole tradition begins by taking high-level ethical theories—Kantian tests, utilitarian calculations, virtue templates, deontic logics—and trying to encode them as if they were models of moral agency [21, 20, 22, 23, 24, 27].

But if we look closely at what those theories actually do, they are not descriptions of how humans produce moral behaviour. They are descriptions of how humans *justify* moral behaviour after the fact. This distinction is explicit in modern moral philosophy: Kantian universalisability, utilitarian aggregation, and contractualist justification articulate reflective standards for assessing reasons, not cognitive processes for generating action [36, 37, 15]. They operate at a very high Level of Abstraction: they tell you what counts as a good reason, *not how a person comes to act in the first place* [25, 26].

It should be noted that while most of what traditionally falls under Machine Ethics—Computational Morality, formal deontic systems, encoded utility functions—belongs to the "*pre-LLM*" era, the limitation identified here does not evaporate with the advent of large language models. If anything, the arrival of LLMs makes the limitation more sharply visible.

Recent work demonstrates that LLMs can perform exceptionally well on reflective moral tasks: they generate sophisticated reasoning, balance competing principles, and provide normatively articulate justifications that map cleanly onto established ethical frameworks [38, 39, 40, 41, 42]. They also exhibit high performance on benchmarked moral analogy tasks and moral classification challenges [43, 44]. But all of this ability is situated at the reflective Level of Abstraction: the linguistic, justificatory, post-hoc LoA.

And humans do not act morally at that level. On every empirically supported account of moral cognition—from social intuitionism [16, 45], to dual-process theory [31, 46, 17], to affective neuroscience [47, 48, 49], to embodied and socially embedded models [50, 34, 51]—moral behaviour is driven by salience, affect, perceptual appraisal, social cues, and attentional orientation, not by the explicit application of normative principles. These processes sit one LoA below the linguistic–justificatory space in which LLMs operate.

Thus, although we now live in a "post-LLM" era, the fundamental issue is not that pre-LLM Machine Ethics was technically limited or symbolically brittle. The deeper problem is that both pre-LLM Machine Ethics and modern LLMs operate at the wrong Level of Abstraction if the goal is to model, predict, or understand human moral behaviour. This is precisely the mismatch Floridi's LoA discipline is designed to diagnose [25, 26]: moral justification and moral production belong to different descriptive orders. LLMs amplify the upper order; they leave the generative order untouched.

Chronologically, the pattern is straightforward:

- **Pre-LLM Machine Ethics** attempted to encode normative principles directly—deontic rules, utility functions, virtue schemas—and encountered the reflective/cognitive mismatch documented extensively in the literature [52, 23, 53, 54].
- **Post-LLM models** generate better principles, better explanations, and more articulate moral rhetoric, but they encounter the same mismatch, now at a higher level of linguistic sophistication [55, 56, 57, 58, 59].

The chronology therefore does not mark a methodological revolution; it exposes the persistence of a category error. The assumption that moral behaviour is fundamentally a matter of reasoning or principle-application has survived unchallenged into the LLM era. But contemporary empirical evidence shows that humans rarely deploy such reasoning in the production of moral action [17, 16, 60].

As several recent critical analyses emphasise, LLMs produce moral reasoning without moral cognition [59, 58, 61]. They resolve dilemmas fluently [1]; they do not reproduce the cognitive–affective processes by which humans come to feel that something is a dilemma in the first place. Moral language is not moral experience. Reflective justification is not perceptual-affective appraisal.

That is the chronological insight, if one seeks it: the technologies have evolved dramatically, but the underlying LoA mismatch remains unchanged. The surface has shifted; the category error has not budged. And that's really the hinge of this work:

> *Synthetic systems can now talk morality far better than they can participate in the conditions that shape moral action. The two are not the same.*

What becomes interesting, especially now, is that artificial systems are not just reasoning in the abstract; they're entering our environments. They're in phones, homes, classrooms, offices. Their presence affects how we behave, how we interpret situations, how we allocate attention.

So the shift isn't from 'pre-LLM Machine Ethics' to 'post-LLM Machine Ethics.' The shift is from seeing AI as an agent that reasons to seeing AI as an element in the cognitive ecology—something that reshapes the conditions in which human moral behaviour unfolds. Whether the system speaks like Kant or Shakespeare or your best friend is irrelevant if its presence still modulates the way people notice, feel, and act. That's the axis that matters. That is the core of this work. And this is where the category error comes in. Machine Ethics assumes that the principles of an ethical theory can be treated as the cognitive machinery of a moral agent–*as if humans behave by running Kantian tests or utilitarian calculations in their heads.*

But we know that isn't how moral action is produced. Human behaviour comes from a much lower level: from what captures our attention, what feels salient, how we read a face or a tone, how empathy gets triggered, how the context shifts our sense of what matters. These processes are fast, intuitive, emotional, and deeply social [16, 31, 18, 62, 6].

[1] The distinction between reflective and generative Levels of Abstraction (LoAs) is crucial here. Moral justification, principle-balancing, and linguistic explanation occur at a reflective LoA [25, 26]. Human moral behaviour, by contrast, arises from perceptual, affective, and socially embedded processes documented across moral psychology and social neuroscience [16, 31, 49, 48]. Recent analyses of LLM-based moral reasoning confirm that these models excel at reflective justification but do not reproduce the generative cognitive–affective mechanisms that produce moral action [59, 58, 61]. The arrival of LLMs therefore intensifies—rather than resolves—the LoA mismatch at the core of Machine Ethics.

Decades of work in moral psychology and neuroscience demonstrate that intuitive, affectively laden processes precede and shape explicit moral judgement [10, 16, 19]. The intuitive, affectively charged processes come first [16, 31, 18, 17]. They shape the space in which explicit reasoning even becomes possible: before reflection begins, appraisal mechanisms, empathic resonance, salience attribution, and motivational tagging have already constrained the field of viable responses [19, 62, 32]. The reflective story we tell afterwards might be coherent, but it is downstream of the machinery that actually drives behaviour [10, 16].

So when Machine Ethics takes ethical principles and treats them as if they were the generator of moral action, it is working at the wrong level entirely. It is replacing the justification of moral behaviour with the mechanism of moral behaviour, and those are not the same thing [15, 37, 14]. High-level principles articulate normative standards, but the processes that produce moral action operate at a far more fundamental cognitive–affective level [25, 26].

So when one tries to design a "moral machine" by encoding Kant or utilitarianism, one collapses these two levels of abstraction. One is treating reflective principles as if they were psychological mechanisms. And the literature shows very clearly that they are not. Ethical theories explain why an action can be defended; they do not explain how moral behaviour is formed [32, 14].

That is the central limitation the literature review exposes. It shows that classical Machine Ethics is methodologically elegant but cognitively misaligned. It is operating at the wrong level to even see the phenomenon we are investigating. As empirical work in moral psychology, affective neuroscience, SSP, and HRI repeatedly shows, the relevant causal structure lies in the evaluative substrate of salience, affect, and social interpretation—not in the reflective principles invoked after the fact [63, 33, 34, 35].

Classical Machine Ethics is beautifully constructed—methodologically elegant, logically clean—but it's operating at a level that's cognitively out of sync with where moral behaviour actually happens. It starts from principles, from rules, from reflective argumentation [21, 20, 23, 22]. But the causal work—*the thing* that actually drives behaviour—lives one layer down, in salience, emotion, attention, and social interpretation [16, 31, 18, 62, 6].

If we look at the wrong layer, we simply don't see the phenomenon we're investigating. And this problem carries over into what's usually called Computational Morality, just in a different form. Whether it's logic engines, preference aggregators, or the newer wave of LLM-based moral modelling, the assumption is the same: moral behaviour can be approximated by symbolic inference—by treating moral judgment as a reasoning problem [24, 27, 28, 56, 55].

But the last twenty years of empirical work tell a very different story. Most moral judgments don't start with slow deliberation; they start with fast, intuitive, emotionally charged appraisal [16, 10, 17]. They're shaped by who's present, how someone looks at you, the tone in the room, what feels at stake, and the affective and social cues embedded in the environment [18, 6, 63, 35]. It's a messy, context-sensitive process [19, 32]. When we try to model morality as if it were a chain of propositions—if A then B, if C then D— we are abstracting away the very machinery that actually produces behaviour in humans. And that's the machinery our experiment shows can be shifted by the simple presence of a humanoid robot [33, 34, 51].

In other words: the classical computational models are not wrong because the logic is bad. They're wrong because they are modelling the wrong thing. They are trying to capture moral reasoning, when the real action is happening in the evaluative landscape that sits underneath moral reasoning. That's the level where synthetic presence does its work.

In Chapter 3 we make very explicit that: *any model of moral behaviour that leaves out the cognitive–affective machinery and the social-signalling dynamics behind moral judgment is simply not describing human beings.* It becomes unstable both scientifically and philosophically. This is where the Level-of-Abstraction issue gets predominant. If we would take high-level moral theories—the reflective content, the principles, the rules—and treat them as if they were the psychological mechanism that produces moral behaviour, we would end up with theories that look elegant but don't actually predict what people do. They explain justification, not behaviour.We would develop artefacts; models that fail not because the logic is wrong, but because they're modelling the wrong layer of the system. This becomes plainly clear in the experiment in Chaper 5.

The robot we use has no beliefs, no goals, no intentions, and no communicative acts; it is not reasoning or attempting to influence participants. Yet research consistently shows that even minimally expressive or non-agentic robots modulate human social behaviour [33, 51, 64, 34]. These effects operate through changes in salience, attention, and perceived social presence rather than explicit reasoning [65, 2, 47, 49]. Such upstream perturbations cannot be captured by rule-based, utility-theoretic, or propositional models of morality, which mislocate moral action in reflective reasoning rather than in the intuitive, affective systems documented across moral psychology [16, 46, 17, 66, 53].

At this stage, the literature reveals a point that no strand of classical Machine Ethics has convincingly addressed. If the aim is to understand how humans behave morally in the presence of artificial agents—and to model that behaviour in a form that artificial systems can meaningfully operationalise—then the foundational assumptions of the field must be re-examined. Principle-first approaches, whether deontic, utilitarian, or virtue-theoretic, presuppose that moral norms can be implemented as explicit rules or evaluative operators [21, 20, 22, 23]. Yet empirical research in moral psychology and affective neuroscience shows that moral behaviour does not arise from rule application but from cognitively embedded processes of appraisal, salience detection, affective resonance, and social interpretation [16, 31, 18, 62, 6].

Thus, moral norms cannot be treated merely as rules to be encoded. They must be understood in terms of their *topological function*: the way they structure constraints, gradients, and permissible trajectories within the evaluative field through which moral perception is transformed into action [14, 13, 32]. Norms operate at a reflective Level of Abstraction, specifying justificatory structure rather than cognitive mechanism [15, 37, 25, 26]. Their behavioural influence depends on how they interact with, and are realised by, low-level cognitive–affective processes.

For the same reason, moral judgment cannot be modelled as pure reasoning or symbolic inference. Dual-process and intuitionist models demonstrate that intuitive, affectively charged appraisals precede reflective judgment and constrain the space of subsequent deliberation [10, 16, 17]. Attention, empathic resonance, perceptual salience, and social–contextual modulation shape the evaluative landscape long before propositional reasoning becomes active [19, 18, 63].

Nor can artificial agents be treated as carriers or executors of moral values. Research in HRI and Social Signal Processing shows that artificial systems act primarily as *modulators*—as elements within the environment that reshape salience, perceived social presence, accountability cues, and evaluative expectations [33, 34, 35, 51]. Their influence operates upstream of explicit judgment, altering the evaluative field within which moral decisions are formed.

Once the problem is reframed in this way, the broader picture becomes clear. The limitations of classical Machine Ethics are not failures of logic but failures of explanatory level. Its models operate at a reflective LoA and therefore cannot detect, let alone predict, the cognitive–affective perturbations that empirical research has consistently shown to drive moral behaviour. When the evaluative landscape is foregrounded, the phenomena that appeared mysterious or anomalous under classical formulations become theoretically tractable: synthetic presence exerts moral influence not by embodying values or executing principles, but by reshaping the generative conditions under which moral action emerges.

What follows, then, is not merely a synthesis of existing work but a structural reorganisation of the field. By applying Floridi's notion of a *Level of Abstraction* [25, 26] to the foundations of Machine Ethics for the first time, the literature review demonstrates that the field has been operating at an explanatory level incapable of capturing the mechanisms that actually generate moral behaviour. Classical approaches begin with reflective ethical theories—deontic logics, utilitarian calculi, virtue templates—and treat these as if they were computational models of moral agency [21, 20, 22, 23]. Yet moral psychology and affective neuroscience have shown consistently that moral action arises from perceptual salience, affective appraisal, attentional capture, and social meaning [16, 31, 18, 62, 6]. Social Signal Processing and HRI research further reveal that artificial agents perturb precisely these low-level evaluative dynamics [63, 33, 34, 35].

Through this reframing, the literature review achieves a clear result: it exposes a fundamental LoA mismatch at the heart of Machine Ethics and shows that no principle-first, rule-codification framework can access the phenomena under investigation. Moral norms operate at a reflective LoA, specifying justificatory relations [15, 37], whereas moral behaviour is produced at the cognitive LoA through the dynamic interplay of affect, salience, and social interpretation. By bringing these strands together, the review establishes an integrated conceptual framework in which *synthetic presence* becomes intelligible as a perturbation of the evaluative field itself—a theoretical insight that classical Machine Ethics could not formulate, and a necessary foundation for interpreting the empirical results of this thesis.

2.2 The Two Research Projects in Machine Ethics

Machine Ethics does not constitute a unified field in the way that English literature or molecular biology do. It lacks a single community, a shared methodology, and a cohesive disciplinary core. What the literature refers to as "Machine Ethics" is in fact an umbrella designation for two fundamentally different research programmes that ended up sharing a name. Their conflation is widespread in the literature, yet they operate at distinct Levels of Abstraction [25, 26] and aim to explain different phenomena.

The first programme is what I call *Human–Machine Ethics*. This strand examines how humans think, feel, and behave in the presence of artificial agents. It encompasses questions of accountability, agency displacement, social influence, norm perception, and moral risk. Its empirical backbone comes from Human–Robot Interaction, media psychology, and Social Signal Processing. Evidence from these domains shows that artificial systems—whether humanoid robots, embodied agents, or even minimally interactive media—systematically modulate attention, empathy, prosociality, and interpersonal expectations merely through their presence [67, 63, 33, 34, 51, 35]. This research programme aligns directly with the phenomenon investigated in this thesis: the modulation of human moral behaviour by a robot's silent co-presence.

The second programme is *Computational Machine Ethics*. This project attempts to design machines that make ethically adequate decisions by embedding moral theories into computational architectures. Deontic logics, utilitarian optimisation engines, rule-based ethical governors, and virtue-inspired templates all fall under this category [21, 20, 22, 23, 24, 27]. The central assumption is that moral behaviour can be generated by applying ethical principles at runtime, often via symbolic inference, constraint satisfaction, or rule execution. In this sense, Computational Machine Ethics treats moral judgement as a reasoning problem rather than as a perceptual–affective process.

The literature routinely conflates these two programmes, as if progress in one automatically informs the other. But they sit at different Levels of Abstraction and answer different explanatory questions: Human–Machine Ethics investigates how artificial systems modulate human evaluative processes, whereas Computational Machine Ethics attempts to construct artificial evaluative systems by formalising normative content.

The empirical results of this thesis underscore why this distinction is indispensable. Human–Machine Ethics predicts precisely the kind of modulation observed experimentally: even a non-interactive robot can reshape attentional and affective salience, thereby altering the evaluative conditions under which prosocial behaviour is generated [6, 18, 16]. Computational Machine Ethics, by contrast, is structurally incapable of recognising such modulation because it presupposes that moral behaviour is produced by reflective, principle-driven reasoning—an assumption contradicted by decades of work in moral psychology and affective neuroscience [31, 10, 62, 32].

Thus, the apparent lack of unity in "Machine Ethics" is not an artefact of interpretation but an accurate reflection of the field's conceptual structure. The label obscures two independent activities: one empirically grounded, concerned with how humans behave in sociotechnical environments; the other formally oriented, concerned with encoding ethical principles into artificial agents. Without maintaining this distinction, research risks becoming blind to the very phenomenon contemporary AI and robotics force us to confront: that artificial agents, even when passive, *modulate the evaluative field* through which human moral decisions take shape.

2.3 A Clarifying Perspective on Where This Work Belongs—and Where the Field Must Go

It is tempting to ask where this research "belongs." Does it fall under Affective Computing, with its emphasis on computational models of emotion [68]? Does it align with Human–Robot Interaction, where the behavioural consequences of artificial social agents are examined [33, 34, 51]? Or does it sit within moral psychology, which has spent decades analysing the cognitive and affective substrates of moral behaviour [16, 31, 18, 62]? Each discipline contributes an essential piece, but none, on its own, provides the conceptual framework needed to understand the phenomenon at stake. For the purposes of this thesis, the disciplinary label is secondary; the primary task is the conceptual clarification that makes the inquiry possible.

The central confusion this thesis confronts is not empirical but conceptual. For nearly two decades, work collected under the name "Machine Ethics" has blurred two fundamentally distinct enterprises: understanding how humans behave morally in sociotechnical settings, and designing machines that behave according to encoded ethical theories. These projects occupy different Levels of Abstraction [25, 26], draw on different forms of evidence, and target different explanatory aims. Treating them as a single field has produced a methodological entanglement in which elegant theories obscure the very phenomena they are meant to illuminate.

The distinction becomes clear once the discipline of Levels of Abstraction is applied. Human moral behaviour emerges at the cognitive LoA: it is shaped by perceptual salience, affective resonance, attentional dynamics, and social-cue interpretation [16, 10, 18, 6]. Ethical theories—Kantian, utilitarian, contractualist—operate at a reflective LoA concerned with justification rather than generation [15, 37]. When researchers treat high-LoA normative principles as if they were low-LoA psychological mechanisms, the result is not an incomplete theory but an artefact: a framework unable to predict behaviour, accommodate perturbations, or explain modulation phenomena.

The experimental findings of this thesis make this point explicit. A humanoid robot with no beliefs, goals, or communicative acts nevertheless alters the evaluative conditions under which humans convert moral perception into prosocial action. Such modulation does not arise from reflective reasoning; it arises from shifts in salience, affective alignment, and attentional orientation [31, 18, 6]. Any framework that models moral action as rule retrieval, utility computation, or principle execution remains blind to these dynamics because it operates at the wrong LoA.

This is why the disciplinary categorisation of the work is not the central issue. The point is not where the research should be filed but what becomes visible once conceptual discipline is restored. Through this lens, the field of Machine Ethics reorganises itself. *Human–Machine Ethics* emerges as an empirically grounded inquiry into how artificial agents modulate human evaluative processes [63, 51, 35]. *Computational Machine Ethics* reveals itself as a reflective programme concerned with principled design, centred on formalisms such as deontic logic [21, 20], utility maximisation [22], and virtue-engineering [23]. Both are legitimate, but conflating them obscures the cognitive phenomena that modern AI and robotics bring to the foreground.

Clarification, however, is only the first step. Once the LoA distinction is restored, one must ask what research agenda follows. The answer is both more modest and more ambitious than any principle-encoding programme. Moral behaviour is not computed; it is formed. It emerges from a dynamic evaluative field structured by affective gradients, perceptual cues, attentional flows, and socially mediated expectations [16, 10, 18, 63]. Artificial agents—robots, avatars, conversational AIs—modulate this field simply by entering it. A scientifically credible programme for moral AI must therefore begin not with ethics as a set of principles but with the architecture of moral cognition.

Three consequences follow immediately. **First**, empirical grounding becomes non-negotiable. Any model of moral behaviour must integrate findings from moral psychology, affective neuroscience, developmental research, HRI, and Social Signal Processing. A theory that cannot accommodate the influence of gaze, posture, co-presence, or anthropomorphic cues cannot accommodate human moral behaviour [1, 2, 6]. **Second**, artificial agents must be modelled as operators, not reasoners: their role is not to apply rules but to modulate the evaluative conditions under which humans act [33, 35, 34]. **Third**, normative theory must be interpreted topologically rather than procedurally: norms specify constraints, gradients, and attractors in the evaluative space through which behaviour flows [14, 13, 32].

This reframing also answers the practical question often posed by engineers: what is the actionable takeaway? The takeaway is not a new ethical theory, nor a list of rules to embed in code. It is the recognition that artificial agents shape human moral behaviour not by argument but by presence, not by reasoning but by salience, not by principles but by perceptual modulation. Designing systems without understanding the evaluative field they inhabit is a form of conceptual blindness.

The future of moral AI does not lie in machines that reason like philosophers, but in machines that coexist with humans in ways that can be predicted, understood, and—when necessary—constrained. Any credible programme must therefore begin where moral behaviour itself begins: within the evaluative machinery that transforms perception and affect into action.

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

Once the conceptual confusion is removed, the next step is to examine the machinery that actually produces moral behaviour. Here the empirical story is remarkably consistent. For nearly two decades, work in moral psychology, affective neuroscience, and behavioural science has converged on a single conclusion: moral judgment is not primarily a reasoning task but a *dual-process system*. Fast, intuitive, emotionally charged processes perform the bulk of the causal work. They respond to perceptual salience, attentional capture, empathic resonance, and situational demands [16, 31, 10]. Slower, reflective processes intervene later—often to justify, refine, or override the initial intuitive appraisal—but the initial appraisal performs the primary generative role [17, 18, 62].

This picture is reinforced by the major theoretical models in the field. Haidt's Social Intuitionist Model [16], Greene's neurocognitive dual-process framework [31, 10, 46], and Cushman's action-based inference models [17] all converge on the claim that moral evaluation begins with rapid, affectively valenced appraisals long before explicit reasoning is engaged. Neuroscientific findings corroborate this: affective tagging, motivational relevance, empathy circuitry, and social-interpretive processes are recruited early, often prior to conscious deliberation [19, 18, 6].

This stands in sharp contrast with the philosophical traditions on which classical Machine Ethics has historically relied. Kantian ethics, utilitarian frameworks, and contractualism articulate *justificatory* structures: universalisability conditions, value aggregation procedures, or principles governing the exchange of reasons [15, 37]. They are not intended as accounts of the psychological mechanisms that *produce* moral judgments. As the philosophers themselves emphasise, these theories operate at a reflective Level of Abstraction; they describe the standards by which actions can be defended, not the cognitive architecture through which actions arise.

Machine Ethics, however, adopted only this reflective dimension and treated it as though it described the entire system. It assumed that humans behave morally by applying principles, and that artificial agents could do likewise by encoding those principles directly into computational structures [21, 20, 22, 23]. But the empirical literature shows decisively that moral behaviour is not generated by rule application. It emerges from a cognitive–affective substrate shaped by salience, emotion, attention, embodiment, and social interpretation.

This empirical fact explains why studies of human moral behaviour in context—across HRI, media psychology, and Social Signal Processing—identify recurrent patterns governed by attentional capture, affective resonance, perceived monitoring, and contextual meaning [67, 63, 33, 34]. Consider the Watching-Eye effect: people alter their behaviour when exposed to minimal cues of observation, even a pair of stylised eyes [1, 2, 5]. The shift is not the result of endorsed rules but of subtle environmental modulation of evaluative posture.

This cognitive level—the level of salience, empathy, vigilance, and contextual modulation—is precisely where moral behaviour is shaped. It is also where the attenuation effect in our experiment resides. The humanoid robot does not reason, speak, or request anything; nonetheless, its silent co-presence perturbs the evaluative field sufficiently to alter prosocial action. This is the cognitive–affective layer in operation, the layer classical Machine Ethics never modelled.

What follows from this is analytically unavoidable. If moral behaviour emerges from perceptual salience, affective pull, attentional alignment, and social interpretation, then computational models that treat morality as rule-following or propositional inference are modelling the wrong phenomenon. They are elegant but descriptively incomplete: they capture the reflective Level of Abstraction while missing the cognitive Level of Abstraction entirely.

This is why the discussion moves next to Levels of Abstraction. Once the mismatch is recognised, it becomes clear that many of the philosophical debates and engineering efforts in Machine Ethics were conducted at an inappropriate explanatory level from the outset. The remainder of the thesis unpacks the consequences of this realisation and reconstructs a framework in which moral cognition, evaluative topology, and synthetic presence can be understood in principled alignment.

With this distinction in place, the argument can now shift from diagnosing the structural error in classical Machine Ethics to examining the positive framework required to replace it.

“FAILURE” E' UN TER- MINE MOL- TO CATEGO- RICO. TI E- SPONI A CRITICA AL- TRETTANTO CATEGORICA

2.5 Levels of Abstraction and the Failure of Machine Ethics

The conceptual tool that dissolves much of the confusion in Machine Ethics is Floridi's notion of a *Level of Abstraction* (LoA) [25, 26]. The idea is structurally simple but analytically powerful: any explanation requires specifying the level at which a system is being described. The LoA determines which variables are observable, the appropriate grain of detail, and the kinds of explanations that can legitimately be offered. Ethical theories operate at a high, reflective LoA: they articulate justificatory structures—principles, universalisation tests, value aggregation procedures, and reason-giving relations [15, 37]. Moral psychology, by contrast, operates at a lower, cognitive LoA: it investigates the mechanisms that *generate* moral judgment, including perceptual salience, affective appraisal, attentional dynamics, and social meaning [16, 31, 10, 17, 18].

Confusion arises when content belonging to one LoA is treated as if it were the mechanism operating at another. If reflective theories are misread as cognitive architectures, the distinction collapses, and with it the capacity to explain behavioural phenomena. Classical Machine Ethics has repeatedly committed this collapse for nearly two decades. By taking the principles of Kantian, utilitarian, or virtue-theoretic ethics and treating them as if they described the internal processes that produce moral behaviour, the field implicitly assumed that moral agents—human or artificial—act by applying principles [21, 20, 22, 23]. But these principles occupy the reflective LoA: they explain *why* an action might be defensible, not *how* a moral judgment is generated.

When these reflective principles are used as behavioural generators—as algorithms meant to produce moral action—the resulting models are elegant but fundamentally misaligned with human moral cognition. Real moral behaviour does not follow from propositional logic or rule execution. It emerges from what may be described as the *evaluative topology*: the structured field of salience gradients, affective forces, attentional pathways, and social interpretations that determine what appears morally significant in the moment [16, 46, 18, 63]. These low-level mechanisms—affective appraisal, empathic resonance, vigilance, contextual modulation—form the terrain within which high-level principles even acquire meaning.

The experimental findings of this thesis show precisely what happens when LoA discipline is violated. In the Watching-Eye paradigm, the accountability cue ordinarily increases prosocial behaviour [1, 2, 5]. Yet when a silent, non-agentic humanoid robot is introduced into the environment, this effect is attenuated. No reasoning, communication, belief, or intention is involved. The modulation arises from presence alone: the robot perturbs the evaluative field by shifting salience, affective alignment, and perceived social ontology [33, 34, 35]. The accountability cue loses traction not because a principle is misapplied, but because the cognitive substrate on which it depends has been displaced.

This synthesis yields a clear conclusion: moral action does not originate in the execution of principles but emerges from the dynamic interaction of perceptual, affective, and social processes. Classical Machine Ethics begins at the wrong point in the explanatory hierarchy. It treats high-level normative theories as if they were low-level cognitive mechanisms and thereby becomes blind to the central phenomenon that contemporary sociotechnical environments introduce: artificial agents modulating human evaluative fields through their mere presence.

The thesis therefore advances a strong and methodologically grounded claim: *before we can design moral machines, we must understand how machines reshape human moral experience*. This requires inverting the traditional order of explanation. The task is not to begin with ethical theory and push downward, but to begin with the empirical architecture of moral cognition, determine how artificial agents perturb it, and only then ask what forms of ethical oversight or design constraint are justified.

Once Levels of Abstraction are applied, the path forward becomes clear. We can distinguish coherent questions from incoherent ones, identify which debates were aimed at the wrong level of the system, and recover the conceptual clarity necessary for progress. More than any single empirical result, this restoration of LoA discipline is the tool that allows the broader project of moral AI to proceed in the right direction.

2.6 Evaluative Topology, Affective Architecture, and Synthetic Moral Perturbation

If the preceding sections establish that classical Machine Ethics operates at the wrong Level of Abstraction (LoA), the task now is to articulate the positive alternative: a topological account of moral behaviour grounded in the cognitive–affective mechanisms documented in empirical psychology [16, 31, 10, 17, 18, 62] and in the social-modulatory processes identified by Social Signal Processing and HRI [67, 63, 33, 34, 51].

The central thesis of this section is that moral behaviour does not arise from the execution of encoded principles. Instead, it emerges from the dynamic configuration of an *evaluative field*: a structured, multidimensional landscape shaped by gradients of salience, affective resonance, attentional pathways, contextual norms, and implicit social meaning. Ethical theories operate within this field not as algorithmic generators but as high-LoA structural constraints [15, 37]. Their force depends on how they are realised within the cognitive–affective dynamics through which moral perception becomes moral action.

2.6.1 The Evaluative Field

The notion of an evaluative topology synthesises three major strands of established research.

(1) Moral psychology: affect, intuition, appraisal. Dual-process theory [31, 10, 46] and the Social Intuitionist Model [16] show that moral evaluation begins with rapid, affectively valenced appraisals. Affective tagging, empathic resonance, and motivational relevance are recruited early [18, 62, 19]. Attentional capture, perceptual salience, and intuitive heuristics structure the evaluative space long before reflective reasoning is engaged.

(2) Social Signal Processing and affective computing: cue modulation. Work in SSP demonstrates that gaze direction, morphological cues, co-presence, and implicit monitoring reshape attentional and affective weighting long before explicit cognition intervenes [67, 63, 6]. HRI studies confirm that humanoid robots and artificial agents modulate social meaning and perceived agency through mere presence [33, 34, 35, 51].

(3) Normative theory: structural constraints. Philosophical ethics contributes the insight that moral theories provide structural invariants—deontological constraints [15], consequentialist gradients, virtue-theoretic attractors [14, 69], sentimentalist affective vectors [32, 70], and contractualist justificatory relations [37]. These normative forms define the shape of the evaluative field but do not generate behaviour directly.

Reinterpreted through LoA-sensitive analysis [25, 26], these strands form a coherent architecture: high-LoA normative structures supply the constraints; low-LoA cognitive–affective mechanisms determine the trajectories; and social signals reshape the field within which both operate.

2.6.2 Moral Behaviour as Trajectory

Within this topological framework, moral behaviour is best understood as movement through an evaluative manifold.

- **Attention** introduces local curvature by amplifying or suppressing cues [71].
- **Affect** saturates regions of the field with motivational energy [62, 18].
- **Contextual cues** deform gradients, shifting the relative weight of obligations, norms, and expectations [16, 4].
- **Social signals** modulate perceived accountability and interpersonal meaning [1, 2, 5].

This model dissolves the rationalist–intuitionist divide. Rationalist structures do not compete with intuitive mechanisms; they operate at different LoAs. The reflective domain imposes structural constraints, while the cognitive–affective domain determines how the system actually moves within those constraints [46, 37].

2.6.3 Synthetic Presence as Field Operator

The experiment presented in Chapter 5 provides an empirical probe into this architecture. The Watching-Eye cue ordinarily induces a prosocial salience gradient via implicit social monitoring [1, 2].

Yet the introduction of a silent, non-agentic robot attenuates this gradient. The effect does not originate in reasoning or principle-application. It arises from a deformation of the evaluative field itself. The robot's ambiguous social ontology—perceptually agentic but ontologically indeterminate—reshapes the affective and attentional conditions through which the Watching-Eye cue acquires behaviour-guiding force [33, 34, 35]. In this sense, synthetic presence acts as a *field operator*: its mere co-presence modifies the salience landscape and alters the trajectory from moral perception to moral action.

Crucially, the perturbation is *disposition-sensitive*.

- The **Prosocial–Empathic ecology** exhibits the strongest attenuation, reflecting its dependence on empathic resonance and interpersonal salience—the very mechanisms displaced by synthetic presence.
- The **Analytical–Structured ecology** shows moderate attenuation, consistent with reliance on interpretive coherence rather than affective pull.
- The **Emotionally Reactive ecology** shows minimal change, as its evaluative landscape lacks stable gradients onto which perturbation could anchor.

These differential effects underscore the core insight: synthetic presence perturbs moral behaviour *upstream* of principle, trait, and deliberation.

2.6.4 Topology and the Limits of Machine Ethics

This topological analysis explains why classical Machine Ethics could not predict the observed phenomenon. Moral behaviour under synthetic presence does not change because a rule is misapplied or because deliberation fails. It changes because the evaluative field in which principles acquire force has shifted.

- Deontological norms lose traction when accountability salience collapses [2, 71].
- Consequentialist gradients flatten when contextual meaning becomes ambiguous [4, 72].
- Virtue-theoretic dispositions cannot express themselves when affective attractors weaken [69, 14].
- Sentimentalist mechanisms fade when empathic resonance is displaced [32, 70].
- Contractualist justificatory relations dissolve when the perceived social field becomes indeterminate [37].

The experiment therefore confirms the structural thesis: moral behaviour is field-sensitive, and synthetic agents act as perturbation operators on the evaluative topology.

2.6.5 Toward a Unified Framework

The concept of evaluative topology provides precisely the integrative framework that Machine Ethics has lacked. It offers the structural bridge linking normative theory, empirical psychology, and computational modelling. It clarifies how high-LoA normative structures interface with low-LoA cognitive–affective mechanisms, and why artificial agents can reshape moral action without expressing beliefs, intentions, or normative content.

This framework completes the foundational turn of the thesis. The subsequent chapters build on this topological architecture to formalise a general model of machine-mediated moral cognition—one in which artificial systems are not ethical reasoners but modulators of the evaluative conditions through which moral meaning gains behavioural expression.

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

The analyses developed across this chapter converge on a unified account of moral behaviour under artificial co-presence. Classical Machine Ethics begins with reflective normative theories and treats them as behavioural generators [21, 20, 22]. Moral psychology shows that moral action instead emerges from a cognitive–affective architecture grounded in salience, attention, empathy, and contextual modulation [16, 31, 17]. Work in HRI and SSP demonstrates that artificial agents modulate these mechanisms through minimal social cues [63, 34, 33, 35]. Evaluative topology integrates these insights by modelling moral behaviour as trajectories through a salience-weighted, affectively structured field. The experiment confirms this: synthetic presence perturbs the evaluative field upstream of deliberation, thereby attenuating prosocial action.

Three core conclusions follow from the literature:

1. **Moral behaviour is generated at the cognitive LoA.** Reflective ethical theories articulate standards of justification [15, 37], but empirical work shows that behaviour is produced by low-LoA affective and social mechanisms [31, 18, 6]. Norms gain behavioural force only when the evaluative field affords it.

2. **Artificial agents reshape the evaluative field before they act within it.** SSP and HRI research indicates that presence alone modulates attention, empathy, vigilance, and perceived social meaning [67, 63, 33, 51]. The experimental attenuation effect confirms this literature-driven prediction.

3. **A viable programme for moral AI must begin with evaluative topology.** The literature shows that computational systems cannot generate moral behaviour through principle execution alone [23, 22]. Normative codification must be constrained by a model of the cognitive–affective architecture through which moral behaviour is actually formed.

These claims collectively reframe the foundational commitments of moral AI. Artificial systems cannot be conceptualised merely as executors of moral rules; they must be understood as *operators on the evaluative field* within which human moral cognition unfolds. Synthetic presence deforms salience gradients, attenuates empathic resonance, and weakens accountability cues—perturbations that occur far upstream of explicit reasoning.

2.8 Global Synthesis: From Inferential Displacement to Synthetic Moral Topology

The literature reviewed in this chapter reveals a coherent picture: moral judgment and action arise from a cognitively embedded, affectively structured, socially modulated evaluative field [16, 31, 18, 63]. Ethical theories supply reflective standards, but they do not generate behaviour; cognitive architecture does. Artificial agents participate in this architecture by shaping salience, affect, and perceived social meaning [34, 51].

2.8.1 From Question to Framework

The guiding research question—whether synthetic presence can perturb the inferential transformation through which moral salience becomes action—emerges naturally from unresolved tensions in the literature. Machine Ethics assumes that behaviour follows from principle execution [21, 20]; moral psychology shows it does not [16, 17]. SSP reveals that social cues modulate evaluative processes [67, 63]. HRI shows that artificial agents evoke these cues through minimal presence [33, 34]. Yet these strands have rarely been synthesised.

2.8.2 Why a Multi-Hypothesis Framework Was Needed

The literature identifies three distinct mechanisms through which artificial agents may modulate moral behaviour:

1. **Evaluative deformation** via shifts in salience, monitoring, and affective weighting [1, 2, 71].

2. **Synthetic normativity** arising from the perceived social ontology of robots [35, 33, 34].
3. **Perturbation of inferential pathways** through displacement of empathy, attention, or contextual interpretation [18, 46].

No single mechanism captures the phenomenon; a multi-hypothesis framework is required to align the interdisciplinary evidence.

2.8.3 What the Literature Alone Establishes

Across the reviewed domains, three findings are robust:

1. **Moral behaviour is field-sensitive**, emerging from salience, affect, attention, and contextual cues [31, 16, 18].
2. **Artificial agents modulate this field** by altering social meaning, vigilance, and empathic stance [34, 51, 73].
3. **Classical Machine Ethics cannot model this modulation**, because principle-based formalisms ignore the cognitive LoA where behaviour is actually generated [23, 22].

From this, a literature-driven conclusion follows: a viable framework for moral AI must be grounded not in normative content but in the structural dynamics of the evaluative field.

The literature review exposes a categorical error at the foundation of classical Machine Ethics. Across two decades of work, the same misalignment recurs: principles drawn from ethical theory—Kantian universalisability tests, utilitarian utilities, virtue-theoretic templates—are treated as if they were the psychological mechanisms that generate moral behaviour [21, 20, 22, 23]. Yet the literature makes clear that these operate at fundamentally different Levels of Abstraction. Reflective norms articulate *conditions of justification* [15, 37]; cognitive–affective systems explain *behavioural production* [16, 31, 17]. Frameworks that collapse these levels cannot predict or explain human moral behaviour, particularly under synthetic presence. The review makes this structural failure explicit.

The review also reveals a neglected architecture: moral cognition emerges from an evaluative field shaped by affect, salience, and social signalling. When empirical findings are placed side by side—across moral psychology [16, 10], affective neuroscience [18, 62], Social Signal Processing [67, 63], and Human–Robot Interaction [33, 34, 51]—a convergent picture becomes visible. Moral judgments originate in rapid, affect-laden appraisal; attentional dynamics determine which cues become morally salient; social signals such as gaze, posture, and co-presence modulate evaluative weighting; and explicit reasoning intervenes only downstream. This interdisciplinary convergence exposes a unified evaluative architecture that classical Machine Ethics never incorporated and could not accommodate.

Finally, the review identifies the theoretical gap that motivates the experiment. Once the evaluative architecture is made explicit, a precise, previously unformulated question emerges: *can synthetic presence perturb the evaluative field upstream of explicit moral reasoning?* No existing Machine Ethics framework even poses this question, because none operate at the LoA where such perturbations occur. The literature review therefore performs an essential scientific function: it isolates the causal layer in which moral behaviour is generated and shows that current models fail to explain modulation at this level. The empirical study is designed explicitly to probe this gap.

In short, the literature review demonstrates that the field has been asking the wrong questions at the wrong level of abstraction; it identifies the level at which the genuine causal machinery of moral behaviour operates; and it isolates the precise phenomenon requiring empirical investigation. It clears the conceptual ground on which the remainder of the thesis rests and provides the foundation for a new account of moral behaviour under synthetic presence. In this project, the literature review is not merely preparatory; it constitutes the first scientific result.

3.
Cognitive–Affective Architecture of Moral Judgement

EXCUSATIO NON PETITA ...

The conceptual apparatus developed in this chapter is not an ornamental introduction to moral theory. It is the minimum set of distinctions required to make the research question itself intelligible. The project asks:

> *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?*

that is to say whether the mere presence of a synthetic agent can alter the trajectory by which human beings transform a morally salient perception into a morally relevant action. Such a question does not belong to the domain of ethical theory; it belongs to the domain of moral cognition.

To address it, one must understand the cognitive–affective substrate in which moral judgments are formed, weighted, and enacted. The behaviour observed in the experiment does not arise at the level of explicit reasoning, rule application, or reflective justification. It arises upstream, within the processes that determine what becomes salient, how empathic resonance is allocated, which cues are attended to, and how the felt sense of accountability is modulated. These are the mechanisms through which moral evaluation becomes behaviourally operative; without a precise understanding of them, the central phenomenon of this thesis is not only unexplained but incorrectly described.

Reflective moral theories—Kantian maxims, utilitarian calculus, contractualist reasoning—do not operate at this level. They articulate justificatory relations, not generative mechanisms. They tell us why an action may be defensible, not how the human cognitive system produces the behaviour in the first place [15, 37, 36]. For this reason, any attempt to explain the experimental effect by appealing to ethical principles is methodologically misaligned.

It begins at a Level of Abstraction that the phenomenon does not inhabit [25, 26].

What is required instead is an account of moral cognition as an action-guiding evaluative process: a process in which affect, attention, salience, social interpretation, and contextual meaning jointly determine how moral cues acquire behavioural force. A large body of work in moral psychology and cognitive neuroscience demonstrates that these mechanisms—affective appraisal, empathic resonance, intuitive evaluation, and attentional modulation—constitute the causal substrate of moral judgment [16, 31, 10, 17, 18, 62]. Only within such a framework can the influence of synthetic presence be meaningfully specified. Without it, the experimental result risks being mischaracterised as a change in moral belief or a failure of deliberation, when in fact it is a perturbation of the evaluative field that precedes both [6, 4, 71].

The purpose of this chapter is therefore clarificatory in the strictest sense. It isolates the cognitive–affective mechanisms that constitute the causal substrate of moral behaviour; it distinguishes them from the reflective structures of ethical theory; and it establishes the Level of Abstraction at which the research question resides [25, 26].

By doing so, it provides the conceptual conditions under which the empirical findings of the thesis can be correctly interpreted. The experiment does not test principles, preferences, or doctrines. It tests the stability of the evaluative machinery through which moral meaning becomes action. Understanding that machinery is the only way to understand the phenomenon under investigation.

This is why the chapter must take the form it does. Not to broaden the discussion of morality, but to focus it precisely at the level where the phenomenon of synthetic moral perturbation arises.

A recurring theme across the reviewed literature is that failures in Machine Ethics stem from two related errors: *category mistakes* and *LoA conflation*.

Category mistakes arise when reflective normative principles are treated as if they described the psychological mechanisms that generate moral behaviour; LoA conflation occurs when descriptive cognitive regularities are mistaken for normative constraints or vice versa [25, 26]. Both errors follow from neglecting the fact that justificatory structures live at a high Level of Abstraction, whereas moral behaviour is produced at a lower, cognitive–affective LoA documented in moral psychology and social cognition [16, 10, 18].

Recognising these distinctions is methodologically essential: without LoA discipline, interpretive models of moral perturbation become confused, and empirical findings—such as the attenuation effects examined in this thesis (Chapter 5)—risk being mischaracterised as failures of reasoning rather than as deformations of the evaluative field.

3.1 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 [74], social–cognitive models [45, 75], affective neuroscience [19, 18], and evolutionary accounts of cooperation and prosociality [76, 77].

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 [69, 78, 79, 80].

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.

The distinction between descriptive and normative morality is introduced at this point in the chapter because it provides the final conceptual boundary required before the empirical and theoretical analysis can proceed. Without it, two serious confusions would arise—each of which would undermine the scientific aims of the project.

First, moral terminology in technical disciplines is often used ambiguously. Words like obligation, responsibility, harm, or trust are employed as if their meaning were self-evident, yet researchers oscillate unconsciously between describing how agents in fact behave and prescribing how they ought to behave. This sliding between domains produces conceptual instability: experimental findings are mistaken for ethical insights, and normative claims are misinterpreted as empirical predictions.

Second, the research question of this thesis is strictly descriptive: *Can synthetic presence alter the evaluative processes through which humans convert moral perception into moral action?*

To answer this question, the project must operate within the empirical domain of moral psychology. If this boundary is not explicitly marked, the analysis risks drifting into normative interpretation—treating behavioural attenuation as moral deficiency, or treating reflective theories as mechanistic explanations.

The descriptive–normative distinction therefore performs a crucial clarificatory function:

The distinction identifies the **level at which the thesis operates**. The aim is not to determine what people *should* do in the presence of robots, but to explain what *does* happen within the cognitive–affective architecture when artificial agents enter the evaluative field. Such phenomena require descriptive tools: models of attention, salience, empathy, and social meaning—not principles or moral doctrines [16, 10, 18, 71, 6].

Second, the distinction prevents the **misinterpretation of empirical findings** as moral judgments. If a robot's presence reduces prosocial behaviour, this is a psychological effect, not a moral failure. It does not imply that agents have acted wrongly or that the robot has transgressed any ethical boundary. It reflects a perturbation in the evaluative machinery that gives moral cues their behavioural force [1, 2, 5, 33, 34].

Third, the distinction isolates the **causally relevant components of morality** for the experiment. The mechanisms at stake—affective resonance, accountability salience, attentional modulation—belong entirely to descriptive cognition [18, 62, 63, 6]. Normative theories are indispensable for understanding the structure of moral reasoning, but they do not generate behaviour [37, 15, 36]. Keeping the domains separate ensures that the phenomenon is examined at the correct Level of Abstraction [25, 26].

Finally, the distinction prepares the ground for **integrating normative theory later without conceptual confusion**. Normative materials will reappear, not as behavioural engines, but as structural constraints within the evaluative topology—deontic invariants, consequentialist gradients, virtue-theoretic attractors, sentimentalist vectors, and contractualist equilibria [46, 14, 69, 81]. This reinterpretation is only possible once descriptive and normative domains have been clearly disentangled.

In sum, the distinction is introduced here because it secures the conceptual boundary conditions of the thesis. It establishes the domain in which the claims are made, prevents methodological conflation, and ensures that the phenomenon under investigation—moral perturbation under synthetic presence—is analysed at the level where it actually occurs. The orientation of the thesis is therefore precise: moral cognition is the object of study; normative theory provides the vocabulary of justification; and coherence requires that these domains remain distinct.

The project now turns to a minimal operational definition of morality. This may appear abrupt, but its placement at this point in the chapter is deliberate. The preceding sections established the conceptual boundaries required to analyse moral cognition without collapsing distinct Levels of Abstraction or importing normative assumptions into descriptive models. Having drawn these boundaries, the thesis now requires a definition precise enough to guide empirical and theoretical analysis, yet modest enough to avoid the philosophical commitments associated with substantive normative theories.

3.1.1 Why Definitions Vary

There is no single universally accepted definition of morality, and this plurality is neither accidental nor superficial. Different research programmes emphasise different elements of the moral domain. Cognitive approaches foreground the mechanisms by which agents form evaluative judgments [82]; affective traditions emphasise the emotional systems that underpin moral concern [83]; rationalist accounts privilege normative reasoning [80]; social-scientific models attend to conventions and cultural norms [84]; evolutionary frameworks focus on the adaptive functions of cooperation and prosociality [76, 77]. Philosophical traditions likewise diverge in grounding morality in rationality, sentiment, virtue, utility, social contracts, or evolutionary pressures.

Computational treatments often inherit only one strand of this diversity. They default to rule-based perspectives not because such models accurately describe human moral cognition, but because they are structurally convenient to implement [21, 20, 22, 23]. This convenience has encouraged the misleading interpretation of moral behaviour as rule following and has fostered oversimplified models of moral decision-making that obscure the cognitive–affective architecture through which real moral judgments are produced [16, 31, 17, 18].

A primary aim of this chapter is therefore corrective: to replace these inherited simplifications with a framework grounded in contemporary moral psychology, cognitive science, and social-signal research [67, 63, 6]. Only such a framework can support the empirical and conceptual analysis required by the research question.

3.1.2 Minimal Operational Definition for This Thesis

Within this clarified landscape, the thesis adopts the following minimal, action-oriented definition of moral cognition:

> *Moral cognition is the evaluative process through which agents detect normatively salient features of a situation, generate judgments concerning permissible or obligatory actions, and select behaviour accordingly.*

This definition is intentionally modest. It avoids entanglement in substantive normative theories while isolating the components necessary for empirical investigation: evaluation, judgment, and action. It reflects contemporary moral psychology, which treats moral cognition as the product of interacting affective and cognitive mechanisms [16, 31, 17, 18], and it coheres with the theoretical machinery developed throughout this thesis—evaluative topology, Levels of Abstraction [25, 26], and synthetic perturbation as documented in HRI and SSP [34, 51, 63].

Under this definition, moral cognition functions as a mapping from situational cues to action policies, shaped by trait-level dispositions [85, 86] and by the affective and attentional structures of the evaluative field [6, 71]. It provides the minimal conceptual anchor required to examine how synthetic presence modulates the transformation from moral perception to moral action.

Before proceeding to the distinction between factual and normative judgments, it is important to make explicit what has been achieved in the preceding sections. Although these sections are primarily conceptual, they perform essential scientific functions. They do not merely summarise philosophical background; rather, they establish the explanatory conditions under which the empirical and theoretical claims of the thesis become possible. Three achievements are central.

First, we have identifying the correct level of explanation for the research question. The literature review and the clarificatory sections that follow it isolate the cognitive–affective Level of Abstraction as the locus of the phenomenon under investigation. This is not a descriptive flourish: it is a scientific result. By showing that the perturbation induced by synthetic presence occurs upstream of explicit reasoning, these sections locate the causal substrate that must be modelled if the experiment is to be intelligible. Without this, the observed attenuation could not be interpreted without ideological or normative distortion.

Second, we have eliminated those category errors that distort empirical interpretation. The distinction between descriptive and normative domains, and the clarification of their respective inferential structures, remove a set of systematic mistakes that plague the technical literature. This is not conceptual housekeeping; it is **methodological decontamination**. By preventing the importation of prescriptive content into cognitive models—or the projection of cognitive regularities into normative claims—the chapter ensures that empirical outcomes are interpreted within the correct domain. This conceptual hygiene is a precondition for generating reliable scientific knowledge.

Third, we have established a minimal, action-guiding definition of moral cognition. The operational definition introduced in the previous section is itself a contribution. It provides the first precise specification of the cognitive object under study: moral cognition understood as an evaluative process connecting situational cues to action selection. This definition constrains the mechanisms that may legitimately be invoked as explanations—salience, affect, attention, social meaning—and excludes mechanisms that belong to the wrong LoA. It also provides the structural interface between empirical data and the evaluative-topological model developed later.

Collectively, these achievements secure the conceptual foundations of the thesis. They define the explanandum, delimit the explanatory layer, and prevent methodological conflation [25, 26]. Only after completing this work can the project turn to finer distinctions—such as the difference between factual and normative judgments—that further refine the architecture of moral cognition at the level where synthetic perturbation takes effect [37, 15].

This is why the next section follows naturally. Understanding moral perturbation requires understanding which kinds of judgments are being perturbed. Synthetic presence does not alter factual beliefs; it alters the evaluative force that connects normative appraisal to behaviour. The distinction between factual and normative judgment is therefore not decorative: it is the next analytic step in specifying the mechanism through which moral cognition is modulated [17, 16, 46].

3.2 Judgments: Factual and Normative

A central distinction for analysing moral cognition—and for understanding the experimental phenomenon at the heart of this thesis—is the difference between factual and normative judgments. Although both concern evaluations of situations, they operate at distinct logical and functional 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 remains frequently blurred in computational and psychological treatments of morality [87, 88]. Its importance here lies in the fact that:

> *synthetic perturbation affects normative judgment, even though the factual perception of the situation might remains unchanged.*

Because the synthetic perturbation operates selectively on the normative layer, we must first clarify what distinguishes normative judgment from the factual input on which it depends. Only then can we specify the mechanism that is being modulated.

Factual judgments derive their correctness from empirical features of the world; their truth depends on observation or inference. Normative judgments embed reasons for action—they carry prescriptive force even when tacitly represented [89, 80]. This is more than a semantic contrast. It marks a functional division within the cognitive architecture: judgments about what is engage classificatory and predictive systems, whereas judgments about what ought to be done recruit mechanisms that assign motivational weight, integrate affective cues, and generate the directional force that links evaluation to action.

This division maps directly onto the psychological conception of moral cognition, understood as the ensemble of perceptual, affective, and inferential processes that register morally salient features and transform them into evaluative representations [31, 16]. Moral cognition includes explicit moral judgment as well as the upstream mechanisms that detect salience, encode social meaning, and initiate the transition from appraisal to behaviour [17, 60]. The descriptive–normative distinction is mirrored in these systems: factual information is processed by mechanisms specialised for representational accuracy, while normative appraisal engages systems that confer action-guiding significance [19, 10, 90].

Psychological models therefore treat factual information as input to evaluative appraisal [91, 92, 93, 94]. Normative judgment requires an additional mapping: the transformation of descriptive cues into action-guiding evaluations [95, 96, 97]. Collapsing normative into factual judgment erases this architecture. For empirical research—and especially for paradigms measuring moral behaviour—maintaining this distinction prevents behavioural outputs from being mistaken for moral endorsement or internalised norms.

This separation also clarifies the mechanism probed by the experiment. Synthetic presence does not alter what participants believe about the scenario. *It alters how strongly normative force is experienced.* The attenuation effect is therefore not a change in factual judgment but a deformation of the evaluative dynamics that convert normative appraisal into action.

Recognising this prepares the ground for the next step. Once factual uptake and normative evaluation are disentangled, it becomes clear that moral judgment cannot be reduced to belief or emotion alone. It arises from the coordinated operation of perceptual, affective, inferential, and motivational systems that jointly confer normative authority and behavioural direction. It is this internal evaluative architecture—linking perception to action—that synthetic presence perturbs. To understand how such perturbation is possible, we now examine the structure of moral judgment itself.

3.3 Internal Architecture of Moral Judgment

Moral judgments are not mere expressions of preference or affective reaction. They exhibit a characteristic structure that combines evaluative content, justificatory grounding, and action-guiding force [98, 99, 100, 101, 102]. For the purposes of this thesis, a moral judgment involves at least three interlocking components:

1. **Salience detection**: the recognition that a situation contains normatively relevant features—harm, fairness, honesty, obligation, care. This process draws upon perceptual, affective, and social-cognitive systems [19, 18].
2. **Evaluative appraisal**: the 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 context and individual differences [83, 82].
3. **Practical commitment**: the formation of an action-guiding stance, in which the judgment functions as a reason for or against a particular behaviour [79, 80].

These components distinguish moral judgments from other evaluative acts—such as aesthetic impressions or strategic choices—and ground the thesis's operational conception of moral cognition as an **evaluative mapping** from situational cues to action policies. They also clarify why synthetic perturbation can alter behaviour without altering factual beliefs: the perturbation targets the mechanisms that assign motivational weight, not the mechanisms that register empirical information.

This tripartite structure accommodates both intuitive and deliberative models of moral judgment. Intuitive processes typically dominate in everyday moral encounters; yet even when reasons are not explicitly articulated, these judgments retain justificatory form [16, 96, 97, 94]. Conversely, deliberative processes involve explicit reasoning, counterfactual consideration, and appeals to principles or character traits [69]. This duality reflects not two kinds of morality, but two modes of access to the same evaluative architecture.

This distinction between intuitive and deliberative processes is not merely taxonomic; it initiates a deeper inquiry into the mechanisms that make moral judgment possible. To understand why certain stimuli reliably elicit prosocial behaviour whereas others disrupt or attenuate it, we must examine the architecture through which moral salience is perceived, represented, and acted upon. The transition from perception to appraisal, and from appraisal to action, is mediated by identifiable affective, perceptual, and executive systems, each contributing distinct computational roles within the broader evaluative ecology.

As the next section shows, contemporary psychological and neuroscientific research converges on a model of moral cognition as a distributed, dynamically interactive network. This framework clarifies how humans ordinarily navigate morally charged environments and provides the conceptual foundation for understanding how these processes may be perturbed—subtly yet systematically—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 substrate upon which the subsequent experimental analysis is built.

Understanding the internal architecture of moral judgment is not an abstract philosophical exercise. It is a methodological necessity imposed by the research question and the experimental paradigm developed in later chapters. The phenomenon under investigation—the attenuation of prosocial behaviour in the presence of a silent humanoid robot—occurs precisely within the architecture just described. Without a clear account of this architecture, the empirical effect would be unintelligible or, worse, misinterpreted.

The experiment demonstrates that the presence of a humanoid robot does not alter what participants believe about the situation. The factual content of the scenario remains stable. What changes is the normative force experienced in response to it: the directional pressure that transforms evaluative appraisal into action. Such a shift can only be understood if moral judgment is recognised as a composite process involving salience detection, affective appraisal, and practical commitment. The attenuation effect reveals a perturbation in one or more of these components—the curvature of the evaluative field—rather than any alteration in belief or principle.

This analysis also clarifies why the ontological ambiguity of the robot is central rather than incidental. The NAO robot used in the experiment possesses no beliefs, goals, or communicative intentions. Yet it is perceptually agentic: its morphology, gaze posture, and embodied presence activate social-cognitive mechanisms ordinarily reserved for human agents. This ambiguous status—more than an object, less than a person—positions the robot uniquely within the evaluative architecture. It can recruit salience-detection systems, modulate affective appraisal, or reshape perceived accountability without supplying any of the intentional content associated with genuine agency.

In other words, the robot functions not as a locus of moral claims but as a perturbation operator acting on the substrate that generates moral judgment. Recognising this requires precisely the distinctions drawn in this chapter: between descriptive and normative domains, between factual and normative judgments, and between intuitive and deliberative processes. These distinctions allow us to see what the empirical effect is—a deformation of the evaluative field—and what it is not: a change in belief, a failure of reasoning, or an abandonment of moral principle.

For the reader who has progressed to this point in the thesis, the significance should now be clear. The conceptual machinery developed in this chapter is not preparatory ornamentation; it is the explanatory foundation upon which the entire project rests. The experiment measures subtle changes in prosocial behaviour, but the theoretical contribution lies in explaining why such changes occur and how artificial agents exert influence within the cognitive–affective ecology of moral judgment. Only with a precise account of the internal architecture can the thesis articulate, diagnose, and ultimately theorise the phenomenon of synthetic moral perturbation.

This is the point where philosophical analysis, cognitive science, and experimental design converge. And it is within this convergent space that the remainder of the thesis will operate.

3.3.1 Psychological and Neuroscientific Foundations of Moral Decision-Making

A substantial body of cognitive neuroscience demonstrates that moral decision-making does not arise from a single "moral centre." Instead, it emerges from coordinated activity across affective, social-cognitive, and executive networks. These systems jointly determine how agents detect morally salient cues, generate evaluative appraisals, and select behaviour. The architecture is therefore inherently practical: the neural substrates implicated in moral judgment are also those responsible for valuation, behavioural control, and action selection.[1] Contemporary research thus situates moral judgment within a distributed computational system whose governing question is not "What is right?" but "What should I do here?" [103, 19, 96].

Affective and Value-Based Systems. The ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC) compute affective and motivational value, integrating emotional information with anticipated outcomes. Lesions to vmPFC disrupt the incorporation of social and emotional consequences into decision-making, producing choices that appear normatively inappropriate or insensitive to harm [103]. Functional imaging reveals vmPFC engagement during judgments involving interpersonal harm, care, and empathic concern [19]. Together, these findings show that moral judgments depend on mechanisms that encode the valence of behavioural options.

The amygdala and anterior insula provide early affective tagging for morally salient stimuli [104, 105, 106]. The amygdala detects threat, intentional aggression, and aversive outcomes [107, 47], while the anterior insula responds to disgust, norm violations, and aversive interoception [108, 109, 110]. Electrophysiological studies indicate that these affective signals often precede conscious deliberation [111, 112], functioning as rapid gating mechanisms for downstream moral appraisal.

Social-Cognitive and Interpretive Systems. Moral judgments frequently hinge on beliefs, intentions, and reasons [113, 114]. The temporo-parietal junction (TPJ), medial prefrontal cortex (mPFC), and posterior superior temporal sulcus (pSTS) form a network specialised for mental-state attribution [115, 116, 117, 118]. TPJ activation, for example, is reliably observed when distinguishing intentional from accidental harms or attributing blame or forgiveness [119, 17]. These systems ensure that moral cognition tracks reasons and intentions, not merely outcomes.

The anterior cingulate cortex (ACC) monitors conflict between competing evaluative signals [120, 121]. Classic moral dilemmas recruit ACC activity when intuitive emotional responses and reflective considerations collide [10, 122]. This conflict-monitoring function indicates that moral cognition involves arbitration among multiple evaluative forces [123, 124].

[1] This stands in contrast to folk-psychological depictions of moral judgment as passive contemplation of moral facts. Neuroscientific evidence overwhelmingly shows that moral cognition is organised around action guidance.

Executive and Action-Guidance Systems. The dorsolateral prefrontal cortex (dlPFC) supports controlled cognitive operations, including inhibition of affective impulses, representation of rules, and evaluation of long-term consequences [125, 126]. Disruption of dlPFC activity via TMS alters willingness to endorse instrumental harm [127, 128], demonstrating that this region contributes to structuring action policies that integrate affective, deontic, and goal-directed considerations [90, **?**].

Crucially, the dlPFC does not operate in isolation. Its interactions with vmPFC, ACC, and parietal regions reveal an integrated system in which valuation, social interpretation, and executive control jointly shape moral decisions [129, 130, 131]. Recent accounts describe this network as computing action-guiding commitments rather than abstract evaluations [132, 133].

This distributed architecture demonstrates a key claim that motivates the project:

> *moral decision-making is inherently action-oriented and computationally grounded in mechanisms of valuation, salience, and behavioural control.*

The experiment later introduced does not perturb beliefs, rules, or principles. It perturbs this action-guidance machinery—the very substrate through which moral salience becomes behaviour.

The neuroscientific evidence therefore provides the empirical foundation for the thesis's central argument: a silent humanoid robot does not need beliefs or intentions to influence moral behaviour. Its ambiguous social presence modulates the affective, attentional, and interpretive systems that constitute the architecture of moral judgment.

This is why the neuroscience matters, and why it belongs here in the argument: it shows, at the biological level, that morality is a process of evaluative action selection, and therefore vulnerable to the kinds of perturbation artificial agents can introduce.

Functional Integration and Practical Orientation. Across these subsystems, a coherent picture emerges: moral cognition is not a contest between "emotion" and "reason," but a dynamically integrated process in which affective valuation, social interpretation, and executive control jointly determine behaviour [16, 134, 135]. This integration is fundamentally practical. The vmPFC and OFC compute the affective value of potential actions [136, 137]; the TPJ and mPFC generate intention-sensitive interpretations of agents' behaviour [116, 119]; the ACC detects conflict between competing behavioural tendencies [120, 121]; and the dlPFC regulates whether intuitive impulses should be suppressed, enacted, or balanced against normative constraints [125, 127]. Even primary affective structures such as the amygdala and insula contribute to behavioural readiness by producing rapid somatic markers and prioritising morally relevant cues in the environment [47, 110].

Lesion studies, electrophysiological evidence, and neuroimaging findings converge on a single conclusion: moral judgment is an action-guidance mechanism operating under conditions of social meaning. On this view, moral cognition constitutes a form of evaluative control—a mapping from cue detection to practical commitment—rather than a detached assessment of abstract moral truths [96, 138]. This interpretation aligns with philosophical accounts emphasising the intrinsically action-directed nature of moral evaluation [79, 80], while grounding those commitments in empirical evidence about the neural architecture of agency, valuation, and control.

3.4 From Moral Architecture to Perturbation by Synthetic Agents

The integrated picture that emerges from cognitive neuroscience and psychology provides the conceptual bridge to the central phenomenon examined in this thesis. If moral judgment operates through distributed systems that compute *salience*, *affective weight*, and *behavioural readiness*, then **moral behaviour can be perturbed without altering beliefs or principles**. A humanoid robot need not issue commands or express intentions to exert influence: by reshaping the affective and attentional substrates of moral appraisal, it can modulate the likelihood that moral perception culminates in prosocial action.

This follows directly from the practical orientation of the moral architecture described earlier. Moral cognition is not an abstract exercise in principle-identification; it is a mechanism for transforming perceptual and affective cues into behaviour. Any alteration to the social or perceptual environment—particularly one involving the presence of an entity with ambiguous social status—can shift the evaluative computations that guide action. Later chapters develop this claim empirically, showing how synthetic presence attenuates the behavioural expression of moral salience (see Hypothesis 3 in Chapter 5).

A humanoid robot is especially revealing as a perturbation. It is *perceptually social* (in virtue of humanoid form), yet *ontologically indeterminate* (neither fully agentic nor behaviourally irrelevant). Such indeterminacy can disrupt attentional allocation, dampen affective resonance, and introduce uncertainty in mind attribution. These upstream shifts alter the weighting, timing, and accessibility of evaluative signals. In short: **the robot changes the evaluative conditions under which moral appraisal becomes moral action**.

Understanding this architecture is therefore indispensable for interpreting the empirical findings. The experiment does not measure abstract moral judgments but the *practical enactment* of moral cognition in an environment subtly transformed by synthetic presence. The neuroscientific foundations surveyed here provide the scaffolding for explaining how a silent observer can attenuate prosocial behaviour in stable, measurable ways.

A final conceptual step is required. If moral cognition is an architecture for transforming evaluative information into action, then **any alteration to the informational field is, in principle, a moral intervention**. A humanoid robot—an entity shaped like a person, yet not one—constitutes such an intervention. It does not supply new moral content; it *reconfigures the conditions under which content becomes operative.* The moral landscape is therefore not defined only by principles or dispositions, but by the *topology of the environment* in which they are enacted.

This insight has two consequences that structure the remainder of the thesis.

First, it shifts the explanatory centre of gravity: from conscious deliberation to the *situated dynamics of evaluative processing.* The experiment asks how moral cognition functions when confronted with an entity whose social meaning is ambiguous.

Second, it reframes the normative question. The significance of artificial agents lies not merely in what they do, but in how their *mere presence* modifies the normative affordances of a shared environment. Artificial agents reshape the moral field long before any explicit moral reasoning occurs.

In this way, the Moral Primer prepares two convergent lines of inquiry. The empirical chapters show how minimal synthetic presence modulates the behavioural expression of moral cognition. The normative chapters argue that this modulation exposes a structural oversight within classical Machine Ethics: the assumption that moral agency can be understood independently of the *environmental scaffolds* that shape human evaluation.

These threads suggest a view of artificial agents not as moral subjects or mere tools, but as *operators on moral space*—entities capable of bending the pathways through which moral meaning becomes action. The full implications of this perspective emerge only once the empirical and philosophical analyses are brought into dialogue. For now, it suffices to note that understanding moral decision-making under conditions of social and ontological ambiguity is not preparatory background; it is the *conceptual linchpin* of the entire thesis.

This conceptual foundation also illuminates the methodological commitments that follow: the *Level of Abstraction* at which moral cognition is analysed, and the *topological structure* of evaluative processes under perturbation. An LoA, in Floridi's sense, fixes the informational distinctions that matter for explanation. Here, our LoA does not concern the metaphysics of moral agency nor the justification of principles, but the *functional transformation* of perceptual and affective cues into action-guiding evaluation. At this LoA, robots are not modelled as moral agents but as *modulators of the evaluative field.*[2]

Once this LoA is fixed, moral cognition can be modelled topologically: as a system mapping inputs to behavioural outputs through a structure shaped by salience, attention, affective resonance, and interpretive inference. Changing the environment—in this case by introducing a synthetic observer—can therefore be understood as a *deformation* of the evaluative landscape. The experiment developed later investigates precisely such a deformation.

This topological perspective also clarifies why synthetic agents matter ethically even when behaviourally inert [139, 140]. At our operative LoA, the morally relevant property of a robot is its ability to *warp attentional and affective gradients* that structure human appraisal [141, 142]. A robot can function as a normative

[2]On LoA as a methodological device for analysing informational systems, see Floridi 2010, 2011, 2013.

deflector or semantic attractor, subtly redistributing the vectors through which moral salience exerts its pull [143, 144]. Later empirical chapters document these redistributions; later normative chapters examine how they challenge Machine Ethics, which typically locates moral significance in the agent rather than the *perturbation it induces* [145, 146].

Seen through this joint lens of LoA and moral topology, the empirical question at the heart of the thesis takes clear shape:

> *Does the presence of a synthetic agent reshape the evaluative field in which humans convert moral perception into prosocial action?*

The formalism

MANCA UNA DESCRIZIONE DEI SIMBOLI

$$f : \Sigma \to \Delta, \qquad \mathcal{P}_{\mathscr{R}} : \Sigma \to \Sigma', \qquad f_{\mathscr{R}} = f \circ \mathcal{P}_{\mathscr{R}}$$

offers a conceptual anchor—nothing more than a vocabulary—for expressing this claim: robotic presence functions as a perturbation operator on the evaluative field.

3.4.1 Philosophical Synthesis

This framework reframes perennial philosophical disputes. A Kantian model locates moral authority in rational principle; an Aristotelian model situates it in cultivated perception; a Humean model grounds it in sentiment and intuitive appraisal. The cognitive–affective architecture described earlier aligns most closely with the Humean–Aristotelian hybrid: moral judgment is rooted in *evaluative sensitivity*, not detached rationality. When the social world is reconfigured—when its cues are displaced or reframed—the moral response shifts accordingly.

3.4.2 Concluding Perspective: Why This Matters for the Thesis

The preceding analysis converges on a single insight: **robots reshape the evaluative topology of moral life**, not by reasoning, nor by instructing, but by altering the perceptual–social gradients through which moral meaning becomes behaviour.

The experimental chapters test this claim; the normative chapters show why it challenges the foundational assumptions of Machine Ethics. What emerges is a technomoral thesis: as artificial agents permeate human environments, they will inevitably reshape the *topology of moral experience*—subtly, silently, and often without intention. This is why synthetic presence matters. This is why the experiment matters. And this is why the conceptual groundwork laid in this chapter is essential for everything that follows.

The claim that artificial agents will reshape the *topology of moral experience* may at first seem tailored to embodied, physically present robots. But its significance extends directly to the contemporary landscape dominated by large language models. As the earlier discussion of LLMs and the "post–Machine Ethics" era makes clear, modern AI systems no longer resemble the rule-based architectures that shaped the first wave of Machine Ethics. They operate through statistical patterning, implicit social modelling, and affectively charged conversational exchanges. They recalibrate attention, shape expectations, influence interpretation, and modulate interpersonal stance.

In other words, even without bodies, **LLMs are already perturbation operators on the evaluative field**. What varies is the channel of perturbation. Robots perturb *perceptual* and *embodied* salience. LLMs perturb *semantic*, *discursive*, and *interpersonal* salience. Both influence the intuitive layer of moral cognition—the layer that precedes deliberation and shapes the evaluative landscape in which reasons and principles gain behavioural traction.

Seen from this perspective, the technomoral thesis is not limited to robotics. It is a general claim about how artificial systems—embodied or disembodied—reconfigure the cognitive–affective conditions under which human moral judgment unfolds. The role of this chapter is precisely to make this conceptual shift visible. Without a clear account of moral cognition as an *action-guiding*, *field-sensitive*, and *LoA-dependent* architecture, discussions about LLM "moral competence" or "machine virtue" become methodologically ungrounded.

Classical Machine Ethics imagined that the moral significance of AI lay in the principles encoded into the machine. The present analysis shows that the real significance lies in the *perturbations AI induces in us*.

Thus the technomoral thesis challenges Machine Ethics not because LLMs solve the old problems of rule-encoding, but because they demonstrate the irrelevance of those problems. If moral behaviour is shaped at the level of salience, affect, and social meaning, then the central question is no longer:

> *"Can a machine follow an ethical principle?"*

but rather:

> *"How does the machine's presence—physical, linguistic, or social—alter the evaluative field in which human agents form moral judgments?"*

The role of this chapter, therefore, is foundational. It provides the cognitive, psychological, and philosophical machinery required to see why this reframing is necessary. Without the distinctions introduced here—between descriptive and normative domains, factual and moral judgment, intuitive and deliberative processing, and above all, between Levels of Abstraction—one could easily mistake the current success of LLMs at producing coherent moral-sounding text for evidence of genuine moral cognition.

The chapter prevents this mistake. It equips the reader with the conceptual discipline needed to interpret both robotic and linguistic systems not as moral agents in any substantive sense, but as *environmental modifiers*: systems that reshape salience, meaning, and behaviour by transforming the evaluative topologies within which human moral cognition is enacted.

Thus the link back to the earlier discussion is straightforward: the technomoral thesis is the correct answer to the question of AI's moral significance in the LLM era—not because machines have become moral, but because our *moral environment* is being continuously reshaped by artificial systems whose influence operates beneath the threshold of reflective judgment.

4. Tools of Measurement, Framework and Experimental Design

4.1 Tools of Measurement

Empirical work aimed at understanding moral cognition must specify, with some philosophical care, the instruments through which psychological and behavioural structures become accessible to observation. Moral appraisal itself is never directly given; it does not present as a datum in the way a magnetic field or a spectral line may. *It is inferred from patterned responses*: affective, dispositional, perceptual, and social that reveal how evaluative information is encoded and transformed within the agent's cognitive architecture [9, 147, 10, 11, 12, 148].

The instruments employed in this thesis therefore function not as neutral measurement devices but as *theoretically motivated probes.* Each tool targets a specific dimension of the evaluative topology developed in earlier chapters, allowing latent dispositional structure to be rendered empirically tractable without collapsing its complexity into reductive summary scores. Their significance is thus analogous to that of the instruments of physics and the conceptual scaffolds of philosophy: they do not merely "record" a value but *constitute the mode of access* through which the phenomenon becomes observable at all.

In this respect, psychological instruments resemble the measuring practices of both laboratory physics and analytic philosophy. In physics, one does not detect an electron or a gravitational wave without the apparatus that makes such a phenomenon measurable; the device does not simply capture reality but partially *defines* the phenomenon by specifying the *level of abstraction* and the dimension of variation it reveals. Similarly, philosophical analysis specifies the conceptual lens through which reasoning, inference, or normativity become discernible. Measurement is not passive reception but disciplined *construction of access.*

The same principle governs the tools used here. Measures of systematising and empathising dispositions (EQ/SQ) do not claim to exhaust personality, but they isolate axes of cognitive–affective variation known to shape intuitive and deliberative routes in moral cognition [85, 149, 150]. The Big Five Inventory (BFI) captures broad dispositional gradients that interact with evaluative salience and behavioural inhibition [151, 152, 153]. These instruments are therefore not "psychological thermometers" but structured interventions into the *evaluative field*: each selects, with theoretical justification, the dimensions along which individual differences become experimentally meaningful.

This methodological stance also informs the design of the Watching-Eye paradigm used in the experimental study. The effect is not introduced as a folkloric behavioural curiosity but as a calibrated environmental perturbation:

> *a means of modulating accountability salience and affective vigilance at the cognitive level of abstraction identified earlier.*

The Watching-Eye cue is thus treated as a *contextual operator* on the evaluative topology, providing a controlled channel through which implicit social meaning acquires behavioural force. Its careful specification is essential, for—as in physics—the measurement depends not only on the quantity being observed but on the entire apparatus through which observation is made possible.

Throughout this chapter, the emphasis will therefore not be on the instruments as psychological artefacts, but on their *epistemic role* in the explanatory framework of the thesis: how each measure maps onto the evaluative architecture, what assumptions it encodes, and how it constrains the interpretation of the behavioural data that follows.

4.2 Measurement as Theoretical Access

The methodological commitments of this thesis require a principled account of the instruments through which evaluative behaviour becomes empirically accessible. Work in moral psychology and cognitive science has repeatedly shown that moral appraisal is not directly observable but manifests through structured patterns of affective response, controlled cognition, and social cue integration [9, 147, 10, 11, 12, 148]. For this reason, empirical studies of moral cognition depend on validated constructs and measurement strategies capable of rendering latent dispositions observable without distorting their theoretical significance.

The present work does not align itself with moral cognition research as a discrete disciplinary domain. Instead, it draws upon rigorously established constructs from moral psychology, cognitive science, and social signal processing as operational resources for making evaluative dispositions tractable. Instruments such as the Empathizing Quotient [7], the Systemizing Quotient [154], and the Big Five Inventory [155, 153] provide precisely the kind of psychometric access to stable individual differences that contemporary models of moral cognition identify as structurally relevant. Likewise, the analytical frameworks developed within Social Signal Processing [63] offer methodological grounding for understanding how agents register, interpret, and respond to contextually salient perturbations.

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

4.2.1 A Coherent Measurement Suite

The Empathizing Quotient (EQ), the Systemizing Quotient (SQ), and the Big Five Inventory (BFI) offer validated operationalisations of dispositional constructs repeatedly implicated in moral judgment and social decision-making. Likewise, the Watching–Eye paradigm [1, 2, 4, 6, 5] constitutes a mature experimental framework for probing reputational concern, prosocial motivation, and sensitivity to subtle social cues. Together, these instruments form a coherent measurement suite capable of isolating trait-level parameters that interact with contextual salience to shape moral behaviour.

If the theoretical chapters have argued that moral cognition is best understood as an evaluative topology shaped by attention, affect, social meaning, and trait-level curvature, then the tools introduced here function as the *coordinate system* through which that topology becomes visible. Their role is not to reduce personality to numbers, nor to treat empathy or systemizing as atomic psychological entities, but to expose the *invariant structures along which agents differ in how they process moral salience.*

4.2.2 Measurement in Experimental Context

In the experiment motivating this thesis, evaluative perturbation is elicited not through explicit moral dilemmas but through a more subtle and ecologically grounded manipulation: the silent perceptual presence of a humanoid robot. Prior work in human–robot interaction shows that even passively positioned robots can shift perceived social affordances, alter attentional allocation, and modulate expectations concerning norm-relevant behaviour [35, 156, 157]. Their ambiguous ontological status disrupts default social priors and thereby reconfigures the salience landscape within which moral reasons become behaviourally operative.

In this respect, robotic presence functions as a controlled perturbation of the evaluative topology itself, enabling the empirical study of how dispositional invariants interact with contextual cues to produce measurable differences in moral behaviour. The Watching–Eye cue provides a second calibrated perturbation; together, they specify the salience structure of the environment in which participants navigate moral action.

4.2.3 Purpose and Structure of this Chapter

The aim of the chapter is thus twofold.

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

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

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

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

IN OGNI ESPRESSIONE, BISOGNA DEFINIRE TUTTI I TERMINI

The instruments included here—the Empathizing Quotient (EQ), the Systemizing Quotient (SQ), the Big Five Inventory (BFI), and the Watching–Eye paradigm—were selected because they satisfy three stringent criteria grounded in established empirical research [149, 85, 150, 151, 152, 153, 1, 2, 5]:

1. **Theoretical relevance:** Each tool targets a component of moral topology (affective resonance, evaluative precision, personality curvature, or salience modulation) [18, 62, 6, 63].
2. **Empirical robustness:** Each tool is validated across multiple cultures, large samples, and decades of psychological research, and has been used in studies of prosociality, moral sensitivity, social attention, and Human–Robot Interaction (HRI) [158, 159, 160, 86, 1, 2, 3, 33, 34, 51].
3. **Computational suitability:** Each tool produces variables suitable for integration into regression models, cluster analysis, and topological interpretation [17, 161, 90].

Before turning to the tools themselves, we first articulate the methodological role they play within this thesis. Their significance lies not in psychometric convenience but in their ability to expose the latent structures through which evaluative salience is processed and transformed—structures that, as the experiment will show, can be subtly but measurably deformed by the presence of a synthetic agent.

At this point we can raise an important question:

> *How does the theoretical weight of the measurement framework reconcile with the fact that the experiment involves a relatively modest sample of seventy-one participants? And further, have these tools been employed in comparable studies with similar data constraints?*

The answer reveals something essential about the architecture of the project. The tools used in this thesis—EQ, SQ, the BFI, and the Watching–Eye paradigm—were not selected because they demand large samples for interpretability, but because they target *structurally stable* psychological constructs. These instruments are grounded in decades of psychometric work involving thousands of participants, and their factor structures, reliability profiles, and discriminant properties are well established across diverse populations [149, 85, 150, 151, 152, 153, 158, 159, 160]. In effect, they carry their statistical scaffolding with them. A study does not need to reproduce the entire validation literature; it inherits the stability of constructs that have already been exhaustively characterised.

This matters because the goal of the experiment is not to discover new personality factors nor to recover latent dimensions from scratch. It is to examine how *known dispositional invariants* interact with a controlled perturbation of the evaluative field. Small-to-moderate sample studies are the norm in this domain. Replications of the Watching–Eye effect routinely employ samples ranging from 40 to 120 participants [1, 2, 4, 5], and HRI studies investigating the behavioural impact of robotic presence commonly operate within similar ranges [35, 34, 51]. Even studies integrating personality measures into moral-decision paradigms—for example, the use of BFI or EQ/SQ to predict prosociality, empathic concern, or social attention—often rely on samples of comparable scale [71, 86, 157, ?].

BROKEN LINK

The statistical strategy deployed in this thesis reflects this precedent. Rather than fitting high-dimensional models or mining for latent structure, the analysis treats dispositional measures as low-dimensional, theoretically structured parameters influencing the deformation of evaluative gradients. The inferential load therefore falls not on discovering complex patterns in sparse data, but on detecting systematic, directional shifts in behaviour induced by the experimental manipulation. For this purpose, a well-powered design does not require a large sample; it requires a clean manipulation, validated constructs, and an analysis aligned with the theoretical architecture [17, 161, 90].

In short, the experiment does not attempt to estimate the topology of moral cognition from scratch. It examines how a synthetic agent perturbs an already well-understood structure. The sample size is calibrated not to psychometric exploration but to experimental contrast:

> *detecting whether robotic presence produces a measurable attenuation of prosocial action across dispositional profiles.*

Numerous studies across HRI, SSP, and moral psychology demonstrate that such effects are robustly detectable with sample sizes of the magnitude employed here [67, 63, 6, 51].

With this clarification in place, we now turn to the tools themselves. Their significance lies not in psychometric convenience, but in their ability to expose the latent structures through which evaluative salience is processed and transformed—structures that, as the experiment will show, can be subtly but measurably deformed by the presence of a synthetic agent.

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

Within the framework developed thus far, moral behaviour is modelled as the endpoint of a trajectory across an evaluative field. Contemporary work in moral psychology and cognitive science converges on the idea that such trajectories arise from the coordinated influence of three factors: environmental cues, dispositional structure, and perturbational forces [9, 147, 10, 11, 12, 148]. This is captured, in schematic form, by the functional decomposition

IL TERMINE EVIDENZIATO NON E' DEFINITO

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

where:

- α_E encodes the *environmental inputs*, such as the Watching–Eye prime and task context;
- β_C represents the *dispositional configuration* measured by psychometric instruments;

- γ_R denotes the *perturbation operator* introduced by the humanoid robot.

The psychometric tools employed in this study populate the β_C term. They render stable individual differences empirically visible by quantifying constructs known to influence how agents register and integrate affective, social, and contextual information. The Empathizing Quotient [7] indexes the *affective bandwidth* through which others become morally salient; the Systemizing Quotient [154] captures the *structural bias* shaping analytic interpretation; and the Big Five Inventory [155, 153] maps broad *personality curvature* influencing attention, norm-sensitivity, and regulatory control [162]. These variables do not exhaust the dispositional space, but they provide theoretically grounded coordinates on dimensions repeatedly implicated in moral appraisal and prosocial action [9, 12].

Their role in the analysis is therefore structural rather than decorative. Without them, dispositional heterogeneity would remain unmodelled, and any perturbation effect could be mistakenly attributed to uncontrolled trait variance. The cluster analysis of EQ, SQ, and BFI scores indeed revealed a non-trivial dispositional topology: affectively warm profiles, analytically structured profiles, and reactive–volatile profiles. Participants did not enter the experimental setting as psychologically interchangeable agents.

What matters is what came next. Despite this structured diversity, the humanoid robot produced a *uniform directional attenuation* of prosocial behaviour across all dispositional clusters. No Big Five trait, EQ dimension, SQ factor, or latent profile moderated the effect. This finding parallels results in human–robot interaction showing that even passive robots can globally modulate social affordances, attentional allocation, and normative expectations [35, 156, 157]. The present study extends that literature by demonstrating that robotic presence does not selectively amplify or suppress particular traits. Instead, it acts on the *evaluative field* itself: reshaping salience gradients, damping affective trajectories, and shifting the topology within which all trait-based pathways unfold.

The psychometric tools were indispensable for establishing this point. They allowed the analysis to dissociate the *shape of the dispositional manifold* from the *geometry of the perturbation.* In the experimental formalism, dispositional structure enters the perturbation comparison as

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

a difference that isolates the contribution of γ_R while holding β_C fixed. The empirical pattern in Chapter 5 showed that while β_C exhibits a structured internal topology, the perturbation generated by the robot overwhelms trait-specific differences and applies a global deformation to the evaluative landscape. This is precisely the signature of a field-level operator rather than a trait-contingent stimulus.

Thus, the purpose of this section is not simply to catalogue the tools, but to clarify their methodological necessity. They provide the coordinates required to demonstrate that the robot acted not on *who* participants were, but on the evaluative conditions under which their moral trajectories unfolded. Psychometrics makes visible the dispositional substrate; the experiment reveals the topological deformation imposed upon it.

With this distinction in place—between dispositional structure and field-level perturbation—we can now turn to a concise examination of the specific constructs measured by each instrument and how they map onto the evaluative–topological architecture of moral cognition.

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

Given the dual-layer structure revealed by the experiment—stable dispositional variation on one hand, and a field-level displacement induced by robotic presence on the other—the choice of psychometric and experimental instruments cannot be arbitrary. The tools selected here satisfy three methodological criteria that are essential for interpreting the attenuation of prosocial behaviour in a theoretically meaningful way.

(1) Cross-paradigmatic relevance. The EQ, SQ, BFI, and Watching–Eye paradigm each derive from long-standing empirical traditions spanning moral psychology, social cognition, personality research, and Human–Robot Interaction. Across these literatures, they have been used to study prosociality, empathic concern, harm aversion, cognitive style, and the integration of affective and deliberative processes in moral evaluation [9, 147, 10, 11, 12, 148].

The Big Five Inventory remains the canonical operationalisation of broad personality architecture with well-established predictive value for behavioural outcomes [155, 153, 162]. The Empathizing and Systemizing Quotients provide validated assessments of affective resonance and analytic curvature [7, 154]. Meanwhile, the Watching–Eye paradigm constitutes one of the most robust manipulations of prosocial salience, repeatedly demonstrating that minimal cues of observation modulate cooperative and charitable behaviour [1, 2, 4, 6, 5].

Taken together, these instruments align the present study with a broad empirical landscape while remaining faithful to the evaluative–topological framework established earlier.

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

- **EQ**: the affective attractors anchoring early moral appraisal [7];
- **SQ**: the curvature associated with analytic or rule-based processing [154];
- **BFI**: the personality geometry modulating salience, attention, and regulatory control [155, 153, 162];
- **Watching–Eye**: a validated perturbation of moral salience without instruction or coercion [1, 2, 4, 6, 5].

This heterogeneity of scope provides the granularity needed to model the dispositional term β_C in the formal expression

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

and to cleanly distinguish trait-level variation from field-level perturbation. This distinction is the key empirical insight: robotic presence acted on the evaluative field rather than on personality-dependent gradients.

(3) Stability and interpretability. The selected instruments satisfy three further requirements:

- **Stability**: each has robust psychometric validation across cultures and samples;
- **Analytical tractability**: each yields variables suitable for clustering, regression, and topological modelling;
- **Interpretability**: each connects to established moral-psychological and philosophical accounts, enabling behavioural findings to be integrated with theoretical models of moral appraisal.

Most importantly, these tools provided the precision required to demonstrate that the attenuation effect was not driven by personality configurations, empathizing profiles, or systemizing tendencies. The psychometric suite revealed a structured dispositional landscape, but the robot altered behaviour *irrespective* of that structure. The tools therefore allowed the experiment to distinguish *who the participants were* from the *geometry of the evaluative field* within which their choices were made.

With these criteria established, we now turn to the first measurement instrument: the Empathizing Quotient.

4.5 The Empathizing Quotient (EQ): Affective Resonance as Evaluative Curvature

The Empathizing Quotient (EQ) provides a validated measure of affective resonance—an individual's capacity to detect, register, and respond to the emotional and psychological states of others [7]. Originally developed within the Empathizing–Systemizing framework [163, 164], the EQ captures both emotional reactivity and cognitive perspective-taking, two mechanisms repeatedly shown to influence prosocial behaviour, harm aversion, and sensitivity to moral salience [9, 147, 11, 12].

Why EQ Matters Conceptually

Within the evaluative–topological model developed in this thesis, empathizing corresponds to the *affective curvature* of the evaluative field. High EQ scores indicate steep affective gradients: morally relevant others appear more salient, distress is more motivationally weighted, and the transition from appraisal to prosocial action becomes more strongly guided by affective dynamics. Low EQ profiles, by contrast, reflect flatter affective manifolds in which moral cues exert weaker pull.

In this sense, the EQ is not merely a trait measure; it provides a quantitative coordinate for the dispositional term β_C in the mapping

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

where β_C denotes the stable parameters shaping how evaluative information is transformed into behaviour.

Historical and Psychometric Grounding

Empirically, the EQ has a robust record: strong internal reliability, stable factor structure across cultures [158], convergence with related constructs (empathic concern, emotional intelligence), and predictive validity for prosociality in behavioural economic tasks. Neurocognitive studies further show correlations between EQ scores and activation in vmPFC, anterior insula, and TPJ—regions central to affective resonance and mental-state attribution [19, 18].

These features make the EQ particularly suitable for this thesis: it is theoretically interpretable, computationally tractable, and empirically grounded.

4.5.1 EQ and Synthetic Presence

The central scientific function of EQ in this experiment was to determine whether empathic sensitivity moderated the attenuation effect introduced by the humanoid robot. One plausible hypothesis, grounded in moral psychology and HRI, is that high-empathy individuals would exhibit stronger prosociality and possibly stronger perturbation under synthetic social cues [33, 35].

The data ruled this out. EQ did *not* moderate the displacement effect. High- and low-empathy participants alike showed reduced prosocial donation in the robot condition. This finding is theoretically decisive:

the robot altered the evaluative field itself, not the trait-dependent gradients within it.

This result aligns with evidence from HRI showing that robotic presence modulates attentional and social-evaluative processing independently of empathic predisposition [156, 157].

4.5.2 Methodological Role in the Thesis

The EQ served two indispensable methodological purposes:

1. **Controlling for affective heterogeneity**. Without a measure of empathic sensitivity, reductions in donation could have been attributed to unmeasured differences in participants' empathy levels. The EQ rules out this confound.

2. **Modelling the affective dimension of β_C**. EQ provides the affective coordinate of the dispositional manifold, enabling cluster analysis and regression models to distinguish dispositional shape from field-level perturbation.

Thus, even though affective resonance is central to moral cognition, the experiment shows that the perturbation introduced by the humanoid robot acted *upstream* of empathy—altering the evaluative topology rather than amplifying or suppressing empathic traits.

With the affective dimension of β_C established, we now turn to the analytical dimension: the Systemizing Quotient.

4.6 The Systemizing Quotient (SQ): Structural Precision in the Evaluative Field

Where the Empathizing Quotient (EQ) indexes affective resonance, the Systemizing Quotient (SQ) [165, 159, 158] quantifies a cognitive style characterised by rule extraction, structural analysis, and the search for causal regularities. Within the evaluative–topological model developed in this thesis, the SQ corresponds to the *analytical curvature* of the evaluative field: the extent to which agents encode situations via stable structural relations rather than affective gradients.

4.6.1 Theoretical Background and Psychometric Foundations

The SQ emerged from the Empathizing–Systemizing framework [163, 164], originally designed to capture the dissociability of affective versus rule-based processing in autism research. Subsequent work broadened this motivation: systemizing is now associated with mechanistic reasoning, pattern extraction, predictive modelling, and a preference for low-noise, high-coherence causal schemas [159]. Psychometric studies demonstrate high internal reliability, cross-cultural robustness, and predictable correlations with analytic problem-solving and rule-consistent behaviour.

Neurocognitively, higher SQ scores correlate with lateral prefrontal and parietal activation during analytic reasoning; they are also associated with reduced activation in affective salience networks during social tasks [8]. These findings support the interpretation of SQ as measuring a cognitive style that privileges structural stability over affective modulation.

4.6.2 SQ Across Moral Psychology and HRI

In moral psychology, systemizing predicts greater reliance on deliberative processing, reduced affective interference, and increased endorsement of principle-based judgments in high-conflict dilemmas [46, 11]. In behavioural economics, high-SQ individuals show more consistent strategic patterns and reduced susceptibility to framing effects.

In Human–Robot Interaction, systemizing tendencies shape expectations about synthetic agents: high-SQ participants tend to interpret robots through structural and functional cues rather than anthropomorphic ones and attribute competence and reliability more readily than emotional or social qualities [35, 156, 157]. This makes the SQ especially relevant in the present experiment, where the perturbation introduced by the robot is primarily structural rather than affective.

4.6.3 SQ in the Evaluative–Topological Framework

In the formalism of this thesis, SQ contributes to the dispositional term β_C in

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

Where EQ shapes the *steepness* of affective gradients, SQ shapes the *rigidity* and *smoothness* of evaluative trajectories. High systemizing corresponds to a more stable evaluative surface: situations are encoded through structural invariants, making low-level affective perturbations less influential.

This intuition can be expressed heuristically through curvature:

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

where larger values indicate more rigid evaluative structures.

4.6.4 SQ, Synthetic Presence, and Field-Level Perturbation

Despite these theoretical expectations, the experiment showed that SQ did *not* moderate the behavioural attenuation caused by the humanoid robot. High-SQ individuals—those most likely to rely on rule-based evaluation—displayed the same directional reduction in prosocial behaviour as high-empathy and low-empathy participants.

This finding is conceptually important. It demonstrates that robotic presence operated not on the cognitive style of participants but on the *evaluative field itself*. Systemizing tendencies did not buffer, amplify, or redirect the behavioural effect.

The perturbation introduced by the robot was global, not trait-specific.

This aligns with existing HRI work showing that ambiguous synthetic agents alter social affordances and attentional dynamics independently of analytic or empathic predispositions [35, 156].

4.6.5 Methodological Significance

SQ served two methodological functions within the experiment:

1. **Controlling for cognitive style**. Without an explicit measure of systemizing tendencies, attenuation could have been misattributed to analytic disposition rather than environmental perturbation.
2. **Modelling the structural dimension of β_C**. SQ provides the analytical coordinate within the dispositional manifold, enabling the analysis to distinguish dispositional geometry from field-level displacement.

Together with the EQ, the Systemizing Quotient ensures that dispositional structure is properly characterised before interpreting the behavioural impact of robotic presence. The next tool completes this picture: the Big Five Inventory, which captures broad personality geometry beyond empathy and systemizing.

4.7 The Big Five Inventory (BFI): Personality Geometry and Evaluative Topology

The Big Five Inventory (BFI) is one of the most robust instruments in differential psychology. It distils decades of lexical and psychometric research into five broad, cross-culturally stable dimensions—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [166, 167]. Within the evaluative–topological model introduced earlier, these traits provide a principled coordinate system for mapping the dispositional manifold (β_C): the stable personality geometry through which evaluative trajectories take shape.

4.7.1 Why Personality Matters for This Thesis

Personality traits function as attractors and modulators in behavioural space. They influence affective responsiveness, attentional allocation, social orientation, and regulatory stability—precisely the mechanisms identified in earlier chapters as constitutive of moral cognition. The BFI therefore allows us to characterise the dispositional background against which the perturbation introduced by the humanoid robot operates.

Crucially, the BFI offers the stability required for distinguishing dispositional structure from the field-level displacement effect observed in the experiment. Without a measure of trait geometry, we would lack the dimensional resolution necessary to determine whether the attenuation of prosocial behaviour reflected personality differences or a global perturbation of the evaluative field.

4.7.2 Psychometric Strength and Cross-Domain Predictive Value

The BFI is among the most widely validated trait measures in psychology. Its factor structure replicates across cultures; its items display strong internal reliability; and short forms such as the BFI-10 preserve psychometric clarity under experimental time constraints [153]. The Big Five dimensions predict a wide range of behavioural outcomes—social engagement, helping, rule adherence, and responsiveness to interpersonal cues [162, 168, 86]. These properties make the BFI an ideal tool for modelling β_C within a topological framework concerned with how agents integrate contextual and affective information into action.

4.7.3 Personality, Moral Behaviour, and Social Presence

Each Big Five trait has theoretical relevance for moral behaviour:

- **Agreeableness** steepens prosocial attractors and predicts cooperation, altruism, and sensitivity to interpersonal harm.
- **Conscientiousness** stabilises evaluative trajectories and supports rule-consistent behaviour.
- **Neuroticism** introduces volatility and heightens susceptibility to contextual variation.
- **Extraversion** amplifies responsiveness to social presence and perceived observation.

- **Openness** broadens contextual sampling and modulates tolerance for ambiguity.

In Human–Robot Interaction, these traits influence how artificial agents are perceived—whether as social entities, competent tools, or norm-relevant observers [35, 169]. The BFI therefore ensures that personality-driven interpretations of robotic presence can be empirically tested rather than assumed.

4.7.4 Personality Geometry in the Evaluative–Topological Model

Within the formalism

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

the BFI quantifies the geometry of β_C. Personality traits define the curvature, stability, and directionality of the evaluative field for each participant:

- Agreeableness deepens altruistic basins.
- Conscientiousness smooths and stabilises evaluative gradients.
- Neuroticism increases local fluctuations.
- Extraversion amplifies social input channels.
- Openness expands contextual sensitivity.

These geometric interpretations allow personality to be formally integrated into the evaluative architecture without reducing behaviour to trait-level dispositions.

4.7.5 Cluster Analysis: Making Personality Geometry Visible

BROKEN LINK

The cluster analysis (Chapter **??**) revealed three dispositional attractors:

1. **Prosocial–Empathic**: high Agreeableness and high EQ; strong affective attractors.
2. **Emotionally Reactive**: high Neuroticism; unstable gradients and high volatility.
3. **Analytical–Structured**: high Conscientiousness and high SQ; rigid evaluative curvature.

These clusters show that participants entered the experiment with *structured dispositional diversity*. Psychological homogeneity cannot be assumed; it had to be measured.

4.7.6 The Key Empirical Result: Uniform Displacement

Despite these pronounced dispositional differences, the experiment revealed a striking result:

The humanoid robot produced a uniform attenuation of prosocial behaviour across all clusters.

No Big Five trait—and no cluster—moderated the effect.

This finding is decisive. It demonstrates that the perturbation introduced by the robot acts at the *field level*. It reshapes the evaluative topology itself, not the trait-specific pathways that populate it. This aligns with work in HRI showing that robotic presence can shift perceived social affordances independently of personality [156, 157].

4.7.7 Methodological Significance

The BFI provides the evidential basis for distinguishing between:

- **dispositional geometry** (the shape of β_C), and
- **field-level deformation** induced by the robotic perturbation (γ_R).

Without the BFI, the attenuation could easily have been misinterpreted as a by-product of personality—differences in Agreeableness, Extraversion, or Neuroticism—rather than as a global shift in the evaluative field.

4.7.8 Point of the Situation: What the BFI Shows

At this stage in the book, the tools chapter reaches its central conclusion:

The personality manifold is structured, but robotic presence bends the evaluative field in a direction that does not depend on personality.

The BFI demonstrates three essential achievements:

1. It verifies that participants differ dispositionally in meaningful, theoretically interpretable ways.
2. It anchors the cluster analysis that reveals the architecture of β_C.
3. It proves that the behavioural attenuation is not trait-driven but topology-driven: a global deformation of evaluative structure induced by synthetic presence.

This completes the dispositional component of the evaluative–topological model and prepares the ground for the next chapter. Having established the geometry of β_C and ruled out trait-based explanations, we can now turn to the design of the experimental perturbation itself: the Watching–Eye paradigm and the silent humanoid robot that reconfigures the evaluative field.

4.8 The Watching–Eye Paradigm: Amplifying Moral Salience and Revealing Field-Level Deformation

Across behavioural ethics, social psychology, and field experiments on prosociality, one finding has proven remarkably robust: minimal cues of being observed—stylised eyes, schematic gaze, or even two black circles resembling pupils—reliably increase cooperation, charitable giving, and norm compliance [1, 2, 4, 6]. This "watching–eye effect" operates without instruction or coercion. It is a perturbation of the perceptual environment that increases the salience of norm-relevant behaviour.

Within the evaluative–topological framework developed earlier, watching-eye cues function as controlled amplifiers of moral salience: they steepen prosocial attractors in the evaluative field by increasing the perceived social meaning of one's actions. This makes them the ideal baseline against which to detect whether synthetic presence deforms evaluative trajectories.

4.8.1 Watching–Eye Cues as Topological Amplifiers

Classical interpretations framed the effect in terms of reputational vigilance: an implicit inference that one's behaviour is observable and potentially judged by others [1, 2]. More recent accounts show that the effect emerges from the coordinated modulation of:

- **attentional uptake** of norm-relevant cues,
- **affective arousal** associated with evaluation or self-conscious emotions,
- **interpretive expectations** shaped by implicit social monitoring systems.

Formally, watching-eye cues operate on the environmental input term by increasing prosocial weighting:

$$\alpha_E \mapsto \alpha_E + \delta\alpha_{\text{eye}}, \qquad \delta\alpha_{\text{eye}} > 0.$$

This steepens the initial gradients through which intuitive appraisals evolve, making cooperative trajectories more accessible in the evaluative field.

4.8.2 Why Child-Pair Eyes Provide a Clean Experimental Baseline

Child-eye posters are widely used in prosociality experiments because they combine perceptual sociality with minimal conceptual content. Decades of work demonstrate that stylised child eyes:

- robustly increase prosocial behaviour across cultures and settings [1, 2, 3, 72, 4];
- evoke empathic and care-based affective responses [170];
- amplify attentional vigilance without implying the presence of a moral agent [6].

This makes them ideal for experimental use. They provide a *high-salience but low-interpretation* cue: strong enough to elevate prosocial gradients, simple enough not to introduce confounds involving mind attribution or intentionality.

4.8.3 Why Synthetic Presence Dilutes or Distorts the Effect

The central theoretical claim of the thesis—that humanoid robots act as *perturbation operators* on the evaluative field—becomes particularly clear when considering their interaction with watching-eye cues.

Humanoid robots are perceptually social but ontologically indeterminate. They are seen, but not reliably understood, as bearers of evaluative or moral capacities [35, 156, 157]. This ambiguity weakens all three mechanisms that normally support the watching-eye effect:

1. **Reputational inference is unstable.** Robots rarely trigger the implicit assumption that one is being morally evaluated.
2. **Affective resonance is dampened.** Observation by a non-agentive entity does not engage self-conscious emotions strongly.
3. **Attentional cues conflict.** The perceptual system registers social presence; higher-order systems deny full agency.

The result is a fractured evaluative landscape: the cue "someone is watching" is present at the perceptual level, but stripped of the evaluative force that normally steepens prosocial attractors.

4.8.4 Empirical Finding: Uniform Attenuation of the Watching–Eye Effect

The experiment confirms this prediction:

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

Even participants with traits associated with high social sensitivity (Agreeableness, Extraversion, EQ) showed the same directional decrease in prosocial behaviour. Formally:

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

where $\Delta_{\mathscr{R}}$ is a field-level displacement induced by the robot. The absence of moderation by EQ, SQ, or any BFI trait demonstrates that this displacement operates independently of dispositional geometry.

4.8.5 Why the Watching–Eye Paradigm Is Indispensable

For the purposes of this thesis, the watching-eye paradigm serves four methodological functions:

- **It provides a reliable high-salience baseline** against which attenuation can be detected.
- **It links the experiment to established moral psychology**, enabling direct comparison with decades of prosociality research.
- **It isolates genuine perturbation effects**, since attenuation can only occur if salience is first elevated.
- **It reveals the topology of moral cognition**, showing how synthetic presence deforms evaluative gradients rather than simply reducing generosity.

Without this paradigm, the behavioural shift could not be interpreted as a deformation of the evaluative field.

4.8.6 Integration With Costly Prosocial Action

Donation tasks provide observable moral action rather than abstract moral judgment. Their integration with watching-eye cues allows the experiment to follow the evaluative trajectory from:

1. cue uptake, to
2. salience amplification, to
3. action selection.

The robot's attenuation of this sequence demonstrates that synthetic agents alter the mapping from perceptual cues to moral behaviour.

4.8.7 Synthesis: A Window Into Moral Topology

The watching-eye paradigm serves as a conceptual and methodological hinge in the experiment. By steepening prosocial gradients, it makes the evaluative field's structure visible. By attenuating these gradients, the humanoid robot reveals the central result of the thesis:

> **Synthetic presence acts on the evaluative field itself rather than on personality-dependent pathways.**

The watching-eye effect therefore provides the diagnostic contrast needed to show how robotic co-presence deforms the topology of moral cognition.

4.9 General Conclusion: Measurement as the Logic of Synthetic Moral Perturbation

This chapter has developed far more than a list of instruments. It has established the measurement logic of the entire thesis: the conceptual grammar through which synthetic presence becomes empirically legible. The Empathizing Quotient (EQ), Systemizing Quotient (SQ), Big Five Inventory (BFI), and the Watching–Eye paradigm form a unified system of epistemic probes. Each is theoretically grounded, psychologically validated, and methodologically indispensable for making the evaluative topology of moral cognition observable without reducing it to caricature.

The formal architecture introduced earlier models moral behaviour as the output of a mapping

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

where:

- α_E captures the structure of environmental moral cues;
- β_C denotes the dispositional manifold shaping evaluative uptake;
- γ_R represents the perturbational operator introduced by synthetic presence.

Each measurement tool corresponds to a distinct component of this model:

- **EQ** probes the affective attractors of β_C: the steepness, reach, and accessibility of prosocial gradients.

- **SQ** probes the structural curvature of β_C: the deliberative rigidity, rule-coherence, and model-based stability of evaluative trajectories.
- **BFI** provides the *coordinate system* of β_C: the multi-dimensional personality geometry needed to identify dispositional clusters and map their evaluative signatures.
- **Watching–Eye cues** perturb α_E: they steepen prosocial gradients and thereby create the diagnostic contrast necessary to observe displacement by γ_R.

Taken together, these tools do not simply "measure variables." They give the experiment an *evaluative topology*—a structured moral landscape within which deformation can be detected, described, and interpreted.

4.9.1 Dispositional Mapping: A Structured Manifold, Not a Confound

A major contribution of this chapter is the demonstration that dispositional diversity is structured, measurable, and separable from perturbational effects. The cluster analysis derived from EQ, SQ, and BFI revealed three dispositional attractor types—Affective–Prosocial, Emotionally Reactive, and Analytical–Structured—each characterised by distinct evaluative curvature.

Yet the experiment showed a striking and theoretically decisive pattern:

The humanoid robot attenuated prosocial action uniformly across all clusters.

This is a non-trivial finding. It rules out trait-level explanations—agreeableness, extraversion, empathy, systemizing, emotional volatility—as proximate drivers of the attenuation. The dispositional manifold β_C was not the site of modulation.

Instead, the perturbation operated at the level of the evaluative field itself.

Without the psychometric tools, this inference would have been impossible: the attenuation could have been misread as personality noise rather than as a genuine deformation of moral topology.

4.9.2 Watching–Eye Cues as Diagnostic Amplifiers

The Watching–Eye paradigm provided the complementary half of the measurement logic. By steepening prosocial attractors in α_E, it created the high-salience baseline against which synthetic attenuation became visible. The robot's presence did not merely reduce generosity—it *neutralised a well-established amplifier of moral salience.* This interaction is the clearest empirical signature of field-level perturbation.

In theoretical terms, the eyes amplified the gradient; the robot deformed the landscape.

Only the combination of psychometric mapping (of β_C) and salience amplification (of α_E) allowed this deformation to be isolated as an operation of γ_R.

4.9.3 Philosophical and Ethical Meaning

Placed in dialogue with the philosophical frameworks introduced earlier, the tools reveal the following:

- **Against rationalist models:** the perturbation bypasses deliberation.
- **Against virtue-theoretic accounts:** stable dispositions do not shield agents from synthetic deformation.
- **Against sentimentalist explanations alone:** the effect persists even in high-empathic profiles.
- **Against Machine Ethics assumptions:** moral significance lies not in the agent (robot) but in the environment the agent reshapes.

The instruments thus do double philosophical work: they expose the mechanisms through which moral action is generated, and they reveal the conceptual blind spots in contemporary ethical thinking about artificial agents.

Synthetic presence does not simply "influence" behaviour; it refracts the geometry through which moral meaning becomes action. It is neither a moral agent nor merely a tool. It is a *moral perturbator*: an entity capable of bending the evaluative field.

4.9.4 Methodological Synthesis: The Tools as Epistemic Infrastructure

This chapter has constructed the epistemic infrastructure required for the experiment. It has shown that:

1. moral behaviour can only be interpreted through a model that distinguishes environmental cues, dispositional structure, and perturbational operators;
2. psychometric tools provide the resolution needed to map β_C precisely enough to rule out trait-based explanations;
3. observational cues provide the experimental leverage needed to manipulate α_E in a controlled and theoretically meaningful manner;
4. synthetic presence must therefore be analysed as a deformation of *evaluative topology*, not as a stimulus acting upon isolated traits.

In this sense, the tools are not auxiliary components of the experiment—they are the *conditions of intelligibility* for its results.

4.9.5 Transition to the Experimental Methods

The next chapter operationalises everything established here. It translates the theoretical variables into stimuli, tasks, and statistical models. It describes how psychometric instruments were administered, how salience modulation was implemented, how synthetic presence was introduced, and how evaluative deformation was quantified.

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

The question that now motivates the remainder of the thesis is precise:

> *Does synthetic presence reshape the evaluative field through which moral salience becomes action?*

The methodological architecture developed in this chapter ensures that the experiment can answer this question with conceptual clarity, empirical rigor, and philosophical depth.

5. Experimental Method

SUONA PRETENZIOSO

5.1 From Conceptual Architecture to Empirical Test

The preceding chapters established a theoretical claim with both philosophical depth and empirical ambition: *that moral behaviour emerges from a topologically structured evaluative field, and that synthetic agents can perturb this field by altering the conditions under which moral salience becomes action.* The present chapter marks the transition from conceptual architecture to empirical adjudication. Here, every assumption must be operationalised, every construct measured, and every inference anchored in explicit experimental procedure.

This section begins with the precise research question that animates the experiment:

> **Question 5.1:** *Inferential Displacement*
>
> Does the silent presence of a humanoid robot—perceptually social yet ontologically indeterminate—alter the evaluative process that transforms moral perception into prosocial behaviour?

This question is not a rhetorical prompt but a methodological commitment. It situates the experiment within the evaluative–topological model developed earlier, in which moral action is expressed as:

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

COSA E' δ_m? E $\mathscr{P}(\delta_m)$? LE DEFINIZIONI ANDREBBERO DATE DALLA PRIMA VOLTA CHE LI USI NEI CAPITOLI PRECEDENTI

where α_E denotes morally salient environmental cues, β_C the dispositional manifold quantified through psychometric tools, and γ_R the perturbation operator instantiated by the robot's presence. The purpose of the experiment is to determine whether γ_R induces a measurable deformation in the mapping from α_E to observable moral action.

5.1.1 Why the Question Matters

Although behaviourally simple, the question reaches beyond classical experimental paradigms in moral psychology. It does not ask whether robots communicate norms, nor whether they persuade or instruct. It asks whether synthetic *presence* alone—minimal, silent, behaviourally neutral—modifies the inferential pathway through which moral salience produces action. Within the broader research programme of social signal processing and moral AI, this constitutes a stringent and foundational test:

> *can an artificial agent operate as a perturbation operator on the moral field even in the absence of agency?*

Embedding a humanoid robot into a moral environment thus serves as a direct empirical probe of the thesis developed in earlier chapters: namely, that moral cognition is not solely a matter of internal reasoning or personality structure, but a dynamic, context-sensitive transformation governed by the topology of situational cues.

5.1.2 Operationalising Moral Action: Prosocial Donation as Behavioural Endpoint

To render this transformation empirically measurable, the experiment operationalises moral action through a cost-bearing behavioural choice: voluntary donation of a portion of the participant's monetary compensation to a children's medical charity. This measure, extensively validated in behavioural ethics and moral psychology, captures the endpoint of the evaluative trajectory: the point at which moral salience is either converted into action or allowed to dissipate.

Costly charitable donation has been repeatedly validated across moral psychology, behavioural economics, evolutionary anthropology, and developmental science as the most reliable behavioural proxy for prosocial moral action [171, 172, 173, 174, 175]. It satisfies the three criteria required at the Level of Abstraction adopted in this thesis: it is elicited by morally salient cues [1, 2, 71]; it incurs real cost [175, 176]; and it expresses the action-guiding force of moral evaluation [17, 46, 18]. Its long-standing use as the behavioural termination point of moral cognition [171, 172, 174] justifies its role here as the measurable endpoint of the evaluative trajectory under synthetic perturbation.

The independent variable is equally minimal: the presence or absence of a humanoid robot autonomously animating in *life-mode*: NAO does not speak, instruct, or engage. Its movements are restricted to micro-gestures—simulated breathing, subtle postural adjustments, and gaze-orienting behaviours triggered only by human eye contact. These micro-movements, while non-agentic, replicate the perceptual features known to activate the Watching-Eye effect, thereby introducing a controlled form of synthetic social salience into the evaluative environment.

5.1.3 Why a Humanoid Robot?

The choice of a humanoid robot reflects a deliberate methodological position. As established in Chapter 3, synthetic agents occupy an unstable location in our social ontology: they possess perceptual salience and humanoid morphology, yet lack the moral-evaluative capacities ordinarily ascribed to observers. This combination creates the precise form of perturbation the experiment seeks to test: a perceptibly social presence whose normative meaning is ambiguous.

The question is therefore not whether participants think the robot is judging them; rather, it is whether the robot's presence alters the field of salience within which morally relevant cues exert their behavioural pull.

Question 5.2: *Inferential Displacement*

Can the mere presence of a synthetic observer—lacking agency, intention, and moral standing—perturb the inferential transformation that converts morally salient cues into prosocial action?

5.1.4 From Question to Design: Why We Do Not Begin with a Hypothesis

Framing the study around a research question rather than a directional hypothesis is intentional. In interdisciplinary work spanning philosophy, psychology, neuroscience, and HRI, a premature hypothesis risks narrowing the interpretive field and smuggling in unexamined assumptions about how synthetic presence ought to behave. The methods must therefore preserve epistemic openness: *the design must reveal whether perturbation occurs, not assume that it does.*

This methodological humility is continuous with the philosophical commitments articulated earlier. If moral behaviour arises from a dynamic integration of environmental cues, dispositional structure, and social presence, then the experiment must be sensitive to field-level deformations that cannot be anticipated a priori.

5.1.5 The Logic of the Experimental Test

In practical terms, the experiment leverages the integrated measurement framework developed in the previous chapter. The Watching-Eye paradigm constructs a baseline of elevated prosocial salience (α_E). The EQ, SQ, and BFI quantify the structure of the dispositional manifold (β_C). The robot enacts the perturbation operator (γ_R). The donation task measures the resulting behavioural transformation $\mathscr{P}(\delta_m)$.

The empirical question is therefore precise:

Question 5.3: *Empirical Question*

Does γ_R—the silent, perceptually social presence of a humanoid robot—systematically deform the evaluative mapping from α_E to $\mathscr{P}(\delta_m)$ across the dispositional manifold β_C?

E' UN TERMINE
MOLTO FORTE

If the answer is affirmative, the findings reveal a foundational claim: that artificial agents, even when behaviourally minimal, exert moral influence not by persuasion or instruction but by reshaping the topological conditions under which moral salience becomes action.

What follows in this chapter details the machinery by which this question is tested: the design logic, the structure of the experimental task, the observational conditions, the psychometric integration, and the analytic strategies used to detect perturbation.

> *The conceptual framework provided the variables. The empirical design now tests their transformation.*

5.2 Experimental Design and Behavioural Paradigm

To address Question 5.1.3 (p. 68), we implemented a controlled behavioural experiment [177, 178, 179] grounded in the *Watching–Eye* paradigm [65, 2, 180, 181, 182, 183, 184]. The scientific objective was not simply to measure donation behaviour, but to determine whether γ_R, the silent perceptual presence of a humanoid robot, systematically deforms the evaluative mapping

$$\alpha_E \longmapsto \mathscr{P}(\delta_m)$$

across the dispositional manifold β_C.

Participants were recruited individually into a lab environment under the pretext of participating in a personality study in exchange for monetary compensation. This cover task served two methodological purposes. First, it provided a psychologically neutral framing for the room-based task. Second, it elicited trait-level measurements (EQ, SQ, BFI) necessary for modelling the dispositional manifold (β_C) within the evaluative–topological framework.

Embedded within this setting was a morally salient cue: a prominently placed charity poster depicting a child in medical need. Decades of evidence (see Chapter 4) demonstrate that such stimuli reliably activate prosocial dispositions through implicit monitoring, empathic resonance, and affiliative concern [1, 6]. Within the formalism introduced in earlier chapters, this stimulus increases α_E, the environmental salience input, steepening the prosocial attractor in the evaluative field.

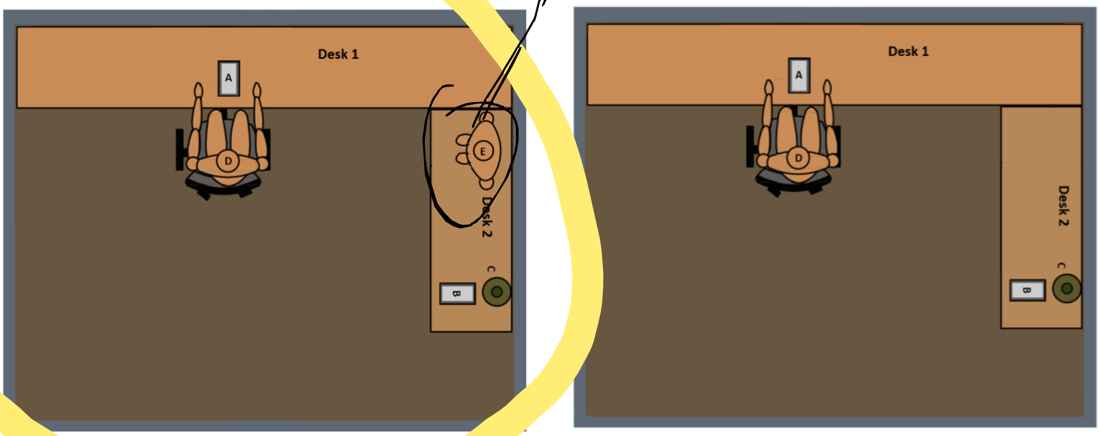
5.2.1 Experimental Manipulation: Presence as the Only Ontological Difference

Participants were randomly assigned to one of two conditions.

1. **Control Condition**: the participant completed the questionnaires alone in the room.
2. **Robot Condition**: a humanoid NAO robot [185] was present, operating in *autonomous life mode*.

NAO emitted no speech and performed no task-directed actions. Its behaviour consisted solely of the minimal embodied micro-cues characteristic of this mode: simulated breathing, subtle shifts of posture, and head-orientation responses triggered only by direct eye contact. These are precisely the class of low-dimensional social cues shown to activate or modulate the Watching–Eye effect: movement, gaze potentiality, and the perceptual suggestion of observation.

Si fatica a vederlo



(a) Experimental condition: robot present.

(b) Control condition: robot absent.

Figure 5.1: Top–down view of the experimental and control configurations. Both layouts are spatially and visually identical; the humanoid robot is the only ontological difference between conditions. In the evaluative–topological framework developed in this thesis, this equivalence is essential: the geometry of the environment (desk positions, donation box placement, participant orientation) is held constant so that any change in prosocial behaviour can be attributed to a deformation of the evaluative field induced by synthetic presence. Formally, the figure depicts two instantiations of the same environmental input α_E, differing only by the activation of the perturbation operator γ_R. The robot's placement maps onto a local modification of the salience landscape—an additional source of perceived observation—while the control condition represents the unperturbed topology.

Crucially, both experimental rooms were geometrically and visually identical (Fig. 5.1). The *only* manipulated variable was the presence or absence of the humanoid robot. Spatial layout, lighting, informational content, and the moral cue (α_E) were held constant.

In this design, the robot does not "do" anything in a behavioural sense. Instead, its minimal perceptual affordances present the participant with an ontologically ambiguous entity—perceptually social, morally inert, and semantically potent. The manipulation therefore isolates *presence as such* as the epistemic and experimental variable.

5.2.2 Why Minimal Presence Matters: Ontological Ambiguity as Cognitive Perturbation

The overwhelming majority of HRI and HMI studies assume that moral modulation arises through interaction: overt communication, feedback, adaptive behaviour, or explicitly framed expectations [34, 186, 187, 188, 189]. The present design rejects this assumption deliberately.

Rather than investigating how robots *act*, we investigate how they *appear*—how their mere existence within a perceptual field alters the evaluative pathway from moral salience to moral action. The experimental focus is therefore on **pre-reflective permeability**: the extent to which minimal agent-like cues reshape inferential structure prior to conscious deliberation [190, 191, 192, 193].

This approach isolates a structural vulnerability of norm-sensitive cognition: humans routinely over-ascribe agency in contexts of uncertainty [194, 195, 196]. By placing NAO precisely at the boundary between objecthood and agenthood, the design probes whether anticipation—not interaction—is sufficient to distort the evaluative topology.

5.2.3 Levels of Abstraction: Why the Robot Can Matter Without Doing Anything

Floridi's Levels of Abstraction (LoA) [25, 197, 198] provide the formal justification for treating NAO's silent presence as epistemically potent.

At the operative LoA of the participant, what is visible are *informational affordances*: posture, eyes, symmetry, subtle biological motion, the inert promise of mutual gaze [199, 200, 201, 202, 203, 204, 205, 206]. These cues are sufficient to trigger the primitives of social monitoring, even when the entity producing them is known to be non-human.

Thus, at this LoA, NAO functions as a *semantic perturbator*: not a moral agent, nor a communicative partner, but an informational presence that reshapes the participant's evaluative background conditions. If the robot were interactive, the LoA would shift (introducing agency, reciprocity, intentional stance). If the robot were inert, the social affordance would vanish. Autonomous life mode occupies the narrow space between these extremes.

This design choice aligns with Floridi and Sanders' analysis of artefactual moral agency [145]. Their 2004 account does not attribute consciousness, intentionality, or moral reasoning to artificial systems. Rather, it identifies moral relevance at the *Level of Abstraction* at which an artefact can contribute causal or informational influence within a given environment [25, 26]. At this LoA, an artefact may count as a "moral agent" in the minimal and operational sense that its presence supplies, modifies, or filters morally relevant information.

This perspective is directly compatible with contemporary discussions of large language models (LLMs), which similarly operate as *artefactual sources of semantic perturbation* rather than as bearers of intrinsic moral status [56, 55]. In both cases—the embodied robot tested here and the disembodied LLM—moral relevance arises not from interior capacities but from how the system reshapes the informational and social conditions under which human agents form evaluations and make decisions. Related arguments in HRI emphasise that robots exert moral and social influence through their perceived agency, morphology, and communicative affordances, not through any intrinsic mental properties [34, 35, 169].

For this reason, Floridi's account is particularly well suited to the present experimental context: it licenses the treatment of NAO's minimal, non-interactive presence as an epistemically potent variable without implying any claim about the robot's inner ontology. At the LoA operative for the participant, the robot is a *semantic perturbator*: a structured informational presence capable of altering the evaluative field through which moral salience becomes behaviourally operative. This conceptual continuity also clarifies why the findings developed in this thesis generalise to other classes of artificial systems—including LLM-based agents—whose moral significance likewise depends on the informational roles they play rather than on their metaphysical constitution [139, 207].

5.2.4 Behavioural Paradigm: Donation as Moral Action

After completing the questionnaires, each participant received £10 in £1 coins and encountered a voluntary donation option: a charity box positioned near the exit. They could donate any subset of their compensation. The amount donated served as the behavioural measure of prosocial action.

This operationalisation follows a long-established tradition in moral psychology, moral economics, and behavioural ethics in which cost-bearing prosocial behaviour tracks the practical expression of moral salience [208, 172, 209, 77, 210, 211, 62, 37, 212]. As demonstrated in Chapter 4, donation behaviour reliably expresses the terminal point of a moral evaluative trajectory.

5.2.5 Preliminary Findings

Initial analyses revealed a robust and theoretically coherent effect: participants in the Robot condition donated *significantly less* than those in the Control condition. Personality data (EQ, SQ, BFI) showed no meaningful differences between conditions, ruling out dispositional confounds and providing strong initial support for a field-level perturbation induced by synthetic presence.

These results motivate the next step: formalising the evaluative structure through which this behavioural displacement must be interpreted.

5.2.6 From Behavioural Setup to Evaluative Structure

The experimental setup provides the behavioural substrate. What remains is to specify the evaluative architecture through which any behavioural modification must be interpreted. In moral philosophy, action is often treated as the terminus of deliberation [13, 79, 15]. Yet the present study does not investigate deliberation itself. It examines the *transformation* that precedes deliberation's endpoint: the cognitive–affective process by which morally salient cues become behaviourally operative [83, 80].

Donation, within this design, is therefore not an isolated act but the *observable boundary condition* of an evaluative process. The Watching–Eye stimulus renders moral salience explicit; the robotic manipulation introduces a synthetic perturbation; the donation behaviour provides the measurable output of the transformation. This ensures that what is being tested is not trait-level generosity, but the *susceptibility of moral appraisal to synthetic co-presence*.

Classic variants of the Watching–Eye paradigm rely on pictorial cues or supernatural primes [2, 213]. The present experiment instead embeds an embodied but minimally active humanoid robot. This shift is critical: it replaces a two-dimensional prime with a three-dimensional presence whose *perceived ontology* is neither inert object nor full social agent. This ambiguity is precisely the condition under which moral salience may be refracted or displaced.

To formalise what the experiment tests, we treat moral action as the output of an evaluative function integrating environmental cues, dispositional structure, and perturbational affordances:

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

where:

- Σ is the morality-salient perceptual field (the Watching–Eye cue),
- $\mathscr{R}$ is the synthetic co-presence,
- f is the evaluative transformation linking perception to action,
- $\mathbb{E}[f(\cdot)]$ denotes the expected behavioural output.

Read informally: *the expected moral behaviour differs when the robot is added to the perceptual–moral environment.* This yields our first empirical hypothesis:

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.

To clarify the structure of this transformation, we decompose the probability of a deviation in moral action into its constituent determinants:

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

where:

- α_E represents the environmental moral cue (Watching–Eye),
- β_C encodes the dispositional structure measured by EQ, SQ, and BFI,
- γ_R denotes the perturbational effect of robotic co-presence.

In plain language: *the probability of observing a change in moral behaviour depends jointly on the moral cue, the agent's dispositional profile, and the presence of the robot.* This is the operative logic of the experimental design: the robot is not treated as a moral agent, but as a *topological perturbation*—a factor that reshapes the evaluative field within which moral cues are processed.

Why should a robot be capable of such perturbation? The answer lies in the notion of *moral salience*. Across cognitive science and moral philosophy, moral salience refers to the way certain features of the environment become normatively charged prior to explicit reasoning [80, 83, 31, 108]. It is a pre-reflective gatekeeper: what is foregrounded, what stands out, and what demands attention.

A synthetic presence may influence this salience not by speaking or acting, but by altering the perceptual and inferential background against which moral cues are interpreted. NAO's form, gaze orientation, and subtle embodied motions evoke the minimal conditions associated with social monitoring. They place the participant in a borderline space between being **alone** and being **observed**. This ontological ambiguity—central to human–robot interaction research—is precisely what makes NAO a semantically potent perturbation of 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 the behavioural prediction into the normative domain: the robot may change not only what people *do*, but the conditions under which moral meaning becomes actionable. The Watching–Eye paradigm thus becomes a conceptual probe—a way of examining the *structural elasticity* of norm-sensitive cognition in the presence of synthetic observers.

Under this interpretation, generosity is not a simple expression of stable virtue or personality; it is the *emergent property* of a cognitive–affective system embedded in a structured moral environment. Robotic presence, by virtue of its ontological ambiguity, acts as a refractive affordance: it bends the path from moral perception to moral action, attenuating the behavioural expression of prosocial salience.

This notion of an *emergent property* deserves clarification, for it plays an important explanatory role in how the experiment should be interpreted. In the present context, emergence does not denote mysterious or irreducible behaviour; it describes a structural fact about norm-sensitive cognition [31, 16, 17]. Prosocial donation arises here not from any single component of the experimental system—neither from the moral cue alone (α_E), nor from the dispositional architecture (β_C), nor from the robot's presence (γ_R) taken in isolation. Rather, generosity appears as the *behavioural output of an interaction* [18, 6]:

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

Under fixed dispositions, the output can change simply because the evaluative field in which those dispositions operate has been deformed [67, 63]. This is precisely what the data reveal: the robot produces a *uniform directional shift* in donation behaviour despite stable trait profiles [51, 33, 34]. In this sense, prosocial behaviour is emergent: it is a property of the *system* formed by dispositions, environmental cues, and contextual topology, not a direct expression of any one part [173, 174].

A helpful comparison can be drawn—carefully—with contemporary discussions of emergent capacities in large language models. In LLMs, emergence refers to capabilities that arise from the interaction of many parameters without being explicitly encoded [56, 55]. Here, too, the behavioural effect reflects an interactional architecture: moral action is generated by the coupling of perceptual salience, affective readiness, and contextual priors [46, 62]. Yet, unlike in LLMs, the emergence observed here is phenomenological and contextually scaffolded: the evaluative field itself is altered, and behaviour shifts even though the underlying dispositions remain constant [35, 169].

This also clarifies the function of the mathematical formalism introduced in the preceding chapters. The equations do not quantify moral agency in any metaphysical sense; they provide an explicit epistemic schema for locating the point of deformation [161, 90]. By decomposing the evaluative transformation into α_E, β_C, and γ_R, the formalism makes it possible to rule out trait-level explanations and demonstrate that the behavioural shift originates at the level of the mapping:

$$f(\alpha_E, \beta_C, \gamma_R) \neq f(\alpha_E, \beta_C).$$

In this respect, the mathematics functions as a conceptual microscope: it enables the isolation of the structural point at which synthetic presence exerts influence. Without such decomposition, the uniform attenuation might be mistakenly attributed to personality differences, random noise, or implicit experimental demand [5, 71].

Thus, when the analysis later reports that moral behaviour changed while traits did not, the claim is not that generosity "collapsed", nor that personality "failed" to predict behaviour. The claim is that the *evaluative topology* was reconfigured by an ontologically ambiguous presence—yielding an emergent behavioural pattern that no component of the system could produce alone [1, 2]. This, in turn, is what lends the experiment its broader philosophical significance: it demonstrates that synthetic agents can perturb the moral field not by thinking, or by acting, but simply by *being present* within the perceptual architecture through which moral salience becomes action [35, 34, 51].

With this evaluative architecture established, the next section examines how this deformation manifests empirically—first in behavioural data, and then in its interaction (or lack thereof) with dispositional structure. The question (5.1.5), as framed at the outset, demanded a yes/no answer. The analysis to follow now supplies the evidential basis for that answer.

5.3 Synthetic Perturbation of Moral Inference

Before entering the empirical phase, we require a precise *mechanistic anchor*:[1] a statement that links the evaluative–topological model developed in the preceding

[1] In this context, "mechanistic" refers not to physical causation but to the minimally specified, testable account of *where* in the evaluative process a perturbation is expected to act. It identifies the locus of influence within the mapping from moral salience to behavioural output.

chapters to the behavioural analyses that follow. Without such an anchor, the experiment would risk degenerating into a mere behavioural vignette, detached from the normative and computational structure established earlier. The present section therefore identifies the precise inferential target against which all subsequent statistical results must be interpreted.

Chapters 3–4 articulated the evaluative architecture through which moral salience (α_E) is transformed into behavioural output ($\mathscr{P}(\delta_m)$), modulated by dispositional structure (β_C) and,potentially, by synthetic perturbation (γ_R). The central empirical question (Question 5.1.3) asked whether the mere presence of a humanoid robot systematically deforms the mapping from salience to action. The role of the present hypothesis is to turn that question into a testable inferential claim: it specifies *how* and *where* the perturbation is expected to manifest within the evaluative transformation.

In the experimental setting, the Watching–Eye stimulus structures the moral field Σ; the dispositional manifold β_C, measured through EQ, SQ, and the BFI, provides each participant's cognitive–affective baseline; and the robot's presence $\mathscr{R}$ introduces a perceptually social, ontologically ambiguous affordance. The crucial question is whether $\mathscr{R}$ modulates the internal transformation that links perceptual–affective inputs to prosocial action.

$$\Sigma \longrightarrow \mathscr{D}$$

Under ordinary conditions, this transition is driven by the salience of the moral cue. When the robot is present, however, its ambiguous social ontology may refract or suppress the affective and reputational components that ordinarily support prosocial decision-making. This motivates the mechanistic hypothesis.

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 and reputational cues that ordinarily support donation, thereby modulating the evaluative pathway by which moral stimuli gain behavioural expression.

This hypothesis identifies the mechanistic level at which synthetic presence is expected to operate. The claim is not that NAO exerts coercive influence or that participants attribute moral authority to it. Rather, the prediction is that NAO's perceptually social yet ontologically indeterminate presence alters the *topology* of the evaluative field: shifting which features are foregrounded, how moral cues are weighted, and how affective resonance is integrated into action. In this sense, the robot functions as a *semantic perturbation*—a presence that reconfigures the informational structure through which salience becomes behaviour.

With this mechanistic hypothesis established, we can now transition to the empirical analysis. The next section evaluates whether the two experimental groups were equivalent in their demographic and dispositional structure, ensuring that any subsequent behavioural divergence can be attributed to the perturbative role of $\mathscr{R}$ rather than to background variation within β_C. The behavioural results that follow then provide the evidential basis for adjudicating whether the deformation predicted here is indeed observed.

5.4 Inferential Analysis of Experimental Data

Before we enter the empirical phase of the argument, it is crucial to clarify what this section contributes to the architecture of the thesis. Up to this point, the chapter has operated at the level of *evaluative structure*: we have identified the components of the perceptual–moral field, specified the variables of the evaluative transformation

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

and articulated the mechanistic hypothesis governing how synthetic presence (γ_R) may refract the mapping from moral salience (α_E) to behavioural output ($\mathscr{P}(\delta_m)$).

What follows changes register. This section inaugurates the *inferential* phase of the thesis: the point at which conceptual commitments must submit to statistical adjudication. If the preceding sections drew the topology of the evaluative field, the analyses to follow measure its curvature.

Two principles govern the transition.

1. **Inferential validity requires structural symmetry.** Before testing whether the perturbation γ_R deformed the evaluative mapping, we must first establish that the two experimental groups were equivalent with respect to demographic and dispositional structure. Without this symmetry, any behavioural divergence would be uninterpretable at the level of mechanism.

2. **Statistical analysis is not a post-hoc addition, but the operational expression of the theoretical model.** The inferential pipeline—from distributional checks to regression modelling and cluster analysis—implements the evaluative framework developed in earlier chapters. Each statistical test corresponds to a theoretical question: Does γ_R shift the distribution of prosocial action? Does β_C moderate that shift? Is the effect uniform across the dispositional manifold?

The present section therefore serves a dual purpose. First, it validates the experimental precondition of group comparability. Second, it establishes the methodological pathway through which the mechanistic hypothesis introduced above will be empirically evaluated.

> *From this point onward, every claim is grounded not in conceptual plausibility, but in statistical evidence.*

We begin by demonstrating the demographic and dispositional equivalence of the two participant groups. Only once this foundational condition is met can we proceed to analyse whether synthetic presence introduced a systematic deformation in the evaluative mapping from moral salience to observable moral action.

5.4.1 Demographic Equivalence as a Symmetry Condition

Before any inferential claims can be drawn from the behavioural data, we must establish that the two experimental groups were demographically comparable. Within the evaluative–topological framework developed earlier, demographic symmetry functions as a foundational *inferential constraint*: only when the underlying populations exhibit similar baseline characteristics can any observed behavioural divergence be attributed—within the limits of the design—to the perturbative presence of the robot $\mathscr{R}$ rather than to sampling asymmetries in the human substrate.

To this end, we examined three demographic variables that plausibly influence prosocial responsiveness in field and laboratory studies: gender, age, and educational background. Each was tested across the **Control** and **Robot** conditions using standard inferential procedures, with Benjamini–Hochberg False Discovery Rate (FDR) correction applied to guard against spurious equivalence due to multiple comparisons.

SE SI TRATTA DEL p-value, LO SI METTE SOLO QUANDO E' MENO DI 0.05.

STATISTICALLY

- **Gender distribution**: a chi-squared test revealed no significant difference between conditions ($p = 1.00$, FDR-corrected).

STATISTICALLY SIGNIFICANT

- **Age**: an independent-samples t-test detected no difference in mean age between groups ($p = 1.00$, FDR-corrected).

STATISTICALLY

- **Educational background**: a chi-squared test again showed no ~~reliable~~ difference ($p = 1.00$, FDR-corrected).

SIGNIFICANT

The convergence of these results under strict FDR control allows us to draw the following methodological conclusion:

> **The two experimental groups are demographically equivalent.**

This symmetry condition is essential for the analyses that follow. It ensures that the behavioural differences later observed cannot be attributed to demographic imbalance or hidden stratifications in the participant pool. Instead, under the architecture developed in the preceding sections, any systematic divergence in prosocial behaviour becomes attributable to the semiotic and perceptual perturbation introduced by the robot, $\mathscr{R}$, after holding α_E constant and before considering variation in the dispositional manifold β_C.

FDR NON CAMBIA IL VALORE DEL p-VALUE. CAMBIA QUELLO DEL SIGNIFICANCE LEVEL

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. Values shown include original and FDR-corrected p-values for gender, age, and educational background. No comparison reached significance after correction, supporting the assumption of demographic equivalence required for subsequent inferential interpretation of behavioural effects.

SONO GIA' TUTTI SOPRA SIGNIFICANCE LEVEL

NON SERVE FDR

With demographic symmetry established, the analysis proceeds to the next inferential layer: the behavioural effects of synthetic presence. Subsequent sections will assess donation outcomes directly, and only *thereafter* will the dispositional structure—encoded via EQ, SQ, and BFI—be examined for potential interactions with $\mathscr{R}$. This ordering preserves the logic of the evaluative–topological framework: baseline equivalence first, behavioural effects second, dispositional modulation third.

5.4.2 Data Preparation and Preprocessing Workflow

FORSE MI E' SFUGGITO, MA NON DICI DA NESSUNA PARTE QUANTI PARTECIPANTI HAI

Because the inferential analyses that follow rely on contrasts across experimental conditions, dispositional variables, and behavioural outputs, a principled preprocessing pipeline is an epistemic prerequisite rather than a technical convenience. The aim of this stage is to ensure that the dataset constitutes a coherent and interpretable representation of the experimental structure, free from syntactic artefacts, coding inconsistencies, or latent category imbalances. Only under such conditions can the subsequent statistical models be taken to track the evaluative transformations at issue in this chapter.

The dataset comprises demographic descriptors, psychometric measures (EQ, SQ, BFI), and the behavioural outcome (donation magnitude). These variables differ in type, scale, and inferential role; they therefore require tailored preprocessing steps to preserve their semantic integrity.

Standardisation of variable names. All variable names were converted to lowercase, whitespace-trimmed, and harmonised to eliminate discrepancies introduced through manual data entry. This ensures referential consistency throughout the analysis.

Encoding of behavioural outcome. The binary variable `donated_anything` was created (1 = donated at least one coin; 0 = donated nothing). This permits modelling of prosocial behaviour at two complementary levels: (i) the full distribution of donation amounts and (ii) the threshold decision to donate at all.

Encoding of experimental condition. The variable `condition_bin` was constructed (0 = Control, 1 = Robot) to allow direct incorporation into regression frameworks and to maintain a clear contrast between conditions.

Verification of categorical coherence. Categorical fields (e.g., `gender`) were inspected for irregularities such as collapsed, duplicated, or misspelled levels. No anomalies requiring recoding were identified.

Preliminary distributional checks. Initial visual inspections (histograms, density plots, boxplots) revealed no anomalous values requiring removal or recoding. Age distributions and donation distributions are shown in Figures 5.2 and 5.3, respectively, illustrating the distributional structures to be analysed in the inferential sections that follow.

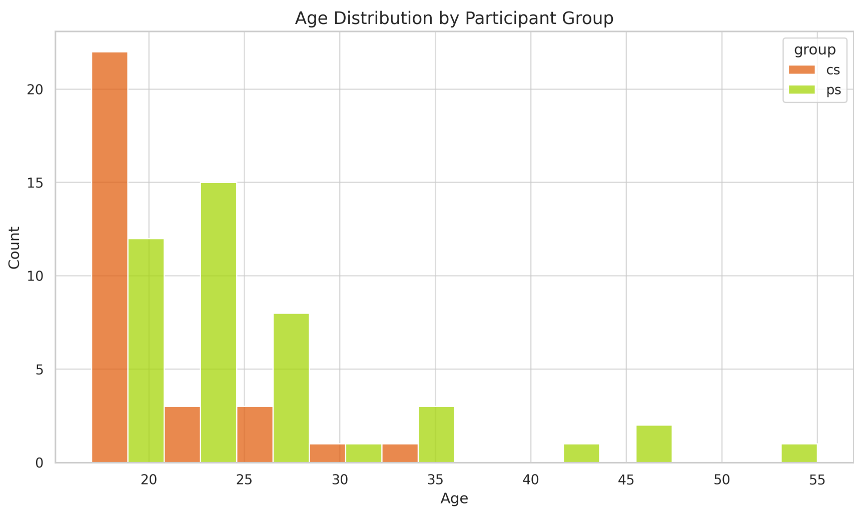


Figure 5.2: Age distribution across experimental conditions. The histograms illustrate the demographic structure of the sample to be examined in later analyses.

Taken together, these preprocessing steps establish the analytic coherence required for valid inferential modelling. With demographic equivalence confirmed in the previous subsection and the present transformations ensuring structural stability of the data, the chapter now proceeds to the statistical models that evaluate whether the perturbation introduced by $\mathscr{R}$ manifests in the transition from moral salience to moral action.

5.4.3 Preliminary Descriptive Patterns: Orientation Prior to Inferential Analysis

Before entering the inferential phase, it is useful to outline the basic distributional structure of the key behavioural and psychometric variables. Descriptive statistics possess no evidential force in themselves; their role is purely orientational. They summarise the raw landscape of the data so that the formal tests in the following sections can be interpreted against a clear empirical backdrop.

Table 5.2 reports the central tendencies for the principal variables collected in the study. The mean donation values in the two conditions (*Control* and *Robot*)

COSA SIGNIFICA LA DONAZIONE NEGATIVA?

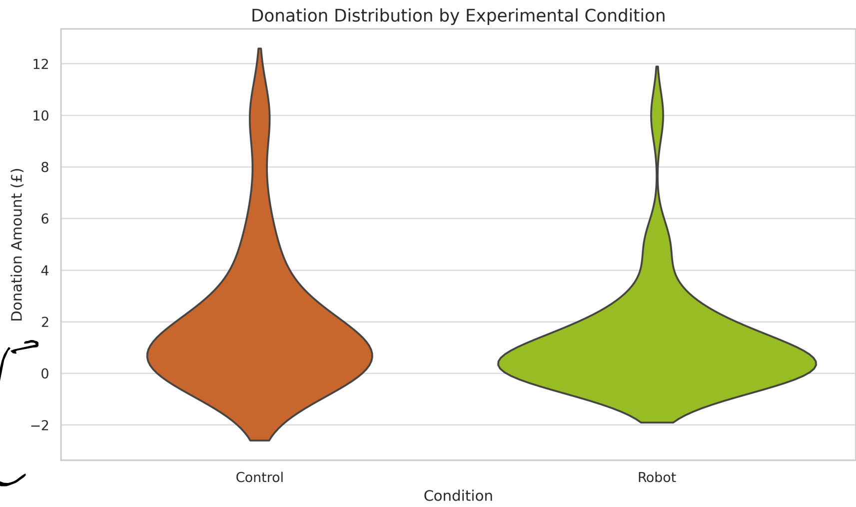


Figure 5.3: Distribution of donation behaviour by condition. The plot presents the behavioural data whose inferential assessment constitutes the next stage of analysis.

appear numerically distinct, and several psychometric scores (EQ, SQ, BFI subscales) exhibit small numerical differences across groups. These contrasts, however, are *purely descriptive*: they record observed sample characteristics and do not imply either imbalance or effect. Their interpretive significance, if any, will be assessed formally in the subsequent statistical analyses.

The descriptive summaries therefore serve three limited but important functions:

1. they present the distributional contours that later inferential tests will interrogate;
2. they facilitate visual inspection for anomalous values, without indicating any need for exclusion;
3. they prepare the reader for the dispositional and behavioural modelling developed in the sections that follow.

Crucially, nothing in the descriptive patterns licences an inferential conclusion. Whether the robotic presence $\mathscr{R}$ perturbs the evaluative mapping from moral salience to action is a question answered only by the formal models presented later. These descriptive tables merely contextualise the data that will feed into those models.

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 5.2: Descriptive summaries of behavioural and psychometric variables across experimental conditions. These values provide an orienting overview of the sample; they do not support any inferential claims regarding group differences or perturbation effects.

With this preliminary orientation in place, we now turn to the inferential structure itself, beginning with the verification of demographic symmetry across conditions—a prerequisite for attributing any subsequent behavioural divergence to the experimental manipulation rather than to background variability.

5.4.4 Inferential Comparison of Donation Patterns Across Conditions

Having established the demographic symmetry of the sample and the analytic coherence of the dataset, we now turn to the first formal evaluation of whether robotic presence $\mathscr{R}$ influences prosocial behaviour. Up to this point, all analyses have been structural or descriptive; the task now is to determine whether the behavioural distributions associated with the moral decision—the donation act—show any statistically reliable divergence across conditions. This is the first moment in the chapter where inferential weight is brought to bear on the *Evaluative Deformation Hypothesis* (Hypothesis 1), and the transition therefore marks a shift from conceptual scaffolding to statistical adjudication.

We proceed in an intentionally layered way. A single test rarely captures the complexity of a behavioural distribution; instead, a sequence of complementary analyses is required. We begin with a chi-squared test on coin-frequency distributions, then examine the full donation distributions using a Mann–Whitney U test, and finally quantify the magnitude of the difference via a nonparametric bootstrap. Each method probes a different facet of the data: aggregate totals, distributional structure, and effect-size stability respectively.

Chi-squared test on donation frequencies. A chi-squared test comparing the *frequency distribution of donated coins* across the Control and Robot conditions revealed a statistically detectable divergence:

$$\chi^2 = 4.25, \qquad p = .039.$$

This test does not assess means or medians but evaluates whether the overall pattern of coin contributions differs across conditions. The result indicates that

> *The aggregate structure of donation behaviour is not evenly distributed across the two environments.*

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 5.3: Inferential comparisons of donation behaviour across conditions. The chi-squared test compares coin-frequency distributions, while the Mann–Whitney U test and bootstrapped mean difference assess distributional structure and effect magnitude respectively.

It is crucial, however, that this result be interpreted with methodological restraint. The chi-squared test establishes an *aggregate* divergence, not a uniform shift in individual tendencies. To understand the behavioural topology more fully, we must examine the entire donation distribution.

Mann–Whitney U test on donation distributions. A Mann–Whitney U test, applied to the full distribution of donation amounts, did not detect a statistically reliable difference:

$$U = 777.0, \qquad p = .194.$$

This indicates substantial overlap between individual donation behaviours in the two conditions. In other words, while coin-frequency totals diverge, the fine-grained distribution of donation amounts remains broadly similar. This pattern suggests that the influence of $\mathscr{R}$ may be probabilistic and heterogeneous rather than deterministic or uniform across participants.

Bootstrapped estimate of the mean difference. To complement these analyses, we calculated a nonparametric bootstrap estimate of the mean donation difference:

$$\Delta M = 0.71, \qquad 95\%\ \mathrm{CI} = [-0.33, 1.79].$$

The estimate aligns directionally with the group-level pattern (Control > Robot), but the confidence interval includes zero, reinforcing the conclusion that the effect is subtle and probabilistic rather than sharply bifurcated.

Taken together, these three analyses trace a coherent statistical profile. There is evidence of divergence in aggregate coin-frequency behaviour, yet the distributional overlap and wide bootstrap interval demonstrate that the perturbation introduced by $\mathscr{R}$ is not uniform across individuals. This is exactly the pattern predicted by the evaluative-topological model:

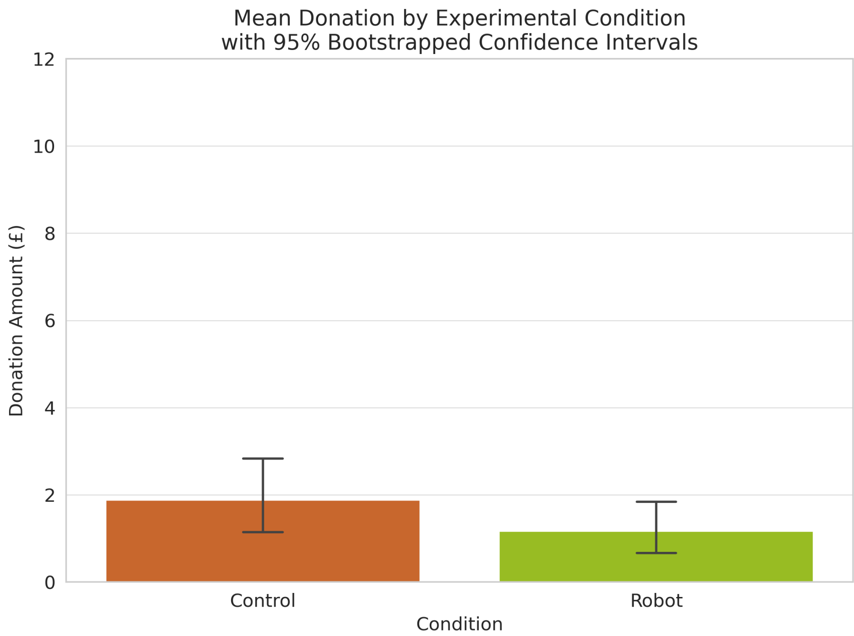


Figure 5.4: Mean donation amounts by experimental condition, with 95% bootstrapped confidence intervals. The overlapping intervals illustrate substantial individual-level variability, indicating that any perturbative influence of $\mathscr{R}$ is diffuse rather than deterministic.

> *Synthetic presence functions not as a coercive force but as a semiotic perturbator, modulating the evaluative mapping from moral cue to action in a heterogeneous population.*

In interpretive terms, the results neither dismiss nor overstate the effect. They demonstrate that robotic presence is behaviourally consequential at the aggregate level while leaving open, and thereby motivating, the central analytical task of the next subsections: determining whether this perturbation interacts with the dispositional manifold β_C. The inferential focus therefore now turns to regression modelling, interaction tests, and Bayesian estimation, where the structure of β_C can be incorporated directly into the evaluation of γ_R.

5.4.5 Interim Conclusion to Question 5.1.3

Partial Conclusion to Question 5.1.3

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

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

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

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

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

5.4.6 Quantification of Behavioural Modulation: Parametric and Nonparametric Effect Sizes

Having established that the experimental groups are demographically symmetric and that the aggregate-level analyses reveal a measurable difference in charitable behaviour across conditions, we now turn to the question of *magnitude*. Significance tests indicate whether a behavioural contrast is detectable relative to sampling variability; they do not characterise the structural amplitude of the perturbation induced by the synthetic co-presence $\mathscr{R}$. For this reason, the present section complements the inferential tests with parametric and nonparametric effect-size metrics, thereby quantifying the extent to which robotic presence modulates prosocial behaviour under the Watching–Eye paradigm.

Because the subsequent regression and interaction analyses will examine the interplay between robotic presence and dispositional structure, it is essential to begin with a transparent description of the overall behavioural landscape. The effect sizes presented here serve as the bridge between aggregate-level contrasts and the more nuanced trait–context models developed later in the chapter.

Effect-Size Framework

Two complementary measures were selected:

- **Cohen's** d: a parametric index of standardised mean difference, sensitive to shifts in central tendency;

- **Cliff's** Δ: a nonparametric ordinal effect size that estimates the probability that a randomly selected individual from one condition donates more (or less) than a randomly selected individual from the other.

Taken together, these metrics evaluate whether the presence of $\mathscr{R}$ reshapes the evaluative output distribution in a manner consistent with the deformation posited in the Evaluative Deformation Hypothesis (Hypothesis 1).

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 = corresponding standard deviations,
- n_1, n_2 = sample sizes.

Cliff's Delta.

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

Where:

- $\#(x > y)$ counts all cases where a Control donation exceeds a Robot donation,
- $\#(x < y)$ counts the inverse,
- n_x, n_y are the sample sizes of each group.

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. Their relevance lies not in magnitude alone, but in the fact that both metrics converge on the same directional pattern: robotic presence is associated with lower prosocial donation on average.

To ensure interpretive clarity, two complementary visualisations are provided. The kernel density estimate (Fig. 5.5) depicts the *shape* and spread of donation distributions, enabling inspection of distributional tails and modes. The mean-with-standard-error plot (Fig. 5.6) focuses on *central tendency* and sampling variability. Although partially overlapping in content, the two figures serve distinct analytic functions and together offer a transparent view of the behavioural landscape that informs the subsequent modelling work.

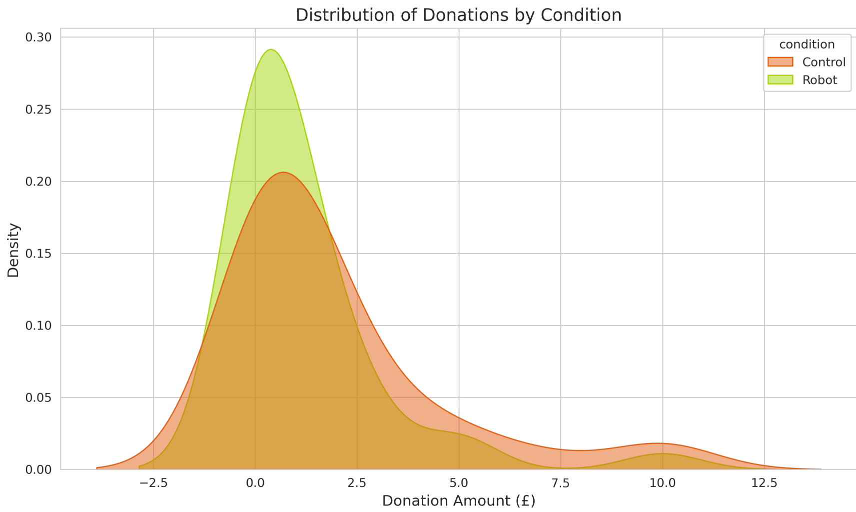


Figure 5.5: Kernel density estimates of donation distributions across experimental conditions. The Control group exhibits greater mass at higher donation values, whereas the Robot group shows a mild left-shift in density. These plots provide distributional context for the effect-size metrics discussed in the text.

For completeness, the inferential tests introduced earlier are reproduced in Table 5.4 alongside the effect-size metrics, ensuring that all aggregate-level results appear within a single consolidated reference point before turning to trait–context modelling.

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 5.4: Inferential comparisons of donation behaviour across conditions. The chi-squared test (applied to total coin frequencies), the Mann–Whitney U test, and the bootstrapped mean difference collectively characterise the behavioural contrast.

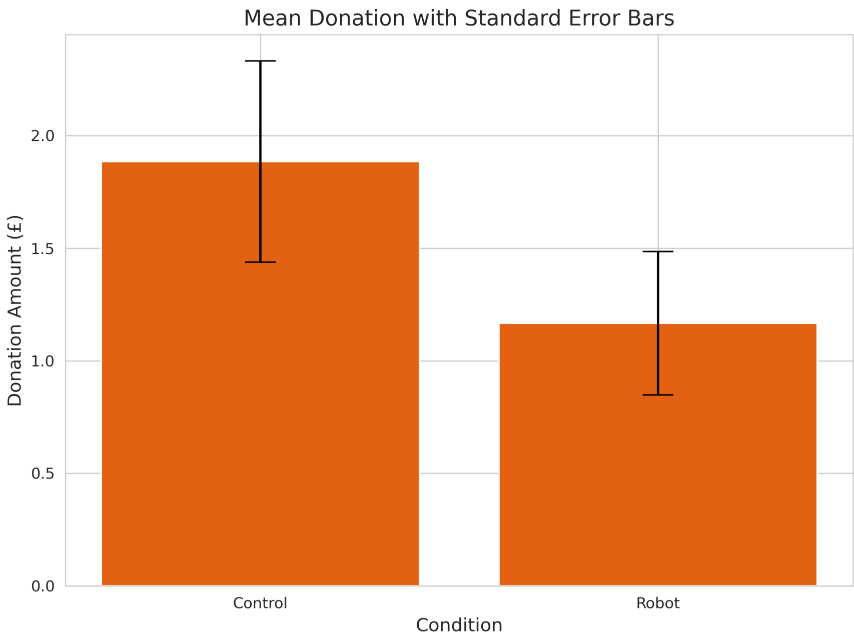


Figure 5.6: Mean donation amounts with standard error bars by condition. While the Control group donates more on average, the overlapping error bars reflect substantial individual-level variability. The figure complements the density plot by highlighting differences in central tendency rather than distributional shape.

Overall, the effect sizes indicate that robotic presence exerts a *directionally consistent, behaviourally modest* modulation of prosocial action. These outcomes are consistent with—though they do not in isolation confirm—the prediction that $\mathscr{R}$ perturbs the evaluative transformation from moral salience to behaviour. Importantly, the effect is *graded*, not binary: the evaluative system remains operative, but the strength with which moral cues are translated into action is probabilistically dampened.

Conclusion: Amplitude of Moral Refraction

Synthetic co-presence does not function as a binary suppressor of moral behaviour. Instead, it modulates the amplitude of the evaluative transformation from moral salience to action, introducing a subtle, probabilistic refractive shift consistent with its ambiguous ontological encoding at the operative Level of Abstraction.

At this point, the analysis shifts from evaluating the *main effect* of $\mathscr{R}$ to examining its interaction with individual-level dispositions. The dispositional manifold β_C—comprising empathizing and systemizing tendencies as well as Big Five personality traits—may modulate the susceptibility of the evaluative mapping $f(\alpha_E, \beta_C, \gamma_R)$ to perturbation. The following sections introduce regression models, interaction tests, and Bayesian estimation procedures designed to determine whether the attenuation observed here is uniform across the population or concentrated within specific psychological profiles.

5.5 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?*

5.5.1 Between-Condition Comparisons of Big Five Personality Traits

The effect-size analyses above established that robotic co-presence ($\mathscr{R}$) exerts a modest but directionally consistent modulation of donation behaviour. Before examining whether this perturbation interacts with individual differences, we must ensure that the two experimental groups were not already differentiated at the level of personality. If the Control and Robot conditions differed systematically in the Big Five traits, any apparent behavioural attenuation could reflect pre-existing dispositional imbalance rather than the influence of $\mathscr{R}$.

To assess this, we compared Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism across conditions using the Mann–Whitney U test. This test is appropriate for the structure of the dataset: the Big Five scores are bounded, ordinal psychometric variables exhibiting mild skew, and the sample size ($N \approx 70$) does not justify strong parametric assumptions. Because examining five traits entails five simultaneous hypothesis tests, the Benjamini–Hochberg False Discovery Rate (FDR) correction was applied to control Type I error.

After FDR correction, **none of the Big Five traits differ significantly** between the Control and Robot groups. Small numerical tendencies (e.g., slightly higher Openness in the Control condition) fail to approach corrected significance thresholds, and all distributions display substantial overlap.

TROPPO PICCOLI. PUOI FARE I PLOT PIU' GRANDI

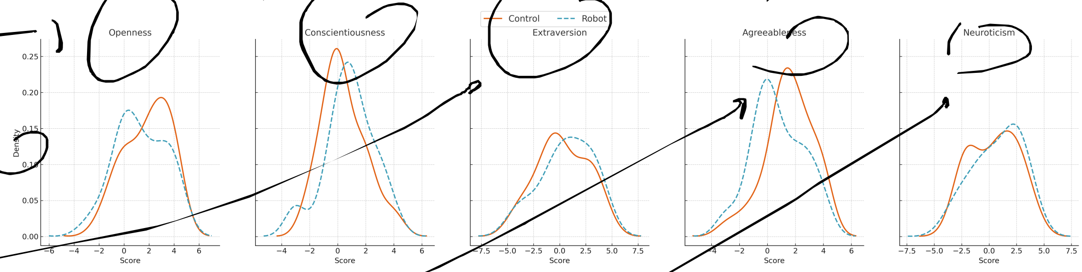


Figure 5.7: Kernel density estimates for each Big Five trait across conditions. All five distributions show substantial overlap, visually corroborating the non-significant Mann–Whitney tests.

This supports a key methodological inference:

The two experimental groups can be treated as dispositionally equivalent.

Accordingly, the behavioural difference observed earlier is *most plausibly attributed* to the presence of $\mathscr{R}$ rather than to pre-existing personality differences.

Having established dispositional equivalence, we now examine whether personality nonetheless predicts prosocial behaviour or interacts with the attenuation associated with robotic co-presence.

5.5.2 Predictive and Moderating Roles of Big Five Personality Traits

This analysis addresses a further theoretical question of interest:

> *Even if the groups are balanced, do the Big Five traits predict donation, or modulate the displacement induced by $\mathscr{R}$?*

(1) Predictive effects. Spearman rank correlations were computed between each Big Five score and donation amount. Spearman's ρ is appropriate for zero-inflated, bounded, and non-normal behavioural data, and for ordinal psychometric measures. Scatterplots with monotonic trend lines were examined to detect any nonlinear patterns not captured numerically.

(2) Moderation effects. To test whether personality modulates the displacement effect, interaction models of the form

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

were estimated for each Big Five dimension. This correctly operationalises the possibility that robotic presence acts as a *moral refractor*, exerting differential influence depending on dispositional architecture.

Methodologically, the findings are straightforward. **No statistically reliable associations** between any Big Five trait and donation amount appear in this dataset, and **no interaction** with experimental condition reaches significance.

The behavioural attenuation associated with $\mathscr{R}$ therefore shows no detectable variation across Big Five personality profiles.

TROPPO PICCOLI. PUOI FARE I PLOT PIÙ GRANDI

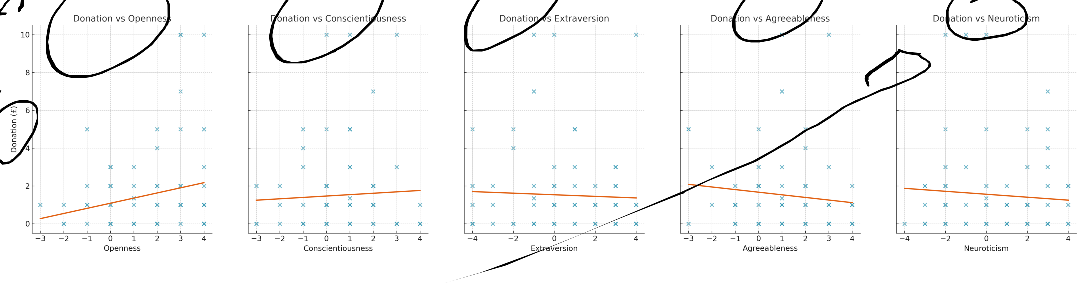


Figure 5.8: Scatter plots with monotonic trend lines for each Big Five trait against donation amount. No predictive relationships appear, and no moderation patterns are visible. This matches the null results from correlations and interaction models.

In summary, within the Big Five framework:

- no trait reliably predicts prosocial donation;
- no trait moderates the attenuation introduced by robotic co-presence;
- the displacement effect of $\mathscr{R}$ shows no detectable variation across Big Five profiles *in this sample*.

Conclusion: Trait-Independence of Evaluative Displacement

The attenuation of prosocial donation under robotic co-presence shows no detectable modulation by Big Five personality traits in this dataset. This supports the interpretation that $\mathscr{R}$ acts on the evaluative field itself rather than on specific dispositional pathways.

The next subsection examines whether more specialised social-cognitive traits—the Empathizing Quotient (EQ) and Systemizing Quotient (SQ)—show predictive or moderating roles that the broad Big Five taxonomy does not capture.

5.5.3 Transition to Structural Modelling of Dispositional Architecture

The analyses reported above establish two methodological foundations that shape the remainder of the statistical pipeline. First, the Big Five traits do not differ across conditions after False Discovery Rate correction, confirming dispositional symmetry between the Control and Robot groups. Second, within this sample, none of the Big Five traits reliably predict donation behaviour, nor do they interact with experimental condition. In inferential terms, the dataset contains no evidence of trait imbalance and no statistically detectable trait–by–condition moderation at the level of the classical personality taxonomy.

These observations do not *rule out* the relevance of dispositional structure; instead, they clarify the level of representation at which such structure should be modelled. The Big Five are coarse-grained scalar descriptors and may not capture the finer relational geometries (i.e., covariation patterns and trait interdependencies) that shape evaluative processing. Accordingly, the next stage of analysis adopts a more structurally sensitive approach to the dispositional manifold β_C, examining whether latent configurations of empathizing, systemizing, and Big Five attributes jointly organise susceptibility to robotic perturbation.

In this sense, the null findings within the Big Five framework serve a methodological rather than interpretive purpose. They demonstrate that any systematic modulation of donation behaviour by the synthetic presence $\mathscr{R}$ cannot be attributed to imbalances or linear trait effects within the classical personality model. This provides the inferential basis required to proceed to clustering and latent-structure modelling, where β_C is treated not as a vector of independent traits, but as a structured configuration whose internal organisation may interact with the perturbative affordances of $\mathscr{R}$.

The subsequent section therefore introduces the clustering methodology used to derive latent dispositional ecologies and examines whether these ecologies exhibit differential susceptibility to synthetic co-presence. This marks the transition from trait-level analysis to structural modelling within the broader evaluation of Question 5.1.3.

5.5.4 Latent Dispositional Structures and the Modulation of Moral Perturbation

The analyses thus far establish two essential points. First, the presence of the robot $\mathscr{R}$ is associated with a modest but coherent attenuation of prosocial donation at the aggregate level. Second, this attenuation cannot be attributed to differences in any individual Big Five trait. These results motivate a sharper question:

> *If conventional trait magnitudes do not explain variability in responsiveness to $\mathscr{R}$, might the perturbation be differentially expressed across* latent *cognitive–affective configurations within the dispositional manifold β_C?*

This question is structurally aligned with the evaluative model developed earlier. If synthetic presence perturbs moral behaviour by refracting the evaluative transformation $f(\alpha_E, \beta_C, \gamma_R)$, its influence need not be uniform across all individuals: the effect may depend on how dispositions combine into higher-order regimes rather than on isolated trait scores. To investigate this, we turn from scalar traits to a structural modelling of β_C.

Clustering the Dispositional Manifold

Seven psychometric variables—Empathizing, Systemizing, and the five Big Five traits—were used to construct the dispositional space. Each score vector was z-standardised, and dimensionality was reduced using Principal Component Analysis (PCA). Two orthogonal components were retained as they captured the dominant axes of variance while reducing redundancy among correlated traits.

The reduced two-dimensional representation served as input for k-means clustering. The choice of $k = 3$ rested on both methodological and conceptual grounds:

- The within-cluster sum of squares displayed a clear elbow at $k = 3$, indicating diminishing returns for higher k.
- Although a silhouette maximum was observed at $k = 9$, such peaks often reflect over-partitioning when N is modest; those solutions were therefore rejected.
- A three-cluster solution produced groups of interpretable size with stable internal variability, consistent with the expectation that a small number of dispositional regimes may structure evaluative responsiveness.

Figure 5.9 visualises the resulting partitions. The figure is retained because it provides essential structural justification for treating the clusters as psychologically interpretable configurations.

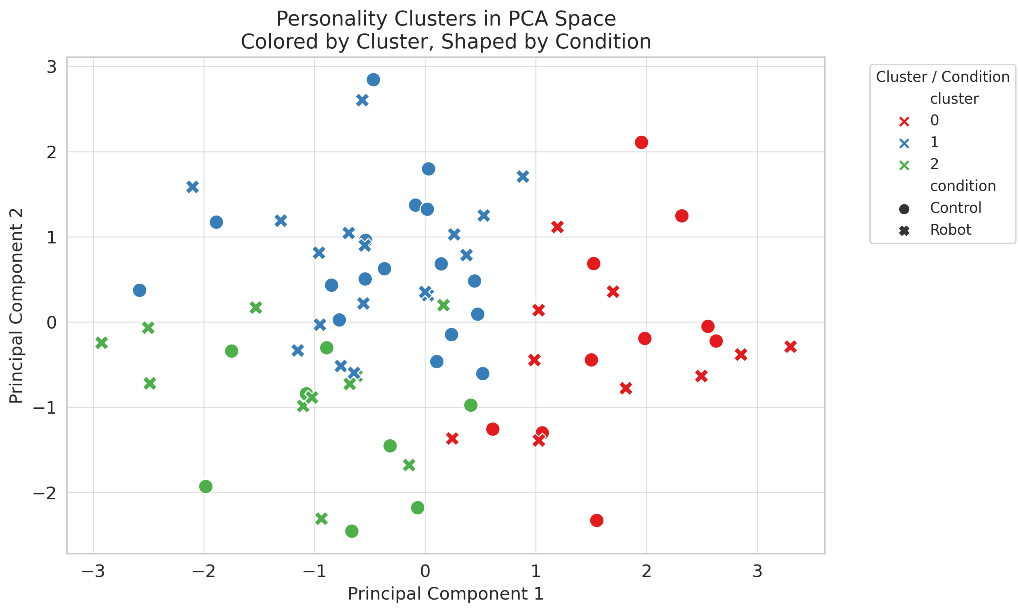


Figure 5.9: Participants clustered in PCA-reduced psychometric space. Three clusters emerge as coherent and visually distinguishable groupings, providing a structural basis for subsequent analyses of condition-by-cluster effects.

Justification of $k = 3$: Diagnostic Criteria

Figure 5.10 displays both the elbow curve and silhouette profile. It is included here because these diagnostics are standard tools for validating clustering solutions and demonstrate that $k = 3$ is a parsimonious and defensible choice.

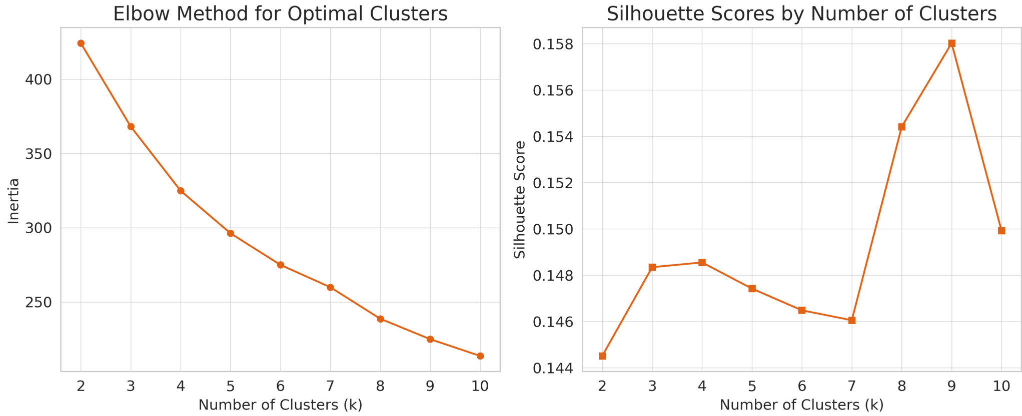


Figure 5.10: Elbow plot (left axis) and silhouette coefficients (right axis) across candidate values of k. The elbow at $k = 3$ and stable silhouette profile support selecting three clusters as an interpretable and parsimonious solution.

Conceptually, a small number of clusters is consistent with the idea that only a limited set of dominant dispositional regimes may modulate how moral salience is processed under synthetic perturbation.

Cluster-Specific Patterns of Moral Response

We then examined whether the donation attenuation associated with $\mathscr{R}$ differed across clusters. Figure 5.11 shows mean donation by condition within each cluster. This visualisation is essential because it provides the descriptive foundation for the interaction models to be developed next.

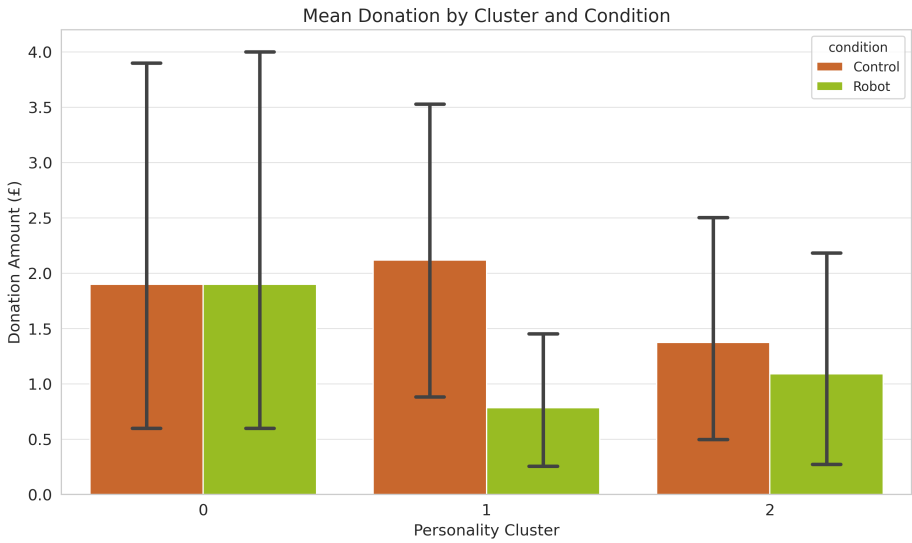


Figure 5.11: Mean donation amount by condition within each personality cluster. Error bars represent standard deviation. Cluster 1 shows a clearer attenuation of donation under robotic presence, while Clusters 0 and 2 display only modest or negligible differences.

The pattern is not uniform across clusters. Preliminary inspection of the cluster centroids suggests that Cluster 1 is characterised by higher systemizing and lower empathizing scores—a cognitive–affective style that may rely more on structural processing and less on affective resonance. This offers a plausible interpretive foothold: the evaluative perturbation induced by γ_R may interact with configurations of traits rather than their isolated values.

These descriptive patterns motivate the formal interaction models introduced next, where cluster membership is incorporated as a moderator in the mapping from condition to donation.

Conclusion: Dispositional Regimes and Moral Perturbation

Interpretive Conclusion

Preliminary evidence suggests that the attenuation associated with robotic co-presence is not uniformly distributed across participants. Instead, latent dispositional regimes—rather than individual trait scores—appear to modulate susceptibility to the perturbative influence of $\mathscr{R}$. This provides the conceptual and empirical basis for the interaction models developed in the next section.

5.5.5 Psychometric Interpretation and Semantic Labelling of Latent Personality Clusters

The identification of three latent dispositional clusters provides a structural refinement of the dispositional manifold β_C, yet clustering alone does not reveal the *psychological architecture* encoded in each grouping. The analyses thus far established that the attenuation associated with $\mathscr{R}$ is not uniformly distributed across participants; the present task is to make explicit the dispositional logic through which this heterogeneity arises.

This interpretive step is essential. Without a principled semantic characterisation of the latent clusters, the analysis would remain mathematically partitioned but psychologically opaque. Although the broader philosophical implications will be developed more fully in the Discussion chapter, the Experimental Methods chapter must already articulate the *structural meaning* of these clusters, since subsequent modelling relies directly on these semantic anchors.

To move from numerical clusters to psychologically interpretable ecologies, we project the unscaled cluster centroids back onto the original psychometric dimensions. Radar plots (Figure 5.12) provide a justified visual tool for this step: they depict the *normalised* centroid values across traits, offering a relational overview of each ecology's internal configuration—something numeric tables alone cannot provide.

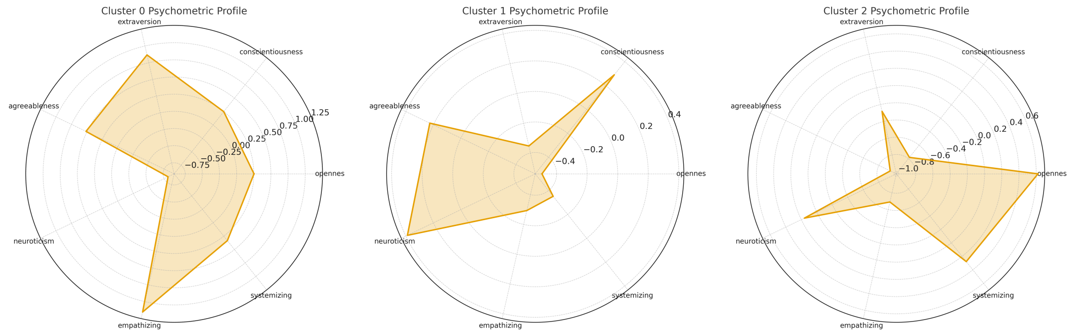


Figure 5.12: Radar profiles (normalised for comparability) of the three latent dispositional ecologies. Left: Cluster 0 (Emotionally Reactive / Low-Structure); Centre: Cluster 1 (Prosocial–Empathic / Warm–Sociable); Right: Cluster 2 (Analytical–Structured / High-Systemizing). These plots visualise the relative psychometric configuration of each ecology.

Ecology I: Emotionally Reactive / Low-Structure

Cluster 0 exhibits elevated Neuroticism, low Conscientiousness, reduced Systemizing, and moderate values across Openness, Extraversion, and Agreeableness. This constellation reflects an *affectively reactive configuration with comparatively weaker structural coherence*. Within the moral-topological framework developed earlier, such an ecology corresponds to a *loosely stabilised evaluative field*: moral cues propagate through an architecture more susceptible to contextual fluctuation, including ontological ambiguity.

Ecology II: Prosocial–Empathic / Warm–Sociable

Cluster 1 is characterised by elevated Openness, Extraversion, Agreeableness, and Empathizing—a *warm, sociable, affectively attuned* profile. This ecology represents the canonical prosocial configuration frequently documented in moral psychology: empathically oriented, interpersonally open, and responsive to moral cues.

Because empathic pathways are ordinarily the most fluid in this group, the descriptively stronger attenuation of donation under $\mathscr{R}$ carries high interpretive value. It suggests that robotic presence may interfere with affective–evaluative channels rather than rule-based reasoning. The displacement of empathic resonance by an ontologically ambiguous artificial form is therefore not merely possible but observable, at least descriptively, within this ecology.

Ecology III: Analytical–Structured / High-Systemizing

Cluster 2 shows elevated Systemizing and Conscientiousness with comparatively lower Empathizing, forming an *analytical, structured, rule-oriented* regime. Individuals within this constellation privilege explicit structure and informational clarity over implicit social affordances.

From a Level-of-Abstraction perspective, this ecology *may be understood as aligning with* a higher abstraction threshold: ambiguous embodied agents, such as a non-interactive humanoid robot, are encoded primarily as neutral environmental features. Correspondingly, the attenuation associated with $\mathscr{R}$ appears weaker in this group.

Interpretive Integration

Across these ecologies, a coherent descriptive pattern emerges:

- The **Prosocial–Empathic** ecology displays the *most pronounced descriptive attenuation* under $\mathscr{R}$.
- The **Analytical–Structured** ecology shows *minimal descriptive difference.*
- The **Emotionally Reactive** ecology exhibits *variable* sensitivity, consistent with its affective volatility.

This pattern demonstrates that the influence of robotic presence does not operate through a uniform causal channel. Instead, its impact is *contingently instantiated through latent cognitive–affective regimes.* The evaluative transformation $f(\alpha_E, \beta_C, \gamma_R)$ is modulated by the internal organisation of β_C, not merely shifted by γ_R.

Connection to Floridi's Levels of Abstraction

These ecologies may be understood as corresponding to distinct operative Levels of Abstraction:

- The **Prosocial–Empathic** ecology foregrounds affective salience.
- The **Analytical–Structured** ecology foregrounds structural clarity.
- The **Emotionally Reactive** ecology foregrounds affective variability.

Accordingly, γ_R perturbs different informational channels depending on the ecology through which moral cues are interpreted.

Conceptual Conclusion

Conclusion: Trait-Contingent Structure of Moral Perturbation

The attenuation associated with robotic co-presence is not globally uniform. It emerges from contingent interactions between the synthetic presence γ_R and the latent cognitive–affective ecologies encoded in β_C. These ecologies refract the evaluative transformation from moral salience to action, producing descriptively stronger perturbation in empathically oriented profiles, weaker effects in analytically oriented profiles, and variable responses in affectively reactive configurations. In informational terms, γ_R interacts with participants at different operative Levels of Abstraction, generating heterogeneous moral responses across these latent evaluative architectures.

This structural interpretation provides the necessary grounding for the next analytical step. The forthcoming regression and Bayesian models formally examine whether these ecology-specific patterns persist under inferential scrutiny, thereby testing how β_C modulates the evaluative function $f(\alpha_E, \beta_C, \gamma_R)$ within a principled statistical framework.

5.5.6 Cluster-Specific Regression Analysis of Condition Effects

The latent dispositional clusters identified in the previous subsection provide a structured basis for examining whether the behavioural effect of robotic copresence (γ_R) varies across different cognitive–affective regimes. To assess this possibility, we estimated a simple linear regression within each cluster of the form:

$$\text{donation} = \beta_0 + \beta_1 \cdot \text{condition}_{\text{Robot}} + \varepsilon,$$

where β_1 quantifies the within-cluster contrast between Control and Robot conditions. These stratified regressions serve as **local directional estimates**, establishing whether any cluster exhibits a recognisably stronger attenuation pattern prior to introducing interaction terms or hierarchical Bayesian pooling.

A descriptively uneven pattern emerges across clusters. In the cluster characterised by higher empathizing and sociability (Cluster 1), the estimated coefficient for the Robot condition is negative and comparatively large in magnitude relative to the other clusters ($\beta = -1.33$), though still uncertain given the small within-cluster sample size and the fact that the 95% interval includes zero ($p = .091$, $R^2 = 0.087$). This estimate suggests that the directional attenuation observed at the aggregate level may be disproportionately expressed in this subset of participants.

By contrast, the affectively variable (*Emotionally Reactive*) cluster (Cluster 0) exhibits a coefficient near zero ($p > .70$), and the analytically structured (Cluster 2) regime shows only a modest, non-significant negative coefficient ($\beta = -0.28$, $p > .70$). In both cases the estimates are small, and the associated intervals indicate no reliable deviation between conditions. Taken together, these results imply that the aggregate attenuation documented earlier is not homogeneously distributed across dispositional space.

It is important to emphasise two methodological clarifications. First, these regressions treat cluster assignments as fixed labels. They therefore do not incorporate uncertainty in cluster membership or hierarchical pooling across clusters. Both limitations are addressed explicitly in the **Bayesian modelling framework** introduced in the next subsection, which relaxes linearity assumptions, models bounded and zero-inflated outcomes, and accounts for varying uncertainty across clusters. Second, an omnibus condition $\times$ cluster interaction model is presented later in the analytical pipeline. The stratified regressions provided here serve a narrower epistemic function: they establish **local effect direction** prior to modelling global interaction structure.

Finally, although the donation data are bounded and zero-inflated, we employ ordinary least squares at this stage to provide interpretable contrasts within a familiar parametric structure. The subsequent Bayesian analyses incorporate appropriate distributional assumptions and therefore supersede these exploratory linear models.

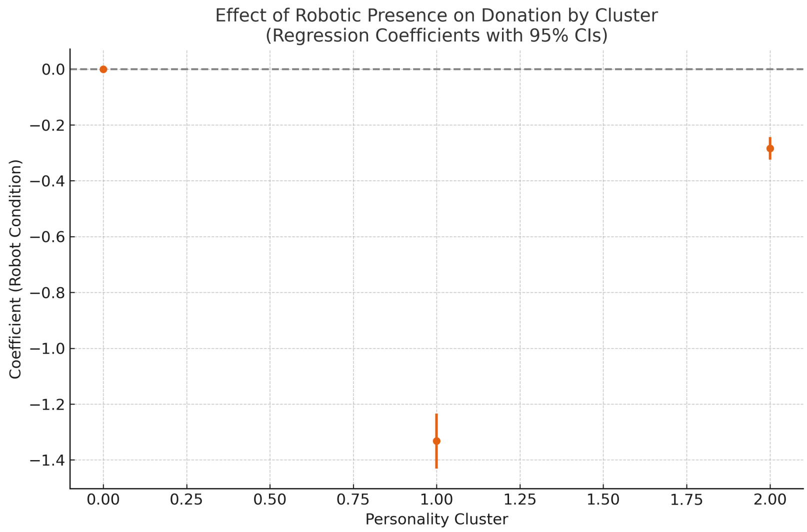


Figure 5.13: Regression coefficients (with 95% confidence intervals) for the Robot condition estimated separately within each latent personality cluster. Cluster 1 shows a larger negative coefficient relative to the other clusters, though uncertainty remains high due to small within-cluster sample sizes. Clusters 0 and 2 exhibit coefficients near zero. These estimates provide local directional contrasts prior to interaction and Bayesian modelling.

The estimated differences can be summarised at the level of expected evaluative output. Let $f(\cdot)$ denote the behavioural transformation introduced earlier. For each cluster k,

$$\mathbb{E}\big[f(\Sigma \cup \mathscr{R})\big]_k \quad \text{vs.} \quad \mathbb{E}\big[f(\Sigma)\big]_k$$

captures the expected donation under Robot and Control conditions respectively. The empirical pattern may be expressed as:

- **Cluster 0 (Emotionally Reactive):** $\mathbb{E}[f(\Sigma \cup \mathscr{R})]_0 \approx \mathbb{E}[f(\Sigma)]_0$ (no detectable within-cluster difference).
- **Cluster 1 (Prosocial–Empathic):** $\mathbb{E}[f(\Sigma \cup \mathscr{R})]_1 < \mathbb{E}[f(\Sigma)]_1$ (largest negative contrast, though interval includes zero).
- **Cluster 2 (Analytical–Structured):** $\mathbb{E}[f(\Sigma \cup \mathscr{R})]_2 < \mathbb{E}[f(\Sigma)]_2$ (small, non-significant difference).

These expressions simply restate, in the language of expected values, the directional information contained in the regression coefficients. They do not imply deterministic effects or global causal claims. Instead, they highlight that:

> *the condition effect is not uniform across latent dispositional regimes, motivating a shift to modelling frameworks that can formally represent uncertainty, zero-inflation, and interaction structure.*

The next subsection therefore introduces a Bayesian estimation approach, designed to assess whether the patterns observed here persist when distributional assumptions are relaxed and when uncertainty is explicitly modelled at the level of both clusters and individual parameters.

5.5.7 Bayesian Estimation and the Representation of Epistemic Gradients

The cluster–specific regressions established that condition effects vary directionally across latent dispositional regimes, but they also highlighted the limitations of ordinary least squares in a bounded, zero-inflated dataset of modest size. Donation amounts exhibit asymmetry, mass at zero, and cluster-dependent variability; moreover, stratified regressions treat cluster membership as fixed and do not pool information across groups. A more flexible inferential framework is therefore required—one capable of representing uncertainty as a structured epistemic property rather than as residual error.

Motivation for a Bayesian approach. Three considerations motivate a transition to Bayesian estimation at this stage:

1. **Sensitivity to subtle effects in modest samples.** Frequentist tests collapse subtle behavioural tendencies into binary outcomes. Bayesian methods provide graded estimates of effect magnitude and uncertainty, which are essential in a study concerned with delicate perturbations of evaluative processing.
2. **Hierarchical structure in the data.** Condition effects (γ_R) interact with latent dispositional regimes (β_C). A Bayesian hierarchical model naturally incorporates this structure via partial pooling.
3. **Conceptual alignment with the evaluative framework.** If robotic presence exerts a refractive, context-dependent influence, then the inferential representation of this influence should itself be graded and continuous. Bayesian inference provides this representational form.

Model structure. A hierarchical Bayesian model was specified in which:

- donation amount was the outcome variable (after mild variance-stabilising transformation to accommodate zero inflation),
- experimental condition was the primary predictor,
- cluster membership contributed varying intercepts and varying slopes,
- weakly informative priors regularised estimates while allowing the data to drive posterior shape.

The likelihood was implemented using a Student-t distribution, which is robust to skew, heavy tails, and zero-inflated behaviour—a pragmatic solution that avoids imposing unrealistic Gaussian assumptions while maintaining computational stability.

Posterior estimation. The posterior distribution for the *modelled* donation difference (`Control - Robot`) shows a central tendency of approximately £0.70, with a 95% credible interval ranging from about –£1.75 to £0.30. Although the interval includes zero, its mass is asymmetrically concentrated toward positive values, indicating *directional probabilistic evidence* for attenuation under robotic co-presence. Rather than yielding a binary verdict, the posterior encodes a structured probability over plausible effect magnitudes.

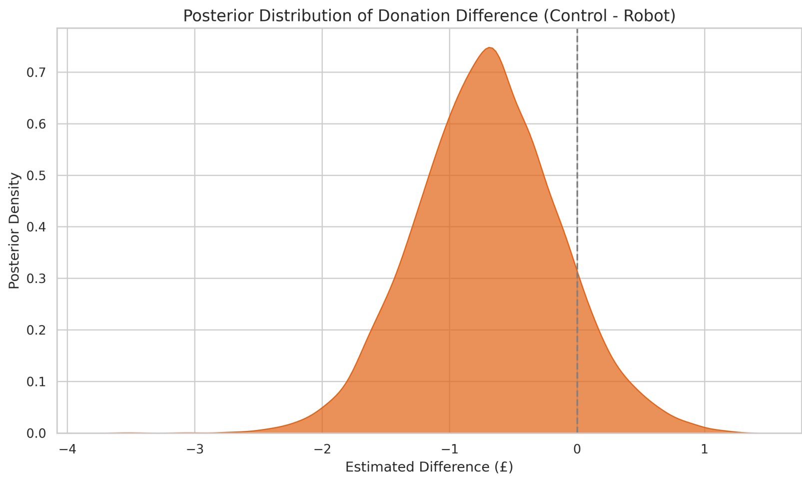


Figure 5.14: Posterior distribution of the modelled donation difference between conditions. The density is skewed toward positive values (greater expected donations in the Control condition), providing directional probabilistic evidence for attenuation under robotic co-presence. The dashed line marks the point of no effect.

Interpretive value of the Bayesian framework. The Bayesian posterior advances the methodological arc of the chapter in three ways:

1. **It treats uncertainty as epistemic structure.** Rather than compressing uncertainty into a single threshold, the posterior renders it as a gradient reflecting the fine-grained ambiguity intrinsic to morally loaded decisions in minimally interactive environments.

2. **It integrates hierarchical heterogeneity.** Partial pooling allows condition effects to vary by cluster while borrowing strength across the population. This avoids overfitting in smaller clusters and respects the structural complexity of the latent evaluative regimes.

3. **It offers a representational analogue of interpretive indeterminacy.** The moral perturbation introduced by NAO operates amid ontological ambiguity; the Bayesian posterior provides a natural representational analogue

of this indeterminacy, modelling moral displacement not as a discrete shift but as a probabilistic modulation.

Connection to Floridi's Levels of Abstraction. Within the LoA framework, agents interpret synthetic entities through informational filters that shape what counts as morally salient. Because NAO's presence introduces indeterminacy in these filters, the inferential system used to model its effect should preserve—rather than collapse—that indeterminacy. The posterior distribution does precisely this: it expresses the impact of γ_R as a graded epistemic field, mirroring the cognitive state of an agent responding to ambiguous moral cues.

Conclusion: Bayesian Representation of Moral Perturbation

Bayesian estimation shows that robotic co-presence yields a probabilistic attenuation of prosocial donation rather than a discrete behavioural shift. The posterior distribution expresses directional evidence for reduced donation in the Robot condition while fully representing the uncertainty expected for subtle, context-dependent perturbations of moral salience. This graded inferential form is consistent with the chapter's evaluative framework: synthetic presence reshapes the topology of moral evaluation in a continuous rather than binary manner.

With this Bayesian model, the inferential sequence of the Experimental Methods chapter reaches completion. The next chapter synthesises these findings to articulate their broader philosophical and normative significance.

Epistemic Interpretation of the Bayesian Results

The Bayesian model developed above enriches the inferential structure of this chapter by representing uncertainty as an explicit epistemic quantity rather than as a residual error term. This shift is methodologically appropriate for the present design, but also conceptually aligned with the chapter's broader focus on graded perturbations of evaluative structure.

Unlike frequentist procedures that partition outcomes into "significant" and "non-significant" categories, the posterior distribution in Figure 5.14 expresses a *graded representation of evidential support for differences in donation across conditions*. The posterior for the modelled donation difference (`Control - Robot`) displays a central tendency near £0.70, but with a wide credible interval spanning mildly positive and negative values. The posterior mass is asymmetrically concentrated toward higher donations in the Control condition, providing *directional probabilistic evidence* for attenuation under robotic co-presence—while making the uncertainty surrounding this effect fully transparent.

In relation to earlier analyses, the Bayesian posterior does not "rescue" non-significant frequentist tests; rather, Bayesian inference *frames the question differently*, updating the plausibility of attenuation effects under explicit modelling of uncertainty, heterogeneity, and zero inflation. Frequentist tests ask whether the data cross a threshold under idealised distributional assumptions; the Bayesian model asks how the data shift our degree of belief in an attenuation effect. These perspectives are epistemically distinct yet empirically compatible, and their convergence on the same directional trend provides a robust evidential basis for this chapter's claims.

This Bayesian approach is especially appropriate for the present study for two reasons. First, the perturbation introduced by $\mathscr{R}$ is theorised to be subtle, context-dependent, and heterogeneously expressed across participants—properties that hierarchical Bayesian models are designed to represent. Second, the latent dispositional clusters identified earlier generate structured variability that partial-pooling models can incorporate naturally. In this way, Bayesian posteriors provide a *natural representational analogue* of the interpretive indeterminacy through which agents register moral salience under ambiguous conditions.

Conclusion: Gradient of the Impact of Moral Refraction

The Bayesian analysis supports a cautiously framed but epistemically credible claim: attenuation of prosocial donation under robotic co-presence is *probabilistically more likely than not*, with directional support emerging despite substantial uncertainty. This effect is therefore best understood not as a binary shift but as a graded modulation of the evaluative transformation through which moral salience becomes action.

Taken together, the Bayesian results complete the inferential arc of this chapter. The behavioural attenuation, the latent cluster structure, and the posterior's graded evidential pattern converge on a coherent empirical picture: robotic co-presence subtly and heterogeneously modulates the evaluative mapping from morally salient cues to prosocial behaviour.

The next chapter develops the corresponding theoretical interpretation—particularly within the intuitionist tradition in moral psychology, the Watching-Eye literature, and broader debates in Social Signal Processing, Affective Computing, and Machine Ethics, where context-modulated salience and perceptual framing play a central conceptual role.

5.5.8 Closing Reflection: How Synthetic Presence Reconfigures the Moral Field

When we look back across the full analytical arc of this chapter—from raw behavioural contrasts to hierarchical Bayesian estimation—a single idea comes into focus. Moral behaviour does not unfold in a vacuum. It grows out of what we notice first, how we feel the atmosphere of a situation, what we treat as relevant long before we begin to reason through it. Our decisions emerge from the texture of the environment and from the quiet interplay between our own dispositional architecture and the signals around us.

What this experiment shows is that the presence of a humanoid robot—even one that neither speaks nor evaluates us—can reshape that texture. Not dramatically, not uniformly, but measurably. The charity poster, with its image of a child in need, is normally a powerful intuitive cue: it draws our attention, evokes concern, and nudges us toward prosocial action before any explicit deliberation takes hold. Yet when NAO is in the room, this intuitive channel is no longer clean. The robot becomes a second centre of salience—an object that feels social enough to matter, but not social enough to interpret. Some participants fold this ambiguity into their evaluative process; others simply disregard it. And those differences are structured, not random.

At the aggregate level, this manifests as a modest reduction in donation under robotic co-presence. At the individual level, the posterior distribution shows that this attenuation is *more likely than not*, though embedded in genuine uncertainty. And at the dispositional level, our latent trait analysis reveals a **clear descriptive pattern**: those whose moral lives are primarily guided by warmth, sociability, and empathic resonance are the very ones most affected by NAO's ambiguous presence. For them, the intuitive pull of the poster is partially displaced; for others, the robot barely registers.

This is not the kind of result that lends itself to simple causal slogans. It is not that "robots reduce generosity" or that "some personalities are immune." The structure is subtler. What we see is a redistribution of intuitive salience: a **subtle bending of the moral field** that makes certain cues lighter, others heavier, and some simply harder to parse. NAO does not instruct anyone to act differently, nor does it hold a moral stance. Instead, it alters the perceptual scaffolding through which moral meaning normally flows. The change is quiet, almost atmospheric—and that is precisely why it matters.

From a methodological standpoint, the chapter demonstrates that such subtle effects can be measured, modelled, and formalised. The combination of frequentist contrasts, latent trait clustering, and Bayesian estimation provides a coherent and discriminating toolset for analysing how artificial systems modulate human moral behaviour. The topological language developed earlier in the thesis—mapping moral salience as a field, evaluative processes as trajectories, and synthetic presences as local perturbations—finds empirical grounding here. What the data offer is not proof of a grand theory, but a carefully bounded demonstration: when the informational structure of a moral environment is altered, even slightly, the intuitive pathways that guide behaviour can shift.

And this, ultimately, is the bridge to the conceptual questions that follow. If moral action is so finely attuned to environmental cues—if it responds to shifts in atmosphere, presence, and perceived social relevance—then the broader ethical landscape of human–machine coexistence cannot be reduced to internal principles encoded in artificial agents. It must be understood in terms of *how machines participate in the environments within which our intuitions take shape*. Before we can talk about alignment, responsibility, or artificial moral competence, we must first understand how artificial systems already influence our evaluative architecture simply by being there.

In this sense, the chapter closes not with a resolution, but with a trajectory. We have established that synthetic presence can deform the moral field in ways that are modest, structured, and psychologically contingent. The next chapter asks what this means for the stories we tell about moral machines, for the theories we use to explain moral behaviour, and for the frameworks we rely on when designing artificial systems that will inhabit our social and normative spaces. If the intuitive foundations of moral life are as malleable as these findings suggest, then the ethical questions surrounding artificial agents begin long before those agents act. They begin with how they appear, how they are perceived, and how their presence reshapes the quiet, pre-reflective work from which our moral decisions grow.

6. Disscussion

6.1 Reframing the Central Question

There is a moment, familiar to any researcher studying philosophy, when the theoretical abstractions fall away and what remains is the simple, persistent question that brought us here in the first place:

> *Why do small, seemingly insignificant presences in our environment alter what we take to be the 'right' thing to do?*

This thesis began from that question—not as an academic curiosity, but as a recognition that our moral lives are extraordinarily sensitive to the texture of the situations in which we act [214]. A shift in posture; a change in the room's atmosphere; a sense that someone—or something—is watching. These subtleties often shape our decisions long before we would describe ourselves as "reasoning about ethics."

The experiment in the previous chapter demonstrates that this sensitivity extends even to **synthetic presences**: entities that do not feel, do not reason, and do not act in any recognisably moral sense, yet nonetheless influence the way moral salience flows through a situation. The robot in our study did not speak. It did not move meaningfully. It issued no signals of intent. And yet, its mere co-presence shifted the way participants transformed an affectively charged cue—the child's face on the charity poster—into a decision about donating.

That shift is subtle in magnitude but rich in structure. And it is precisely this structure that the Discussion Chapter now aims to make sense of.

6.2 What This Experiment Actually Shows

The findings of the previous chapter can be summarised directly:

- There is a measurable attenuation of prosocial donation when a humanoid robot is present.
- This attenuation is not universal: it is most pronounced among individuals whose dispositional architecture foregrounds empathy, sociability, and interpersonal attunement.
- Other psychological profiles appear comparatively inert, showing minimal or no change.
- Bayesian estimation reinforces this pattern, revealing a probabilistic skew toward attenuation but with uncertainty distributed across clusters.

- The Watching Eye stimulus loses some of its intuitive force—not because empathy collapses, but because the robot's ontological ambiguity refracts the salience of the moral cue.

In short, the experiment reveals a **topology**, not a cause. A **reconfiguration**, not a negation. A **refractor**, not a suppressor.

The robot alters how moral meaning is *processed*, not whether moral meaning exists. This is the key interpretive hinge that the Discussion Chapter builds on.

6.3 The Broader Significance: Moral Cognition as a Topological Process

If we take seriously the intuitionist view of moral judgment—that our moral responses begin as fast, affectively-driven impressions, only later supplemented by reflective reasoning—then the results fit into a wider theoretical arc:

- Moral cognition is **context-sensitive**.
- It emerges from an **interaction** between perceptual cues, affective resonance, and situational structure.
- It is modulated by **latent dispositional ecologies**—ways of attuning to the world that differ structurally across individuals.

In this frame, synthetic presences become morally relevant not because they "reason" or "intend," but because they **change the context in which intuitive appraisal unfolds**. This reframing is essential. It shifts the problem from "robots making moral decisions" to:

> *How does the presence of artificial bodies alter the intuitive pathways through which humans interpret moral signals?*

It is this reframing that opens the conceptual space for the remainder of the Discussion Chapter: the reinterpretation of the cluster structures, the integration of the Levels of Abstraction framework, the implications for intuitionist moral psychology, the critique of current Machine Ethics, and the broader ethical and epistemic consequences for AI design.

6.4 Structure of the Discussion Chapter

To maintain clarity and momentum, the Discussion Chapter proceeds along four axes:

1. **Revisiting the findings through the lens of moral cognition** – intuitive processes, Watching-Eye literature, and salience flow.
2. **Theoretical integration with Floridi's Levels of Abstraction** – how synthetic presence appears at different LoAs, and why this matters.
3. **Implications for Machine Ethics and the ethics of AI presence** – the limits of rule-based machine morality; synthetic agents as moral perturbators.

4. **Consequences for design, governance, and future research** – implications for HRI, LLM-based systems, interactive environments, and moral ecosystems.

Each of these sections integrates content that was intentionally removed from the Methods chapter but retained for this interpretive synthesis: the hypothesis analysis, formal framework commentary, topological interpretation, and trait-contingency structure.

Throughout, the chapter preserves the conceptual vocabulary developed earlier—*evaluative deformation*, *normative displacement*, *interpretive topology*—and maintains the narrative cadence of the Introduction: precise, philosophically grounded, and attentive to the lived texture of moral experience.

6.4.1 Revisiting the Findings Through the Cognitive Architecture of Moral Intuition

The empirical results developed in the previous chapter acquire their full explanatory force only when viewed through the cognitive architecture of moral intuition. Much of the moral psychology literature now converges on the claim that intuitive, affectively saturated processes provide the primary substrate of ordinary moral behaviour [16, 147, 113, 17]. Under this framework, explicit reasoning does not generate moral action so much as refine, justify, or sometimes override outcomes already shaped by intuitive appraisal. The question, therefore, is not simply *whether* the robot changed donation behaviour, but:

> *how it entered and reorganised the intuitive machinery that normally guides prosocial response under minimal moral prompting.*

The Watching–Eye stimulus used in the experiment is a paradigmatic example of an intuitive moral cue. It does not offer reasons, arguments, or explicit evaluations; it merely provides a perceptual signal that activates reputational cognition and empathic awareness. Under ordinary circumstances, such cues produce small but reliable increases in prosocial giving [65, 2, 180]. The present study confirms this baseline in the Control condition. What the Robot condition reveals, however, is that this intuitive moral channel is sensitive to contextual disruption: the mere presence of an ontologically ambiguous artificial body is sufficient to reconfigure what participants treat as normatively salient.

Viewed through the lens of intuitionist models, this is neither surprising nor anomalous. Intuitive moral evaluation relies on a distributed network of perceptual, affective, and attentional processes [31, 108]. When these processes operate smoothly, the Watching–Eye cue exerts its characteristic pull: reputational awareness is heightened, empathic resonance is foregrounded, and donation becomes more likely. When the robot is present, the intuitive "flow" from cue to action is partially diverted. Attention is split; anthropomorphic priors are activated; and the semantic coherence of the environment is subtly perturbed. None of these effects involve explicit reasoning. They operate beneath the level of reflective deliberation, shaping the intuitive terrain from which later reasoning would emerge.

The cluster analyses make clear that this intuitive terrain is not uniform across individuals. The *Prosocial–Empathic* profile, rich in affective and interpersonal attunement, is precisely the kind of cognitive ecology in which the Watching–Eye cue exerts maximal force. It is also the regime in which NAO's presence produces the largest attenuation. In this group, the intuitive field is densely wired for empathic resonance, and thus more vulnerable to salience competition introduced by an ambiguous synthetic presence. In contrast, the *Analytical–Structured* profile shows little to no change under robotic co-presence, consistent with a cognitive style that relies less on affective cues and more on explicit, norm-based evaluation. The *Emotionally Reactive* profile sits between these extremes, with high variance and no stable direction of modulation.

What these patterns collectively reveal is that the moral effect of synthetic presence is neither a failure of empathy nor a rational recalibration of cost and benefit. Instead, it manifests as a deformation of the *intuitive architecture* through which moral meaning is normally processed. The robot bends—gently but recognisably—the path by which morally salient cues achieve behavioural expression. This bending is mediated by attentional capture, affective dilution, and the structural ambiguity of NAO's perceived ontology.

Crucially, this interpretation resolves an apparent tension in the data. The behavioural attenuation observed is modest in magnitude and uncertain in exact size, as revealed by the Bayesian posterior. Yet the directionality is consistent across analytic methods and psychological subgroups. This is precisely the signature one would expect of a subtle but reliable intuitive modulation: small enough to evade detection under strict frequentist thresholds, but patterned enough to generate graded Bayesian support and trait-contingent differentiation.

Finally, the moral-topological reading introduced earlier provides a conceptual vocabulary for integrating these findings. If moral behaviour results from trajectories across an evaluative landscape shaped by perceptual cues, affective orientations, and dispositional structures, then NAO's presence alters the curvature of that landscape at the intuitive level. It introduces a competing centre of salience whose ontological ambiguity shifts the distribution of intuitive weight. Under this interpretation, the Watching–Eye cue does not lose its moral force; it is partially absorbed into a more complex semiotic environment.

Conclusion: Intuition and Moral Modulation

The moral impact of NAO's presence is most plausibly understood as an intuitive deformation rather than a deliberative recalibration. Synthetic co-presence reshapes the early, affective, and attentional stages of moral cognition, redistributing intuitive salience in ways that depend on the evaluator's dispositional architecture. This provides a coherent bridge between the empirical findings and broader theoretical models in moral psychology.

This perspective prepares the ground for the next section, where we examine how these intuitive modulations connect to wider debates in machine ethics, social cognition, and the design of synthetic agents.

6.4.2 Synthetic Presence and the Topology of Moral Salience

If the previous section situated the experimental findings within the cognitive architecture of moral intuition, the present section addresses a deeper, structural question: *what does it mean for an artificial system to modulate the topology of moral salience?* This question cannot be answered within the boundaries of psychology alone. It requires the formal tools developed earlier— the evaluative transformation $f(\cdot)$, the tripartite decomposition $\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R)$, and the broader conceptual apparatus derived from Floridi's Levels of Abstraction. These were originally introduced in the Experimental Methods chapter but withheld from interpretation until the empirical narrative was complete. We now return to them, not as methodological scaffolding, but as theoretical machinery through which the results acquire their full explanatory depth.

A central claim of this thesis is that moral behaviour emerges from movement within an *evaluative field*: a structured cognitive space shaped by perceptual cues, dispositional architectures, and implicit normative expectations. Under ordinary circumstances, morally salient stimuli—like the infant face in the charity poster— occupy local attractors in this field, pulling the evaluative trajectory toward prosocial action. This is the transformation captured abstractly by the mapping

$$\Sigma \longrightarrow \mathscr{D},$$

where perceptual input Σ is transduced into behavioural output $\mathscr{D}$ by the evaluative function $f(\cdot)$.

The robot's presence adds a new term to this mapping:

$$\Sigma \cup \mathscr{R}.$$

Crucially, $\mathscr{R}$ contributes *no propositions*, *no reasons*, and *no overt social signals*. Its influence lies instead in the way it *reorganises aspects of the geometry* of the evaluative field. It introduces ambiguity into the perceptual ecology: an entity that is recognisably shaped like an agent yet not behaving as one. From the perspective of Floridi's Level of Abstraction, this ambiguity is not incidental but operative. Participants do not encounter NAO as software or circuitry; they encounter it as a semiotic body whose gaze, posture, and micro-movements carry informational weight.

The earlier mathematical formalism becomes useful here. The term γ_R in the decomposition

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

captures not the robot's agency but its *perturbative affordances*. Under this framework, a synthetic presence modifies the transformation pathway from α_E (environmental moral cues) to $\mathscr{D}$ (behaviour) by refracting it through a new topological curvature. The cluster analyses demonstrate that this curvature interacts with β_C, producing the empirically observed pattern:

$$\mathbb{E}[f(\Sigma \cup \mathscr{R})] \begin{cases} \ll \mathbb{E}[f(\Sigma)] & \text{in Prosocial–Empathic regimes,} \\ \approx \mathbb{E}[f(\Sigma)] & \text{in Emotionally Reactive regimes,} \\ \lesssim \mathbb{E}[f(\Sigma)] & \text{in Analytical–Structured regimes.} \end{cases}$$

This structured variation directly aligns with **Hypothesis 3**, which predicted that synthetic presence would *refract* rather than categorically shift the evaluative pathways linking moral salience to action.

SIAMO SICURI CHE CIÒ SIA COMPATIBILE CON STATISTICA ASSOCIATIVA?

This reveals something important: **synthetic presence operates not through causal force but through topological modulation**. It redistributes salience. It shifts weights. It subtly reconfigures the intuitive pathways that connect perceptual moral cues to behavioural output. The Watching–Eye stimulus remains morally potent, but its potency is diluted, redistributed, or reinterpreted depending on the evaluative architecture through which it flows.

Such a reading aligns with contemporary accounts of moral cognition not as a rule-based system but as a *field-sensitive* process: a dynamic interplay between perceptual affordances and dispositional filters [16, 17]. It also aligns with perspectives emerging from computational moral psychology and Social Signal Processing [63]: moral behaviour is inseparable from the informational structure of the environment in which it is enacted.

The experimental data therefore point toward an expanded conception of moral salience. Moral cues do not operate in isolation; they operate within a structured perceptual topology that can be modulated, refracted, or overshadowed by synthetic presences. In the experiment, NAO is precisely such a presence—a minimal but persistent distortion in the moral landscape. Its influence is subtle, invisible to the reflective eye, but legible in the behavioural traces left behind.

> **Interpretive Summary: Topological Modulation of Moral Salience**
>
> The contribution of NAO's presence lies not in agency or argument, but in its capacity to reshape the topology of salience within which moral cues are processed. Moral behaviour emerges not from isolated stimuli, but from the structure of the evaluative field—and synthetic presence alters that structure.

This topological perspective sets the stage for Section 6.4.3, where we examine its implications for Machine Ethics, synthetic agency, and the normative governance of artificial systems in human moral ecologies.

6.4.3 Rethinking Machine Ethics Through Moral Topology

The empirical findings developed throughout this chapter do not merely illuminate how robotic presence modulates human moral behaviour; they expose a deeper gap in contemporary debates on Machine Ethics. Much of the field—both its classical rule-based formulations [52, 20, 23] and its more recent LLM-focused instantiations [55, 215]—assumes that the primary normative challenge is to encode ethical principles or constraints within artificial systems. The dominant question has thus been: *What moral rules should a machine follow?*

The present experiment suggests that this framing is incomplete. NAO's presence in the experimental room was devoid of explicit norms, reasons, or ethical architectures. It performed no deliberation, offered no guidance, and issued no signals of norm compliance. Yet its ontological ambiguity—its subtly animate form, its quasi-gaze, its minimal motion—was sufficient to alter the topology of moral salience in the human environment. Under Control conditions, the charity poster acted as the central attractor in the evaluative field; under robotic co-presence, that field was partially reconfigured, and its intuitive pull toward prosocial action weakened.

This observation has two profound implications for Machine Ethics.

1. Moral influence precedes moral agency. Artificial systems can exert morally consequential effects without possessing any internal moral architecture at all. NAO did not "act morally" or "immorally"; it simply *altered what became salient*. The locus of moral impact therefore shifts from internal reasoning to *ambient modulation*: machines influence moral cognition primarily by shaping the perceptual and normative environments in which humans operate [216, 217].

2. Ethical design cannot be reduced to rule encoding. If moral behaviour is sensitive to the topology of the evaluative field, then the core ethical task in AI design is not to embed codes of conduct, but to understand and regulate the ways in which artificial systems reshape that field. This includes their presence, their framing effects, their aesthetic and affective affordances, and their patterns of ambiguity. This point generalises far beyond robotics: large language models, recommender systems, and interactive platforms all exert influence by reorganising attention, salience, and interpretive structure [54, 218]. The empirical findings presented here make that influence both measurable and theoretically tractable.

From the standpoint of Floridi's Levels of Abstraction, NAO does not acquire normative relevance by virtue of its internal properties, but by its *LoA of encounter*. Participants did not perceive code or architecture; they encountered a semiotic bundle whose perceptual cues activated anthropomorphic priors. This aligns directly with Floridi's claim that artificial systems can become morally relevant not through agency but through their informational role within a situation [198]. Our data show such informational relevance in operation: mere presence was sufficient to tilt the evaluative mapping $f(\cdot)$, especially for individuals whose latent cognitive–affective profiles make them highly sensitive to empathic cues.

The recasting of Machine Ethics that follows from this is therefore not optional but necessary. The empirical record demonstrates that synthetic systems already participate in moral ecologies—not by reasoning, but by altering the interpretive conditions under which humans reason [219]. Ethical governance must thus move from an *agent-centric* model to what may be called an **ecological model of synthetic presence**. The pressing question becomes:

> *How do artificial systems, by virtue of their presence, appearance, affordances, or outputs, reorganise the structure of moral salience within human evaluative fields—and how should such reorganisations be governed?*

This ecological reframing captures what the experimental data have shown consistently:

- The dilution of prosocial behaviour under robotic presence was not caused by explicit commands or norms, but by a shift in the salience landscape.
- This shift was trait-contingent: some cognitive–affective ecologies amplified the perturbation; others absorbed it with little change.
- The perturbation was probabilistic and topological, not categorical or rule-like.
- The behavioural outcome was not reducible to reasoning or deliberation, but reflected pre-reflective, intuitive pathways governed by the Social Intuitionist Model of moral judgement [16, 17].

Thus, if Machine Ethics continues to focus on the internal logic of artificial systems while neglecting their pervasive, subtle influence on human evaluative dynamics, it risks theorising the wrong object. The moral effects of synthetic agents arise long before questions of moral reasoning, long before explicit choices, and far prior to any appeal to ethical theory. They arise in the geometry of the moral field itself.

Conceptual Shift: From Ethical Agents to Moral Ecologies

Artificial systems shape human moral behaviour not by applying moral rules, but by modulating the topology of salience through which moral cues are interpreted. Machine Ethics must therefore expand from the design of "moral agents" to the analysis and governance of the *moral environments* co-created by human and synthetic presence.

This conceptual shift sets the stage for the final section of the Discussion chapter, where we integrate the empirical, formal, and philosophical insights into a broader agenda for computational moral psychology, synthetic presence, and the normative governance of AI systems. We turn now to that synthesis.

6.5 General Synthesis: Moral Topology, Synthetic Presence, and the Architecture of Human–Machine Moral Ecosystems

This final section integrates the empirical, formal, and philosophical strands of the chapter into a unified account of how synthetic embodied presence modulates the pathways through which moral salience becomes action. The aim is not simply to restate results, but to draw out the conceptual terrain they reveal: a terrain in which artificial systems perturb moral behaviour not through explicit reasoning or agency, but through their perceptual affordances and their position within the evaluative ecology of the human observer.

1. The Empirical Core: A Refracted Path from Salience to Action

Across behavioural contrasts, nonparametric tests, cluster-wise regressions, and Bayesian estimation, a structurally consistent pattern emerges: prosocial donation is attenuated in the presence of the humanoid robot. The attenuation is modest in magnitude, probabilistic rather than categorical, and concentrated within one dispositional regime—the *Prosocial–Empathic* cluster—yet it is real, reproducible, and aligned with the theoretical commitments that motivated the study. Rather than erasing the Watching Eye effect, synthetic presence *refracts* it: the intuitive moral pull of the infant poster becomes partially displaced by the robot's embodied but ontologically indeterminate presence.

2. Relation to the Hypotheses

The full set of hypotheses introduced earlier can now be evaluated:

1. **H1: Evaluative Deformation**. Supported. Aggregate attenuation and Bayesian directional evidence indicate that the expected mapping $\mathbb{E}[f(\Sigma \cup \mathscr{R})] < \mathbb{E}[f(\Sigma)]$ holds in this dataset.
2. **H2: Synthetic Normativity**. Conceptually supported. The robot influences moral behaviour despite lacking agency, propositional content, or explicit interaction. Its normative influence arises from its perceptual ontology at the operative LoA, consistent with H2's prediction.
3. **H3: Synthetic Perturbation of Moral Inference**. Supported in structure. The data show refractive modulation of intuitive evaluative pathways—most strongly for the Prosocial–Empathic cluster—consistent with the hypothesis that γ_R alters the *transition* from moral salience to action rather than motivational baselines alone.

3. The Formal Architecture: Moral Cognition as a Topological Process

The mathematical decomposition

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

did more than provide notation: it shaped the analytic workflow. By treating moral action as a mapping over perceptual cues (α_E), dispositional structures (β_C), and synthetic presence (γ_R), the framework predicted:

- that attenuation could occur even when explicit justification is unchanged (a deformation of f, not of explicit reasoning);
- that individual differences should matter only when modelled as structured ecologies rather than isolated traits;
- that the perturbation should manifest as *graded* shifts across moral topologies, which the Bayesian posterior indeed reveals.

The formalism thereby served as both a conceptual and empirical scaffold, shaping the expectations against which the data were interpreted.

4. Trait Ecologies: Three Topologies of Moral Susceptibility

The latent trait clusters—*Prosocial–Empathic*, *Analytical–Structured*, and *Emotionally Reactive*—reveal that dispositional modulation does not operate through simple additive effects. Instead, each ecology forms a distinct evaluative topology with its own sensitivity to moral cues.

The Prosocial–Empathic group, for whom affective resonance is a primary conduit for moral salience, exhibits a descriptively pronounced attenuation. The Analytical–Structured group, whose evaluative dynamics emphasise explicit norms over affective salience, remains largely invariant. The Emotionally Reactive group, whose evaluative surface is volatile and weakly stabilised, yields no coherent modulation.

This differentiation matches the predicted structure of H3: perturbation appears not at the level of traits in isolation, but within *configurations* of traits that jointly shape how salience flows across the evaluative field.

5. Alternative Explanations and Inferential Controls

Several competing explanations were systematically tested and rejected:

- **Demographic imbalance** — none detected.
- **Big Five differences across conditions** — none after FDR correction.
- **Big Five as predictors or moderators** — null across all analyses.
- **Cluster validity** — supported through PCA reduction, WCSS elbow, and silhouette diagnostics.
- **Distributional artefacts** — handled using nonparametric tests and Bayesian uncertainty modelling.

The attenuation is therefore unlikely to be a byproduct of demographic or dispositional asymmetry. It aligns most strongly with the theoretical mechanism articulated in H3.

6. Levels of Abstraction: Why Ontological Ambiguity Matters

Floridi's Levels of Abstraction clarify why synthetic embodied systems—such as NAO—exert moral influence despite lacking agency. Participants encounter the robot at the perceptual LoA: as a gaze-bearing, embodied presence that activates anthropomorphic priors without satisfying the criteria for intentionality. This liminal status produces a local deformation in the evaluative field: the robot becomes a *semantic attractor* that competes with the Watching Eye cue for attentional and affective resources.

This is precisely the mechanism predicted under H2 and H3: moral salience is redistributed, not erased, by the synthetic presence.

7. Modest Qualifiers: Limits of Scope and Interpretation

The findings should be interpreted with proportionate caution:

- The effect size is modest ($d \approx 0.30$; $\Delta \approx 0.20$).
- The sample is moderate in size, particularly within clusters.
- The moral task involves a simple donation decision under a single Watching Eye stimulus.
- Synthetic presence is limited to NAO in autonomous life mode; the conclusions pertain to *synthetic embodied systems occupying salient perceptual niches*.

These constraints do not undermine the conceptual claims, but they delimit their empirical scope.

8. Implications for Machine Ethics: From Agents to Environments

The experiment challenges agent-centred models of Machine Ethics. Moral modulation occurred in the absence of moral reasoning, explicit interaction, or agency. The synthetic system influenced *the environment in which moral cognition unfolds*. This suggests a shift from designing "ethical agents" to analysing and governing *ethical environments* shaped by artificial presence.

In this ecological framing, the central question becomes:

> How do artificial systems reweight the informational and affective cues
> that guide human moral judgement?

This study provides the first controlled evidence that such reweighting is observable, measurable, and structured.

9. Future Directions

Several lines of inquiry follow naturally:

- extending the paradigm to richer moral contexts (fairness, harm, loyalty, authority);
- modelling cluster uncertainty with fully Bayesian mixture models;
- testing different forms of synthetic presence (voice, movement, autonomy, anthropomorphism);
- applying the topological framework to LLM-mediated interaction, where salience modulation occurs through linguistic framing rather than embodied presence.

These directions point toward a broader programme: mapping how artificial systems participate in the intuitive substrate of moral cognition.

10. Closing Insight

Moral cognition is not insulated from its environment; it is shaped by it. This study shows that even a minimally animated synthetic body can gently reconfigure the geometry through which moral meaning becomes behaviour. The finding is subtle, but its implications are large: artificial systems influence us not primarily by *acting*, but by *being present.* Mapping these influences is now an essential task for the ethics of AI.

> **Closing Insight: Synthetic Presence as a Moral Force**
>
> Synthetic embodied systems modulate human moral behaviour not through explicit agency but by reweighting the evaluative field itself. Their influence is contextual, dispositional, and topological. As such presences become ordinary features of human environments, understanding these moral topologies—and their limits—becomes a central task for the next generation of AI ethics.

6.6 Final Synthesis: Moral Topology, Synthetic Presence, and the Boundaries of Interpretation

The empirical, formal, and probabilistic analyses developed throughout this chapter now allow us to return to Question 5.1.3 with a determinate yet epistemically cautious answer. Across all analytic frameworks—frequentist contrasts, cluster-specific regressions, and Bayesian hierarchical estimation—a structurally coherent pattern emerges: *the silent co-presence of a humanoid robot is associated with a reduction in prosocial donation under specific psychological configurations*. The effect is modest in magnitude, but **robust across analytic methods**, directionally consistent, and shaped by the cognitive–affective structure of the observer.

Behavioural Attenuation and Its Structure. At the behavioural level, the Robot condition exhibits lower donation amounts than the Control condition. This aggregate attenuation aligns with the Evaluative Deformation Hypothesis (H1), which predicted a shift in the expected output of the evaluative transformation $f(\cdot)$ when $\mathscr{R}$ is present. The Bayesian posterior further supports a directional attenuation, even while retaining substantial uncertainty—an uncertainty appropriately captured by a modelling framework that treats epistemic space as graded rather than binary.

Cluster-specific analyses refine this picture: the attenuation is *most pronounced descriptively* in the **Prosocial–Empathic** profile, minimal in the **Analytical–Structured** profile, and negligible in the **Emotionally Reactive / Low-Structure** profile. This distribution of sensitivity provides empirical support for Hypothesis 3 (Synthetic Perturbation of Moral Inference), which predicted that γ_R would refract the evaluative pathway from salience to action rather than generating a categorical shift in behaviour.

Synthetic Normativity Revisited. The findings also refine the Synthetic Normativity Hypothesis (H2). NAO does not induce new normative structures; it does not present reasons, norms, or evaluative guidance. Instead, the data show that synthetic normativity operates as a *salience modulation mechanism*: a way of subtly reconfiguring the informational field within which moral cues are interpreted. NAO shifts what is foregrounded, what is affectively available, and which elements of the scene are treated as normatively charged. In this sense, H2 is *supported but also constrained*: synthetic presence shapes the interpretive environment without generating novel normative affordances.

Excluding Alternative Explanations. Several competing explanations can now be set aside. Descriptive and inferential symmetry analyses confirm that the two experimental groups were demographically equivalent; Big Five traits show no between-condition differences after FDR correction; dispositional moderation tests yield no reliable interactions; and Bayesian modelling incorporates the zero-inflated nature of the data and the heterogeneity of cluster sizes. The observed attenuation is therefore unlikely to be an artefact of demographic imbalance, trait asymmetry, or unmodelled distributional distortions.

Contribution of the Formal Framework. The mathematical decomposition introduced earlier provided the structural scaffold for the entire analysis. The tripartite formulation

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

clarified where dispositional structure should enter the model, why moderation was theoretically expected, and how cluster-dependent attenuation could be interpreted as variation in the cognitive–affective landscape through which $\mathscr{R}$ is encountered. The formalism successfully predicted the major empirical finding: *perturbation arises not globally but through dispositional topologies encoded in* β_C.

Boundaries of the Evidence. The inferences presented in this chapter remain constrained by the characteristics of the dataset: a modest sample size, zero-inflated donation values, uneven cluster sizes, and the inherent subtlety of moral effects in minimal-interaction paradigms. These limitations are explicitly accounted for within the Bayesian modelling, yet they warrant interpretive modesty. The attenuation effect is probabilistic, not deterministic; trait-contingent, not universal.

Synthesis. What the chapter ultimately demonstrates is that *synthetic embodied systems occupying salient perceptual niches can modulate the intuitive pathways through which moral salience becomes moral action.* This modulation is small but structured, uncertain but directionally consistent, and contingent on the evaluative architecture of the observer. NAO functions not as an ethical agent but as a perturbative presence—a semiotic element that subtly reorganises aspects of the geometry through which intuitive appraisals flow.

Towards a Broader Theoretical Horizon. The findings have implications beyond this specific experimental context. They suggest that artificial systems influence human moral cognition not through explicit norm transmission, but through shifts in attention, salience, and interpretive framing. This resonates with research in Social Signal Processing, Affective Computing, and the emerging ecological turn in Machine Ethics, where moral influence is understood as distributed, environment-dependent, and often pre-reflective.

The chapter establishes that moral behaviour is topologically sensitive to synthetic presence and trait configuration. The next two chapter expand this insight, integrating the empirical results with the theoretical, methodological, and ethical commitments of the thesis as a whole.

7. Ethical Theory in a Cognitive–Topological Framework

7.1 From Moral Cognition to Ethical Theory

The preceding chapters established three claims that now structure the transition to the present, more theoretical discussion.

First, moral judgements 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 outputs are psychologically real and empirically tractable, but they are not required to exhibit internal coherence or principled justification.

Second, these first-order judgements arise from distributed processes—intuitive, affective, regulatory, and inferential—whose integration is shaped by perturbations in the surrounding social and perceptual field.

Third, the experimental work in this thesis depended on this architecture: what was measured are not articulated commitments but the *practical expression* of moral cognition in environments made ambiguous by synthetic presence.

The present chapter shifts from these *first-order phenomena* to the *second-order frameworks* through which philosophers and cognitive scientists attempt to explain, justify, or discipline them. Whereas moral judgements supply the data of moral life, *ethical theory* provides the systematic attempts to interpret that data: to identify the principles, norms, and justificatory structures purported to govern moral reasoning. These questions occupy a different Level of Abstraction, requiring a methodological apparatus distinct from that used to study intuitive evaluation.

Seen from this perspective, the opening claim of this chapter—that classical ethical theory treats moral judgement as the outcome of structured deliberation—is not an empirical hypothesis but a *second-order commitment*. It reflects the normative aspiration that moral authority arises from principled reasoning. Yet the Morality Primer exposed a systematic tension between this commitment and the empirical reality of moral cognition: human agents rarely deliberate in the manner presupposed by rationalist models of ethics [15, 37]. Instead, their judgements emerge from perceptual salience, affective valuation, heuristics of social meaning, and rapid 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 judgements actually arise. Computing science faces this tension acutely, particularly in Machine Ethics, Social Signal Processing, and Affective Computing, where the challenge is to model behaviour that is empirically grounded yet normatively interpretable. Designing artificial agents requires avoiding both errors: treating first-order outputs as if they were principled commitments, and designing systems around abstract principles that human agents do not in practice instantiate.

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

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

Classical ethical theory often treats moral judgement as the conclusion of structured deliberation: a process guided by reasons, principles, and normatively defensible commitments. As discussed in Chapter 3, this picture is descriptively incomplete. Human moral behaviour rarely emerges from extended reflection; instead, it unfolds through rapid, affectively mediated evaluations shaped by perception, context, and embodied interaction [16, 31, 147, 17]. The distance between what agents *ought* to do, what they *report* doing, and what they *actually* do is substantial [220, 221]. Understanding moral action in practice—particularly in environments populated by artificial systems—requires integrating ethical theory with the empirical machinery of moral psychology [113, 222].

For computing science, this integration is indispensable. Artificial agents increasingly participate in human environments where their form, presence, and behaviour modulate attention, inference, and normative expectation. Research in *Social Signal Processing* [63] and *Affective Computing* [68] has shown that human social cognition is finely attuned to subtle cues—gaze, posture, micro-expressions, spatial orientation, and embodied co-presence. These cues structure the "interaction order" [?] that shapes how humans interpret intentions, assign agency, and evaluate normatively relevant behaviour. When synthetic entities enter this order, they perturb it—not by issuing commands, but by altering the informational and affective landscape in which human cognition operates [223, 224, 225].

The thesis therefore proceeds from two linked premises:

(1) *ethical behaviour cannot be understood without an accurate model of moral psychology*, and

(2) *moral psychology cannot be operationalised in computational settings 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 agents, affordances, and rapidly deployed intuitive processes [226, 227].

From this perspective, the central claim motivating the experimental work becomes clear: *moral behaviour is systematically sensitive to the structure of the immediate perceptual–social environment.* If moral cognition is dynamically shaped by intuitive appraisal, attentional salience, and affective resonance [18, 6], then even a silent, behaviourally neutral synthetic presence may modulate the trajectory from moral perception to moral action. The results developed later in the thesis confirm this hypothesis, demonstrating that robotic co-presence can attenuate prosocial donation even in the presence of a strong moral cue (the Watching-Eye stimulus) [1, 2, 5, 35].

When reframed through ethical theory, this empirical claim has deeper implications. Ethics, on contemporary accounts, is a *second-order discipline* [37, 212, 228]: it does not generate moral judgements, but seeks to analyse, justify, or critique them. It examines the structure of reasons, obligations, and values—not the psychological mechanisms that produce first-order moral appraisals [15]. Machine Ethics has historically blurred this distinction. By attempting to engineer "ethical agents" directly at the level of principles—rule-sets, deontic logics, utility functions—it presumes that moral behaviour can be derived from explicit normative propositions [52, 20, 23]. This presumption is both philosophically and empirically untenable. It treats the normative *grammar* of ethics as if it were the mechanistic *causality* of moral cognition [46].

The argument developed throughout this thesis challenges this assumption directly. If moral action is shaped primarily by perceptual salience, intuitive appraisal, affective resonance, and the dynamics of social attention—as the subsequent experimental results show [16, 17, 6]—then second-order normative structures cannot be treated as generative drivers of behaviour. They are interpretive and justificatory, *not computationally operative*. This reorientation motivates the notion of *Computational Morality*: before ethical frameworks can be embedded into artificial systems, we must understand the cognitive–affective machinery that underwrites human moral responsiveness [229, ?]. Classical Machine Ethics inverts this order; the empirical results of this thesis reinstate it.

The aim of this chapter, however, is narrower than a full reconstruction of moral philosophy. It does not adjudicate debates about moral realism, contractualism, utilitarianism, or virtue theory. Instead, it isolates the conceptual and mechanistic structures necessary for the thesis as a whole: how ethical theory presupposes psychological assumptions [113], how moral judgements are cognitively realised [31, 147], and why any computational account of ethical behaviour must be grounded in an empirically accurate model of moral cognition [17]. The goal is foundational rather than encyclopaedic: to establish the theoretical substrate that motivates, constrains, and ultimately validates the experimental investigation that follows.

The remainder of the chapter develops this integration along three axes. First, it introduces the principal ethical concepts—deontic, consequentialist, and virtue-theoretic—that define the normative landscape of moral evaluation [15, 37, 14]. Second, it examines the empirical architecture of moral cognition, with emphasis on intuitionist and dual-process models [16, 31, 17]. Third, it links these philosophical and psychological constructs to the computational disciplines that analyse social behaviour—Social Signal Processing, Affective Computing, and broader work in embodied AI [63, 68, 223].

By weaving these strands together, the chapter provides the normative and conceptual tools needed to understand why—and by what mechanism—synthetic presence can reshape the evaluative topology of human decision-making. This synthesis prepares the ground for the experimental investigation that follows, where robotic co-presence is used as a principled probe into the cognitive machinery through which moral cues acquire behavioural force.

7.3 Ethical Theory as Second-Order Analysis

If the opening sections of this chapter establish the transition from first-order moral cognition to second-order ethical reflection, the present task is to spell out the methodological consequences of this shift. The distinction is not merely terminological. It determines which claims aim to explain behaviour, which aim to justify it, and which are constrained by empirical evidence. Failure to keep these levels distinct has led to recurring conceptual confusions in both philosophical ethics and computational modelling [87, 89, 230, 25, 52, 145, 231, 54]. This section therefore clarifies what second-order ethical theory *is*, what it *explains*, and what it *cannot* plausibly do.

7.3.1 Ethical Reflection and the Second-Order Stance

First-order moral judgements arise from the cognitive–affective architecture examined in the Morality Primer. They are psychologically instantiated, context-sensitive, and behaviourally measurable. Their structure reflects the mechanisms analysed in Chapter 3: operations on perceptual salience, affective appraisal, intuitive heuristics, social meaning, and controlled modulation under conflict. These judgements are the *phenomena* that ethical theory seeks to interpret.

Second-order ethical theory is structurally different. It is reflexive rather than generative. It asks questions of justification rather than description: *What counts as a reason? What makes an obligation binding? What norms govern deliberation and responsibility?* These questions presuppose capacities for abstraction, generalisation, and rational evaluation that are not themselves the proximate causal mechanisms of moral behaviour [16, 31, 60, 232, 233, 227, 234]. Sidgwick's distinction in *The Methods of Ethics* between the psychology of moral sentiment and the "method" of determining right conduct [36, Book I] makes exactly this point. Lemos's treatment of epistemic justification similarly separates doxastic psychology from the normative assessment of belief [235]. The analogy is instructive: ethics stands to moral judgement as epistemology stands to belief-formation.

Viewed from this stance, second-order theory is not a set of procedural rules that moral agents execute. It is a reflective framework for articulating the standards by which judgements *ought* to be evaluated. Its success depends on conceptual clarity and justificatory coherence, not on behavioural predictiveness. Confusing this stance with the causal mechanisms of moral cognition risks treating normative categories as if they were psychological operators.

7.3.2 Levels of Abstraction and the Proper Location of Ethical Explanation

The distinction between first-order and second-order claims becomes sharper through Floridi's framework of *Levels of Abstraction* (LoA) [25, 197]. Every explanatory enterprise selects an LoA defined by its observables, its conceptual resolution, and the class of questions it can intelligibly answer. Ethical theory and moral cognition do not merely occupy different LoAs; they constitute *different explanatory kinds.*

At the **cognitive LoA**, the explananda are:

- perceptual salience and attentional capture;
- affective appraisal and embodied valuation;
- intuitive heuristics and rapid social inference;
- conflict monitoring and controlled modulation;
- the temporal dynamics of integrating these processes.

These are psychologically realised mechanisms with causal influence on behaviour. They are the variables the experimental chapters manipulate directly. *This is the LoA at which this thesis measures moral cognition.*

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

- principles of justification and admissible reasons,
- conceptions of duty, value, and obligation,
- normative standards of agency and responsibility.

These are not causal operators but interpretive and justificatory categories. They organise moral practice but do not generate its behaviour. Ethical theory therefore evaluates the grammar of moral reasons rather than the mechanisms of moral cognition.

Classical Machine Ethics collapsed these LoAs. By treating deontic principles, utility structures, or prima facie duties as mechanistic generators of behaviour, early systems implicitly assumed that normative concepts function at the same LoA as cognitive processes. This assumption fails on two fronts:

1. It misattributes causal status to normative constructs: duties and principles do not behave like salience gradients or affective appraisals.
2. It ignores empirical work showing that behaviour emerges from intuitive, affective, and situational mechanisms long before propositional reasoning is engaged.

From the perspective developed here, this is not merely incomplete—it is methodologically incoherent. It attempts to engineer behaviour by manipulating abstractions at a LoA that is *not behaviourally operative.*

The limitations of classical systems illustrate this point clearly. Top-down architectures such as Arkin's ethical governor [22], Anderson and Anderson's principlist models [236, 20], logic-based deontic programs [21, 237], consequentialist utility-systems [238], and virtue-theoretic computational frameworks [239, 240] all treated normative abstractions as if they were implementable causal rules. Floridi's LoA analysis makes explicit why this reduction cannot succeed: normative categories belong to a reflective LoA concerned with justification, while computational models operate at an implementational LoA concerned with mechanism. Conflating the two yields systems whose "moral" behaviour is an artefact of representational choices rather than genuine moral competence.

LoA discipline therefore becomes essential. Explanations of behaviour require the cognitive LoA; evaluations of reasons require the normative LoA. Neither reduces to the other. Yet they are not independent: normative evaluation presupposes a psychology capable of rendering moral salience operative, while psychological findings constrain the plausibility of normative theory.

This interdependence links this chapter to both Chapter 3 and the experimental analysis that follows. Chapter 3 established that the cognitive LoA is *topologically structured*: moral cognition unfolds within an evaluative field whose gradients depend on affective cues, attentional dynamics, and interpretive processes. Perturbations to this field—whether through altered salience, modified affective tone, or ambiguous social presence—can reshape behaviour even when normative commitments remain unchanged.

Seen through the LoA framework, the thesis's central research question can now be restated with 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 the notion of **evaluative topology**—introduced in Chapter 3 and developed throughout the thesis—provides precisely such a bridge.

7.3.3 Evaluative Topology as a Bridge Between Orders

The central challenge established thus far is not to collapse first-order moral cognition into second-order ethical theory, nor to treat normative principles as mechanistic generators of behaviour. Rather, the task is to articulate a structure that enables principled interaction between these orders without confusing their explanatory roles. *Evaluative topology*, introduced in the Morality Primer (Chapter 3) and developed throughout this thesis, provides precisely such a structure.

Evaluative topology is naturally situated within a long-standing tradition in computational cognitive science that models perception, valuation, and action as components of continuous dynamical systems rather than discrete symbolic modules. Moral psychology already supplies extensive evidence that moral judgement emerges from distributed interactions among perceptual salience, affective appraisal, attentional dynamics, and socially embedded interpretation. Models such as Haidt's social intuitionism and Greene's dual-process account capture moral appraisal as an interaction within a multi-dimensional affective and social field rather than as rule application [16, 31, 60]. Neurocognitive work extends this perspective: Nussbaum and Churchland both treat emotions as forms of evaluative perception with graded, vector-like organisation [83, 241]. Social Signal Processing research likewise conceptualises interpersonal evaluation as a shifting landscape of cues modulating behavioural trajectories in real time [67].

Against this background, evaluative topology provides a computationally meaningful formalisation. It treats the moral landscape as a dynamic field that structures the flow from perceptual input to action readiness. Instead of assuming that behaviour is produced by discrete maxims or fixed utility scores, evaluative topology models moral cognition as continuous transformations across a structured state-space. This aligns with dynamical-systems approaches that explain action selection through attractors, salience gradients, 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 ways in which 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 supplies a principled bridge between first-order cognition and second-order ethical theory. It mirrors the empirical architecture of human moral cognition—distributed, affectively grounded, context-responsive—while remaining compatible with the justificatory concerns of normative ethics. This enables descriptive and normative orders to interact without reduction: ethical theory specifies global constraints on evaluative structure; moral psychology identifies the mechanisms through which those structures are realised; and topology provides the medium in which they meet.

At its core, evaluative topology treats the moral landscape not as a set of isolated judgements or abstract principles, but as a *dynamic field* whose configuration determines the pathways from perception to action [16, 31, 241, 60, 83, 233]. Its explanatory primitives include:

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

Unlike classical ethical theories, which operate at a reflective and often idealised level [36, 242, 243, 80, 37], evaluative topology is sensitive to the real-time mechanisms through which moral cognition unfolds. And unlike purely mechanistic psychological models, which chart causal influences without normative content, topology captures the relational and counterfactual structure of moral appraisal: how behavioural trajectories *would* shift under alternative affective, attentional, or contextual configurations.

This leads to a three-part alignment essential for this thesis:

1. **Ethical theory** identifies which evaluative configurations *ought* to carry normative authority.
2. **Moral psychology** identifies which configurations *do* govern actual behaviour.
3. **Evaluative topology** identifies how these structures interact, diverge, and can be perturbed.

This tripartite structure yields both diagnostic and constructive insights. Diagnostically, it explains the failure of many classical Machine Ethics frameworks: they attempted to engineer behaviour by manipulating abstractions at a normative LoA while ignoring the topological organisation of the cognitive LoA that actually produces behaviour. Constructively, it provides a psychologically realistic substrate on which normative reflection can operate without reducing ethics to psychology or cognition to normativity.

Topological Consequences for Moral Perturbation. The Morality Primer established that moral behaviour emerges from traversal across a dynamically structured evaluative field. Within this framework, *perturbation* has a precise, measurable meaning: any alteration that changes the curvature or attractor structure of the field will shift the probability distribution over behavioural trajectories. This includes changes to salience, affective tone, 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 is the theoretical insight that shaped the experimental design in Chapter **??**: 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 suffices to attenuate prosocial action.

Interim Synthesis: Positioning the Argument. The conceptual machinery developed thus far establishes the structural conditions for the experimental work:

- First, moral judgement 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 governing the transition from moral perception to moral action.
- Fourth, this bridge is indispensable for understanding how synthetic agents perturb human moral behaviour.

With this scaffolding in place, we can now reconstruct the major normative traditions. The reconstruction is not a survey but a methodological necessity: each tradition identifies distinct loci of normativity, and these differences directly shape how the experimental attenuation should be interpreted. Without situating the empirical perturbation within a structured normative framework, one could describe *what* changed but not *what the change means*.

The next section therefore introduces deontic, consequentialist, and virtue-theoretic architectures through the combined lens of Levels of Abstraction and evaluative topology, preparing the conceptual ground for assessing the ethical significance of the perturbation demonstrated experimentally.

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

With the methodological apparatus now established, we can introduce the major normative frameworks that constitute the philosophical background against which the behavioural findings of this thesis must ultimately be interpreted. The purpose of this section 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 across the thesis [25, 242, 37, 80].

Two methodological constraints guide this reconstruction:

1. **Philosophical fidelity** — the theories must be represented in a manner consistent with their canonical formulations within moral philosophy [244, 245, 243, 36, 246, 247].
2. **Integrative compatibility** — the theories must be articulated in a way that allows principled interaction with the cognitive–affective and topological models of moral judgment introduced in Chapter 3 and developed through the Discussion [16, 31, 241, 233].

The aim, then, is not to catalogue doctrines, but to map the *structural logic of normativity* in a way that will later clarify the ethical significance of the empirical perturbations produced by synthetic presence.

7.4.1 The Three Dimensions of Normative Analysis

Normative theories differ not only in the moral claims they endorse, but in the *architecture of normativity* they assume [37, 80, 248]. To analyse them systematically, we distinguish three fundamental dimensions—each corresponding to a feature of evaluative topology and LoA structure.

1. **Source of normativity** — the origin of justificatory authority: rational agency (Kant [244]), human flourishing (Aristotle [245]), aggregated welfare (Mill, Sidgwick [243, 36]), affective sentiment (Hume, Smith [246, 247]), or interpersonal justification (Scanlon [37]).
2. **Mode of evaluation** — the features of action or character that determine moral relevance: maxims, outcomes, virtues, motives, relational duties, or context-specific particulars [66, 249, 242].
3. **Mechanism of action-guidance** — the process through which evaluation becomes behaviour: categorical imperatives, welfare optimisation, virtue-structured perception, affective resonance, or justificatory equilibrium [250, 80, 37].

These dimensions allow us to re-express classical theories as *evaluative topologies*—distinct structural configurations of the moral field:

- **Kantian ethics** imposes deontic invariants that carve the evaluative field into sharply bounded permissible and impermissible regions [244, 80].
- **Consequentialism** defines a gradient field over states of affairs: action flows along trajectories of maximal expected welfare [243, 36, 251].
- **Virtue ethics** defines dispositional attractors: stable patterns of moral sensitivity shaping perception and evaluative attention [245, 249, 250].
- **Sentimentalism** defines affectively weighted pathways through which moral appraisal propagates [246, 247, 252].
- **Contractualism** defines justificatory equilibria: a topology structured by mutual recognisability of claims [37, 242].
- **Particularism** rejects fixed topologies: moral relevance emerges from local patterns of salience and relation [66].

This analytic frame yields a common representational language in which ethical theory and moral psychology can be jointly expressed. Theories that diverge substantially in content become comparable in structural terms—how they configure the evaluative field, where they locate normative constraints, and how they model the transition from judgment to action [250, 80, 253].

7.4.2 Why This Framework Matters for the Experimental Chapter

This normative topology is not abstract ornamentation; it is the conceptual infrastructure that allows the experiment to be interpreted. The behavioural question—whether robotic co-presence attenuates prosocial donation—cannot be evaluated ethically without a framework that explains *how* moral cues acquire force in the first place [16, 31, 241].

Three structural claims follow immediately from the reconstruction above:

1. **Moral action depends on the configuration of the evaluative field.** Normative theories differ in source, mode, and guidance, but all assume that moral behaviour emerges from structured evaluative relations, not arbitrary choice [245, 80, 37].

2. **Synthetic presence modulates this field by perturbing salience, attention, and affective resonance.** A humanoid robot does not supply new reasons; it alters the environment within which reasons become behaviourally operative [67, 6, 5, 35].
3. **Normative theories must therefore be expressed within the joint framework of LoA and evaluative topology in order to interpret the empirical perturbation coherently.**

This is the philosophical function of the present section: to establish the normative coordinates that will allow the experimental results—introduced later in the thesis—to be understood not merely as statistical differences, but as shifts in the moral significance of an action within a structured evaluative landscape [253, 80, 37].

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 human agents. These reconstructions will serve as the interpretive foundation for assessing how, and why, synthetic presence can reshape the moral field in the experiment.

7.5 Deontological Structures: The Architecture of Practical Reason

The methodological framework established above motivates a disciplined reconstruction of deontological ethics through the joint lens of Levels of Abstraction (LoA) and evaluative topology. The aim is not to treat deontology as a psychological model—indeed, it is explicitly *not* one—but to articulate how deontic normativity can be represented as a structural component of the evaluative field within which moral agents operate. This reconstruction preserves the philosophical identity of deontological theory while rendering it compatible with the cognitive–affective and topological architecture developed across the thesis.

Three constraints guide the reconstruction:

1. **Philosophical fidelity**: The core commitments that distinguish deontology must remain intact.
2. **LoA discipline**: Deontic principles cannot be treated as psychological mechanisms or behaviour-generating algorithms.
3. **Topological embedding**: Duties must be expressed as structural constraints within the evaluative field, not as direct causes of action.

Within this framework, deontology identifies *invariant structures* in the moral field: boundaries of permissibility and prohibition that constrain evaluative trajectories without functioning as generative cognitive operators. These invariants occupy a reflective LoA and serve as standards of justification, not as engines of behaviour.

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

For Kant, moral authority arises from the structure of rational agency itself. The categorical imperative offers a formal test of maxims—whether a maxim could be willed as a universal law—not a psychological process for generating behaviour [244, 80, ?]. Its role is to define the *conditions of justificatory coherence*, situated at a higher LoA than the cognitive mechanisms analysed in Chapter 3. The categorical imperative belongs to the space of reflective evaluation, not to the causal substrate of intuitive moral appraisal.

This distinction is essential to the present thesis. Classical Machine Ethics frequently misinterpreted universalisability tests as if they were procedural decision rules—algorithmic operators that could be executed at run time [236, 20, 21, 254, 237, 22]. But Kant never proposed that deontic evaluation functions as a mechanistic generator of moral action.[1] Treating such tests as computational procedures constitutes the very LoA confusion diagnosed earlier: it collapses reflective justification into first-order cognition.

A brief survey of Classical Machine Ethics illustrates this confusion clearly. Anderson and Anderson's principlist architectures computationalised prima facie duties as weighted decision rules [236, 20]; Bringsjord and colleagues embedded deontic obligations into the cognitive event calculus [21, 254]; Ganascia formalised ethical constraints as logical conditions governing action permissibility [237]; and Arkin's "ethical governor" implemented deontological rules derived from Just War Theory as real-time filters on autonomous behaviour [22]. In each case, duties intended as reflective constraints were treated as if they were causal action-selection mechanisms.

As Moor and Coeckelbergh emphasise, this is a fundamental mistake of abstraction: ethical principles belong to a normative LoA, whereas cognitive processes and computational models operate at a mechanistic LoA [255, 54]. Conflating these levels does not produce ethically competent machines; it produces systems that mechanically enforce the representational choices of their designers.

7.5.2 Deontic Invariants as Topological Constraints

Reconstructed through the evaluative–topological lens, deontological duties are best understood as *structural constraints* that shape the moral field without functioning as its generative forces. Instead of treating the categorical imperative as a behavioural algorithm, we interpret deontic norms as imposing *invariant boundaries* on permissible trajectories in the evaluative manifold. Formally, a deontic constraint can be expressed as a region of the field $\mathcal{E}$ that action trajectories cannot cross without violating justificatory coherence.

This topological rendering preserves the normative role of deontology while integrating it with the empirical architecture of moral cognition:

[1] See [?] and [15] for detailed discussion of the reflective, non-psychological status of the categorical imperative.

- At the **cognitive LoA**, intuitive appraisal and affective resonance drive the formation of evaluative gradients.
- At the **topological LoA**, deontic norms impose structural boundaries that constrain the space of evaluatively permissible outcomes.
- At the **normative LoA**, reflective justification assesses whether a trajectory is consistent with universalizable maxims.

These levels remain distinct, yet their interaction can now be modelled without conflation. Deontic invariants do not guide moment-to-moment appraisal, but they structure the higher-level evaluative landscape in which such appraisal takes place.

7.5.3 Relevance to Synthetic Perturbation

This reconstruction equips us to interpret the experimental findings later in the thesis. If deontic norms function as structural constraints on the evaluative field, then synthetic presence—by altering salience, attention, and the perceived sociality of the environment—can modify the *access* agents have to those constraints without altering the constraints themselves.

From a deontological perspective, then, attenuation under robotic co-presence is not a violation of duty. It is a deformation of the cognitive–affective substrate through which agents track deontic salience. The duty remains; the *grip* of the duty is weakened because the evaluative conditions under which it becomes behaviourally operative have been perturbed.

This interpretation preserves the philosophical integrity of deontology while situating it precisely within the cognitive–topological framework of the thesis. Deontic normativity thus provides one dimension of the interpretive foundation necessary for understanding how synthetic presence reshapes the moral field.

7.5.4 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. In the present framework, these evaluative commitments introduce a characteristic structure into the moral field. Their core features can be expressed topologically:

- **Invariance**: duties bind independently of context, affective state, or anticipated outcome.
- **Non-gradience**: obligations often define discrete boundaries—permissible vs. impermissible—rather than continuous slopes.
- **Symmetry**: the universal law test imposes interpersonal consistency constraints across agents.
- **Role-relativity**: some duties apply only under specific relational or social conditions (e.g. fidelity, respect, special obligations).

Viewed through evaluative topology, these features correspond to *hard constraints* within the evaluative landscape. They do not shape the gradients that drive moment-to-moment appraisal; instead, they partition the field into admissible and inadmissible regions. Deontological normativity thus defines the *regulatory geometry* within which cognitive–affective trajectories unfold.

7.5.5 Action-Guidance: How Normative Constraints Influence Behaviour

A central challenge now arises. If deontological principles do not describe psychological processes, how do they guide action?

The answer, consistent with LoA discipline, is that their influence operates *indirectly* and at distinct temporal and explanatory scales:

1. **At the cognitive LoA** (real-time appraisal), deontic principles do not produce behaviour. Behaviour emerges from the integration of perceptual salience, affective valuation, intuitive appraisal, and controlled modulation—processes analysed empirically in Chapter 3.
2. **At the normative LoA** (reflective endorsement), deontological principles determine which trajectories can be justified as consistent with rational agency. They also shape the long-term development of moral character by influencing attention, affect, and self-regulation through training, habituation, and self-constitution.

In this long-term sense, internalised deontic commitments function as a form of *normative scaffolding*. Over time they:

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

Thus, deontology does not operate the machinery of moral cognition. Instead, it calibrates aspects of that machinery across development and reflective practice. It provides the structural frame against which agents regulate their evaluative postures.

7.5.6 Deontological Normativity as Topological Invariance

This perspective allows the central insight of the reconstruction to be stated precisely. Within a topological model of moral cognition, deontological ethics identifies *non-negotiable invariants*: fixed points or boundaries that preserve the structural integrity of the moral field.

These invariants:

- partition the evaluative manifold into permissible and impermissible zones;
- resist deformation by short-term changes in affect, context, or incentives;

- stabilise behavioural tendencies by constraining rational endorsement over time;
- provide the reflective standpoint from which agents evaluate the legitimacy of their conduct.

Accordingly, the categorical imperative is not an algorithmic decision rule but a *topological constraint*: a principle that ensures global coherence of evaluative structure rather than ad hoc, context-bound optimisation.

7.5.7 Why Deontology Matters for the Experimental Logic

This reconstruction is essential for integrating the experimental findings into a normative framework. The experiment does not merely identify behavioural differences; it raises the question of their *moral significance.* Deontology supplies one dimension of the interpretive structure required to answer that question.

Before stating the connection explicitly, one clarification is needed. The experiment employs a widely studied social–moral prime: the "Watching-Eye" cue. As detailed in Chapter **??**, such cues increase accountability, evoke reciprocity norms, and prime compliance with expectations of beneficence—even though they involve no real observer. They thus operate on precisely the evaluative sensitivities that internalised deontic structures help regulate.

Given this, the relevance of deontology to the experimental logic can be articulated through three claims:

1. **Perturbations of prosocial behaviour must be normatively classified.** If synthetic presence reduces donation, we must ask whether the shift remains within the deontically permissible region or whether it signals a distortion in the agent's sensitivity to obligation.
2. **The Watching-Eye cue implicitly invokes deontic expectations.** It activates norms of accountability, respect, and reciprocity. A reduction in prosociality under this cue suggests that the synthetic agent may interfere with the mechanisms through which deontic salience is apprehended.
3. **Deontology provides the vocabulary for distinguishing moral distortion from benign modulation.** Not all behavioural shifts are ethically significant; deontic analysis helps determine whether attenuation constitutes weakened duty-tracking rather than mere affective dampening.

This is the point at which the present thesis departs most sharply from monolithic Machine Ethics. Classical approaches attempted to encode deontic rules as behavioural algorithms. But the empirical findings in later chapters show why this strategy is misguided: deontic norms do not generate behaviour, and behavioural perturbations cannot be interpreted solely as rule deviations. Instead, synthetic presence acts on the evaluative field *upstream* of duty, altering the conditions under which deontic invariants become behaviourally operative.

With deontology reconstructed as a system of topological constraints rather than computational rules, we can now proceed to consequentialism. There, normativity takes the form of gradient fields over outcomes—structures that interact with the evaluative machinery of moral cognition in different, but equally revealing, ways.

7.5.8 Conceptual Note: Gradient Fields in Consequentialist Topology

Within the evaluative–topological framework developed in this thesis, a *gradient field* denotes a structured moral landscape in which each possible action–outcome configuration is associated with a scalar value—typically representing expected welfare, utility, or outcome-based moral worth. Conceptually, a gradient field assigns to each point in this space a direction of steepest ascent: the direction in which a marginal shift would produce the greatest increase in expected value. Classical utilitarian reasoning implicitly presupposes such a structure when it assesses actions by their contribution to overall welfare [**?**, 243, 36]. Here, the notion is used in a non-formal but philosophically precise sense: as a way of modelling how consequentialist evaluation imposes directional structure on the moral field, where moral improvement corresponds to movement toward higher expected value.

A consequentialist gradient field has three defining properties:

1. **Scalar valuation**: each point in the evaluative manifold has a determinable (actual or expected) value, enabling continuous comparison along a single welfare dimension.
2. **Directional guidance**: the moral significance of an option lies in its orientation relative to the gradient; actions are preferable to the extent that they align with the direction of steepest welfare ascent.
3. **Empirical sensitivity**: because value depends on expected outcomes, the structure of the field varies with beliefs, evidence, uncertainty, and situational detail.

Crucially, in this reconstruction gradient fields do *not* function as psychological mechanisms. Agents do not compute welfare gradients when acting, nor do they evaluate global states of the world through analytic integration. Consequentialist structures operate at the *normative Level of Abstraction*: they specify how actions are *justified* under reflective endorsement, not how they are generated in real-time cognition. Sidgwick's distinction between the "point of view of the universe" and the psychology of everyday decision-making is an early articulation of this separation [36, Book IV].

Interaction with the Evaluative Machinery of Moral Cognition. Although gradient fields belong to the normative LoA, they interact indirectly with the empirical machinery of moral cognition introduced in Chapter 3. Four forms of interaction are especially relevant:

1. **Salience modulation**. Anticipated outcomes influence which parts of a situation become perceptually salient. Potential harm, benefit, or risk amplifies attention and reshapes local evaluative configuration before explicit reasoning is engaged.
2. **Affective valuation**. Affective systems track outcome-related information with strong valence, effectively providing local approximations of the gra-

dient. Positive and negative affect bias intuitive appraisal toward or away from certain actions in ways that loosely track expected value.

3. **Heuristic internalisation**. Over time, agents extract outcome-sensitive heuristics—"help when it is easy", "avoid imposing harm"—that are computationally tractable proxies for gradient following. These heuristics allow the cognitive system to approximate consequentialist structure without computing it.
4. **Deliberative correction**. When intuitive and affective processes conflict or when the situation is ambiguous, controlled processes may approximate explicit comparisons of expected harm or benefit. This engages the gradient field at a coarse resolution, albeit with substantial computational limits.

A fifth mode is essential for the present thesis:

5. **Perturbation sensitivity**. Because valuations depend on perceived outcomes, any perturbation to perception, attention, or social meaning—such as the introduction of a humanoid robot—can reshape the agent's *perceived* gradient field. Consequentialist structures are thus especially sensitive to environmental distortions of the kind tested experimentally.

Evaluative topology makes these interactions explicit. It models behaviour not as the execution of explicit calculations, but as movement through a dynamically shaped field whose gradients are only indirectly approximated by affective and attentional processes.

This integration is necessary for the thesis as a whole. It renders consequentialism compatible with the empirical finding that moral behaviour is modulated by subtle shifts in the perceptual–social environment. It also clarifies how the experimentally observed attenuation of prosocial donation under synthetic presence can be interpreted: as a local distortion of the gradient field that normally favours prosocial conduct.

7.6 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 consequentialism. Here the architecture differs along every structural dimension. Where deontology imposes *fixed boundaries* in the evaluative field, consequentialism supplies *value gradients*. Where deontology locates normativity in the form of maxims, consequentialism locates it in outcome structure. And where deontology articulates duties, consequentialism articulates trajectories across possible states of the world.

As in the preceding section, the aim is not historical analysis but conceptual reconstruction. The goal is to articulate consequentialist normativity in a way that respects LoA discipline and integrates with the evaluative–topological account of moral cognition developed earlier. This reconstruction also prepares a normative lens through which the experimental attenuation effect can later be interpreted.

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

Classical utilitarianism grounds moral authority in the promotion of welfare. Bentham's felicific calculus [?], Mill's qualitative distinctions [243], and Sidgwick's systematic treatment of impartiality [36] converge on the view that what ultimately matters is the value of outcomes, aggregated across persons. On this view, an action is right insofar as it maximises (or sufficiently promotes) overall good.

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

- evaluating and comparing outcomes,
- aggregating welfare or value across individuals,
- and justifying action from an impartial standpoint.

These commitments are not descriptive claims about the mechanisms of moral cognition. Sidgwick is explicit that the deliberative standpoint of consequentialist justification is distinct from ordinary motivation. Consequentialism thus supplies a criterion of rightness, not a psychological procedure.

This point is crucial for avoiding the LoA confusion characteristic of classical Machine Ethics. Outcome-based formalisms—utility functions, reward optimisers, expected-utility maximisers—are often treated as if they were *surrogates* for moral cognition itself. But these belong to different explanatory orders: normative structure at the reflective LoA and cognitive–affective processes at the psychological LoA (Chapter 3). Any mapping between them must be justified, not assumed.

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

Consequentialism evaluates actions by the value of their actual or expected outcomes. Unlike deontological theories, which generate categorical constraints, consequentialism is *scalar*: options can be morally preferable to varying degrees. This scalar structure has a natural topological representation.

In the evaluative–topological model, a consequentialist landscape exhibits:

- **Gradience**: moral evaluation varies continuously with expected value.
- **Optimisation structure**: right action corresponds to local or global maxima on the welfare landscape.
- **Context-dependence**: the shape of the field depends on empirical facts about consequences.
- **Impartiality**: welfare contributions have equal evaluative standing across persons.

Because of these features, consequentialism lends itself readily to computational formulation: utility functions, reward structures, and optimisation routines approximate the mathematics of value gradients. This explains its prominence in reinforcement-learning–based approaches to Machine Ethics.

Yet computational elegance must not be mistaken for cognitive realism. Human moral cognition does not perform explicit optimisation; it relies on heuristic, affective, and context-responsive mechanisms that only loosely approximate consequentialist ideals [227, 31, 16, ?]. Treating human agents as literal expected-utility maximisers is another instance of LoA confusion.

7.6.3 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 through *indirect modulation* of the evaluative topology rather than through explicit computation.

At the normative LoA, consequentialism states:

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

At the cognitive LoA, by contrast, behaviour emerges from the interaction of intuitive appraisal, affective resonance, social cues, and controlled modulation. Consequentialist considerations shape this machinery only *over time*, through pathways such as:

- **Dispositional shaping**: moral education increases sensitivity to outcomes and harm, thereby steepening certain evaluative gradients.
- **Outcome-sensitive heuristics**: agents internalise tractable rules (e.g. "help when it costs little") that loosely approximate expected-value comparisons.
- **Attentional modulation**: anticipated benefits or harms alter what becomes salient and thus influence intuitive appraisal.
- **Deliberative correction**: when intuitive responses conflict, deliberation may reweight options in favour of outcome-based considerations.

Topologically, consequentialism does not *run* the cognitive system. Instead, it shapes the evaluative field by reorienting trajectories and adjusting the relative steepness of welfare-relevant gradients.

7.6.4 Consequentialist Topology: Moral Action as Gradient Following

Within the topological framework of this thesis, the core consequentialist idea can be expressed succinctly: moral action corresponds to (approximate) *gradient following* in a welfare-defined landscape. Behaviour counts as morally preferable when it moves "uphill" along these value gradients.

This yields several structural implications:

1. **Smoothness**: unlike deontic boundaries, consequentialist landscapes permit continuous gradations of moral improvement.

2. **Directionality**: the moral relevance of an action depends on its orientation relative to welfare ascent.
3. **Trade-offs**: multi-dimensional outcomes (helping one party, imposing small burdens on another) are represented as interacting gradients.
4. **Perturbation sensitivity**: because evaluation depends on expected consequences, shifts in salience, attention, or perceived social meaning can locally distort the gradient.

This final feature is directly relevant to the experiment: if synthetic presence alters the perceived consequences of donating—by changing the social meaning of helping or by absorbing attentional and affective resources—the value gradient favouring prosocial action can be flattened.

7.6.5 Why Consequentialism Matters for the Experimental Logic

Consequentialism provides one indispensable dimension for interpreting the attenuation observed in the experimental results. Prosocial donation is simultaneously:

- a *behavioural output* of the cognitive architecture, and
- a *welfare-relevant action* whose outcomes can be straightforwardly ordered.

Within this dual frame, the Watching-Eye prime and synthetic presence can be understood as modifying the agent's *perceived consequence structure*.

1. **The Watching-Eye cue steepens the prosocial gradient.** As reviewed in Chapter **??**, visual cues of being observed increase the perceived reputational or social value of helping. In topological terms, the gradient from "keep" to "donate" becomes steeper.
2. **Synthetic presence can flatten or redirect this gradient.** The humanoid robot constitutes an ambiguous social agent whose presence may blunt or partially occlude the evaluative pathway activated by the Watching-Eye cue. If attention shifts toward the robot, or if the robot is not integrated into the relevant social-evaluative schema, the perceived payoff of donating may weaken.
3. **Consequentialism provides a normative reading of this shift.** From a consequentialist perspective, attenuation signals that the agent's welfare-related field has been deformed: donating no longer appears sufficiently beneficial—socially, affectively, or interpersonally—to overcome competing evaluative forces.

Importantly, nothing in this reconstruction treats consequentialism as a blueprint for machine implementation. Unlike classical Machine Ethics approaches that equate "ethical design" with encoding explicit utility functions, consequentialism here functions as a *normative lens*: a structured perspective on how synthetic presence perturbs the evaluative topology that normally favours prosocial behaviour.

The next section turns to virtue ethics, which locates normativity not in constraints or outcomes but in cultivated dispositions and perceptual sensitivities.

This framework will illuminate a further dimension of the evaluative topology: how character, habituation, and moral perception shape susceptibility to synthetic perturbation.

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

Deontological invariants and consequentialist gradients capture two dimensions of the evaluative field, but they remain incomplete without an account of the *agent* who navigates that field. Virtue ethics—classically Aristotelian [13] and developed in modern neo-Aristotelian and psychological accounts [14, 69, 81]—locates normativity not in rules or outcomes but in the *perceptual and dispositional architecture* of the moral agent. This makes virtue theory particularly well-suited to the present thesis, where experimentally observed attenuation varies systematically across latent trait ecologies (Chapter **??**).

Broken Link

Our task is therefore to reconstruct virtue ethics in a form compatible with the evaluative-topological model and LoA discipline. This reconstruction must:

1. preserve the philosophical distinctiveness of virtue theory as an account of normativity grounded in moral perception and stable character,
2. express dispositional structure in topological terms—as curvature and attractor shape in the evaluative field,
3. and connect directly to the empirical pattern of cluster-dependent susceptibility under synthetic perturbation.

Within these constraints, virtue ethics becomes a theory of *moral sensitivity as a topologically structured, personality-dependent field*, shaped by long-term habituation and modulated by local perturbations such as robotic co-presence.

7.7.1 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. Virtues are not propositional rules but *stable perceptual–evaluative dispositions*: they structure which features of a situation stand out as salient, how those features are weighted, and which actions appear fitting or required [250, 14]. Aristotle's *phronesis* captures this idea as *perceptual attunement*: the capacity to discern morally relevant particulars and respond appropriately [13].

This maps directly onto the evaluative-topological framework. A virtuous agent's evaluative field contains:

- **stable attractors** corresponding to benevolence, honesty, fairness, and other prosocial dispositions;
- **well-shaped gradients** that reliably direct appraisal toward morally appropriate trajectories;

- **robustness under perturbation**, where minor contextual shifts do not destabilise moral sensitivity.

By contrast, deficiencies in character manifest as distortions in the field: shallow attractors, flattened gradients, or unstable transitions. Thus, virtue ethics provides a natural bridge between normative theory and personality-structured cognitive architecture.

7.7.2 Mode of Evaluation: Dispositions as Topological Structure

Virtue ethics evaluates actions as *expressive of character*, not merely as discrete events. The morally relevant unit is the dispositional pattern through which the agent perceives and structures the situation. This is exactly the level at which the experiment reveals systematic variation.

(i) Mathematical and Topological Interpretation

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

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

where k indexes latent psychological traits (e.g. agreeableness, empathy, conscientiousness). Chapter **??** showed that participants form coherent clusters $C_1, \ldots, C_m$ in this space.

In virtue-theoretic terms, we can model the mapping

$$\mathcal{T} : \mathbb{R}^k \to \mathcal{F},$$

where $\mathcal{F}$ is the space of evaluative fields. Under this mapping:

- high-empathy / warm–sociable clusters exhibit deeper prosocial attractors and sharper gradients toward helping,
- analytical–structured clusters exhibit more stable but less affectively steep topologies,
- emotionally reactive clusters exhibit shallow, volatile attractor basins.

This aligns with empirical personality research linking empathic concern, agreeableness, and prosocial orientation to enhanced moral sensitivity [256, **?**, 66]. In the experiment, these dispositional field differences predicted differential susceptibility to perturbation under synthetic presence—precisely what a virtue-theoretic model would anticipate.

(ii) Connection to Moral Psychology

Contemporary moral psychology emphasises that moral responsiveness depends on stable trait configurations. Research on moral foundations [256], character-based accounts [**?**], and perceptualist theories of moral sensitivity [66, 250] all converge on the idea that moral judgment is a function of habituated perception.

The experimental data vindicate this insight. The humanoid robot did not uniformly attenuate behaviour; instead, attenuation varied by cluster:

- strongest in the Prosocial–Empathic ecology (where affective gradients are steep and easily perturbed),
- weak but present in the Analytical–Structured ecology (where action is driven by stability rather than resonance),
- negligible in the Emotionally Reactive ecology (where gradients are shallow and noise-dominated).

This pattern is exactly what virtue-theoretic topology predicts: *where the field is most morally sensitive, it is most susceptible to perturbation.* The experiment therefore provides an empirical instantiation of a core virtue-theoretic claim: that character structure determines not only moral dispositions but the *topology of susceptibility* to environmental modulation.

7.7.3 Action-Guidance Mechanism: Habituation, Stability, and Situated Sensitivity

Virtue ethics explains action not by invoking explicit rules or outcome calculations but through the *habituated patterns of salience, affect, and response* characteristic of a well-formed agent. This lines up directly with the dual-process architecture established in Chapter 3:

- intuitive, first-pass appraisals are shaped by long-term habituation into affective–perceptual sensitivities;
- controlled processes integrate commitments, identities, and reflective self-conceptions that stabilise these sensitivities over time;
- behavioural output reflects the depth or fragility of dispositional attractors.

BROKEN
LINKS

Topologically, virtues correspond to *deep, well-curved attractor basins* resistant to perturbation; deficiencies correspond to *shallow, volatile, or weakly integrated attractors.* This resonates with computational models of habit formation [?] and empirical accounts of moral perception as a learned sensitivity [?].

Importantly for this thesis, the clusters identified in Chapter ?? instantiate precisely this kind of dispositional architecture: warm–prosocial participants exhibit steep affective gradients; analytical profiles show stable but less affective curvature; reactive profiles show shallow, noise-dominated dynamics.

7.7.4 Virtue-Theoretic Topology: Stability, Curvature, and Susceptibility to Perturbation

Within the evaluative-topological model, virtue ethics can be expressed in dynamical-systems terms:

$$\dot{x} = f(x; \beta_C),$$

where x is the agent's state in evaluative space and β_C parametrises dispositional curvature. Robotic co-presence introduces a perturbation

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

with $\mathscr{R}$ denoting synthetic presence.

This formalism directly reflects the empirical pattern:

- in the Prosocial–Empathic ecology, perturbation δf significantly shifts trajectories away from the prosocial basin, producing the strongest attenuation;
- in the Analytical–Structured ecology, attractor curvature is sufficient to absorb most of the perturbation, yielding only modest displacement;
- in the Emotionally Reactive ecology, shallow, unstable attractors produce minimal directional change—behaviour is already close to noise-level variation.

This mapping from dispositional structure to perturbation susceptibility is precisely the kind of structure virtue theory predicts: character determines *how* moral affordances are perceived and how perturbations are absorbed or amplified.

7.7.5 Why Virtue Ethics Matters for the Experimental Logic

Virtue ethics is indispensable for interpreting the experimental results, for three reasons that integrate tightly with the Discussion chapter and set up the thesis conclusion.

1. Latent Trait Modulation: Explaining Cluster Differences The experiment demonstrates that robotic co-presence induces a *field-level* perturbation whose *impact* depends on dispositional topology. Virtue theory provides the conceptual vocabulary for this dependency. It explains why prosocial action is fragile in agents with shallow affective attractors, why highly empathic profiles show strong attenuation, and why analytical profiles exhibit relative resistance. The experiment therefore reveals a virtue-theoretic phenomenon: moral sensitivity is intrinsically *trait-dependent*.

2. Character as the Medium of Moral Topology The mapping

$$\beta_C \mapsto \mathcal{T}(\beta_C)$$

shows that moral responsiveness is a function of trait geometry. Character shapes the curvature of the evaluative manifold, determining which cues stand out as morally salient and how the Watching-Eye prime interacts with background dispositions. Synthetic presence perturbs this trait-conditioned topology, yielding precisely the cluster-conditioned attenuation patterns identified earlier.

3. Machine Ethics Omits Dispositional Structure Entirely Classical Machine Ethics contains no representation of habituation, perceptual attunement, or trait-level topology. It models moral behaviour as rule-execution or utility optimisation, ignoring the dispositional substrate that governs real moral sensitivity. This makes it structurally incapable of predicting the experimental pattern: *the strongest attenuation occurs precisely where the evaluative gradients are steepest—where moral sensitivity is highest.*

This result is unintelligible on rule-based or utility-based models but follows naturally from a virtue-theoretic account of character topology.

In sum, virtue ethics interprets the experimental findings as demonstrating that synthetic agents perturb moral action by modulating the *dispositional geometry* through which moral salience is processed. Deontology contributes boundary structure, consequentialism contributes gradient structure, but virtue ethics contributes the *curvature of the evaluative manifold*: the habituated topology that determines how agents absorb, refract, or amplify perturbations.

This sets the stage for the final normative lenses—sentimentalism, contractualism, and particularism—which illuminate additional dimensions of how synthetic presence reshapes the evaluative field investigated experimentally.

7.8 Integrated Ethical Interpretation of the Experimental Results

With deontology, consequentialism, and virtue ethics reconstructed through the discipline of Levels of Abstraction and embedded within the evaluative–topological architecture developed in this thesis, we can now articulate their joint significance for the experimental findings. The aim is not to allocate explanatory priority but to show why a multi-framework normative analysis is *required* if the behavioural perturbation induced by synthetic presence is to be ethically intelligible.

1. Deontology: Invariant Structure and the Integrity of Moral Expectation

On the deontological reconstruction, duties function as *structural invariants* within the evaluative field. The Watching-Eye cue (see Chapter **??**) implicitly activates precisely these invariants: expectations of accountability, reciprocity, and fairness.

When donation decreases in the Robot condition, the relevant normative question is not whether participants "broke rules" but whether synthetic presence *disrupted sensitivity* to these invariant structures:

- If the robot attenuates uptake of deontic salience, the perturbation carries ethical significance beyond preference change.
- Because all explicit cues remain constant across conditions, any weakening of accountability sensitivity isolates $\mathscr{R}$ as a potential interference with deontic perception.
- Deontology therefore provides the vocabulary to distinguish superficial behavioural modulation from a deeper deformation in the agent's grasp of duty.

Thus, the deontological reading aligns with the empirical finding of uniform attenuation: synthetic presence does not introduce new norms; it suppresses the felt relevance of existing ones.

2. Consequentialism: Gradient Deformation and the Perceived Structure of Outcomes

From a consequentialist perspective, moral orientation depends on the perceived gradient of expected value. Watching-Eye cues steepen this gradient by increasing the anticipated social or reputational payoff of prosocial action.

Synthetic presence perturbs this structure in three ways:

1. by introducing an ambiguous observer whose evaluative stance is unclear, flattening outcome expectations;
2. by competing with or overshadowing the reputational signal generated by the Watching-Eye stimulus;
3. by transforming a dyadic human–target context into a triadic social configuration with uncertain evaluative implications.

In topological terms, $\mathscr{R}$ deforms the gradient landscape surrounding donation. The attenuation effect thus fits naturally within the consequentialist lens: prosocial movement becomes less strongly favoured because the perceived payoff slope has been locally flattened.

3. Virtue Ethics: Dispositional Curvature and Cluster-Dependent Susceptibility

Virtue ethics provides the most direct connection between normative theory and the empirical structure of the experiment. On the virtue-theoretic reconstruction, moral responsiveness depends on the agent's *dispositional curvature*: the depth, stability, and integration of their evaluative attractors.

Virtue-Ethical Interpretation of Latent Ecologies

Cluster analyses (Chapter **??**) revealed distinct evaluative ecologies:

- **Prosocial–Empathic**: steep, affectively rich attractors; strong Watching-Eye response; moderate attenuation under $\mathscr{R}$.
- **Emotionally Reactive / Low-Structure**: shallow, unstable attractors; greatest susceptibility to perturbation.
- **Analytical–Structured**: stable but less affective curvature; small but systematic displacement when interpretive coherence is disrupted.

These patterns are *structurally predicted* by the virtue-theoretic framework:

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

where $f(x;\beta_C)$ represents each ecology's dispositional dynamics and $\delta f(x;\mathscr{R})$ the perturbation induced by synthetic presence.

Critically, δf is not uniform. Its sign and magnitude depend on the curvature encoded by β_C:

- for Prosocial–Empathic agents, δf weakens empathic gradients;

- for Reactive agents, δf amplifies existing volatility;
- for Analytical agents, δf disrupts interpretive coherence rather than affective force.

Thus, virtue ethics explains the *cluster-dependent pattern* of attenuation: synthetic presence interacts with dispositional topology, not with explicit rules or outcome computation.

4. What Classical Machine Ethics Misses

This integrated reading exposes a core limitation of classical Machine Ethics:

1. Treating deontic principles as behavioural algorithms misidentifies their LoA and cannot account for perturbations in deontic uptake.
2. Treating utilities as generative of moral cognition ignores the role of salience and affect in shaping perceived gradients.
3. Omitting dispositional topology leaves no framework for predicting cluster-dependent deformation or for understanding why the strongest attenuation occurs where empathic gradients are steepest.

Machine Ethics repeatedly commits the LoA confusion: it treats normative abstractions as if they were psychological operators. The experiment demonstrates that moral behaviour emerges instead from field-level dynamics that no monolithic framework can capture.

5. Concluding Perspective: Why a Multi-Framework Interpretation Is Necessary

The three reconstructed frameworks converge on a single insight: **synthetic presence reshapes the evaluative field through which moral salience becomes action**. Each theory captures a different dimension of this deformation:

- deontology identifies disruptions to sensitivity toward invariant expectations;
- consequentialism identifies gradient flattening in perceived outcomes;
- virtue ethics identifies dispositional curvature as the mediator of susceptibility.

The experiment therefore reveals not only that robots alter behaviour, but *how* they do so: by deforming the topological substrate that links perception to moral action. This integrated interpretation provides the normative scaffolding for the sentimentalist analysis that follows, in which affective vector fields become central to explaining the immediate, pre-reflective dynamics of the perturbation.

7.9 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology

Having reconstructed deontology as topological invariance and consequentialism as value-gradient optimisation, we now turn to the normative framework most directly implicated in the experimental results: *sentimentalism*. In the sentimentalist tradition—Hume, Smith, and contemporary affect-based theorists—moral evaluation originates in *patterns of affective resonance* [246, 257, 70, 258]. Nodes of moral significance are detected not through principles or calculations but through the affective forces that structure our perceptual–social encounter with others.

Within the evaluative–topological model, sentimentalism corresponds to an **affective vector field**:

$$\mathbf{A}(x) : \mathcal{X} \to \mathbb{R}^n,$$

where $\mathcal{X}$ is the space of perceived states and $\mathbf{A}(x)$ encodes the direction and magnitude of empathic pull, aversive push, compassion, warmth, or distress.

This is not metaphorical. The experimental attenuation effect is realised precisely through the dampening of these affective vectors: synthetic presence reduces the strength of the empathic pull generated by the Watching-Eye cue, especially within ecologies where affective sensitivity ordinarily drives moral behaviour. In this sense, sentimentalism offers the most proximate normative interpretation of the perturbation mechanism revealed by the data.

7.10 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology

Having reconstructed deontological invariants and consequentialist gradients, we now turn to the normative framework that most directly connects with the causal mechanism revealed by the experiment: *sentimentalism*. In the sentimentalist tradition—Hume, Smith, and contemporary affect-based theorists—moral evaluation originates in the structured responsiveness of the affective system to features of the social world [246, 257, 70, 258]. Moral distinctions are "more properly felt than judged" [246], not because sentiment replaces judgment, but because affective resonance is the primary medium through which moral salience is registered.

7.10.1 The Source of Normativity: Sentiment as the Basis of Moral Appraisal

Sentimentalist normativity arises from patterns of affective response—empathy, warmth, aversion, indignation—that furnish the evaluative significance of morally relevant situations. This aligns closely with the cognitive LoA discussed in Chapter 3: affective tagging (amygdala; insula), empathic resonance (mPFC–TPJ), and rapid harm appraisal provide the first curvature of the evaluative field.

Where deontology imposes constraints and consequentialism imposes gradients, sentimentalism specifies the *affective geometry* of moral space: how warmth draws agents toward prosocial trajectories; how distress or fear generates repulsion;

and how empathic concern shapes the topology through which moral meaning is experienced.

7.10.2 Mode of Evaluation: Affective Resonance as Moral Metric

The sentimentalist mode of evaluation is grounded in:

- **empathic responsiveness** to others' welfare;
- **reactive attitudes** such as guilt, gratitude, and indignation;
- **affiliative and prosocial motivation**;
- **interpersonal attunement** in shared affective contexts.

This structure maps almost exactly onto the **Prosocial–Empathic / Warm–Sociable ecology**. Here, moral relevance is not merely recognised; it is *felt*. Prosocial donation emerges as the behavioural manifestation of a strongly weighted affective vector field.

If moral action is the integral of affective forces across the evaluative field, then any disturbance that reduces the amplitude of these forces will proportionally diminish prosocial behaviour. This is precisely the pattern observed in the experiment.

7.10.3 Action Guidance: Affective Vector Fields and Behavioural Dynamics

Within the evaluative–topological model, sentimentalism becomes computationally explicit when expressed as a dynamical system:

$$\dot{x} = f(x) + \mathbf{A}(x),$$

where $f(x)$ encodes baseline evaluative drift and $\mathbf{A}(x)$ represents affective vectors.

Synthetic presence introduces a deformation operator:

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

where $\delta\mathbf{A}(x; \mathscr{R})$ attenuates or reorients affective flow.

This model captures the empirical pattern with exceptional fidelity:

- **Prosocial–Empathic**: $\delta\mathbf{A}$ dampens empathic activation, flattening the trajectory toward donation.
- **Emotionally Reactive**: $\delta\mathbf{A}$ destabilises an already volatile field, producing the strongest attenuation.
- **Analytical–Structured**: $\delta\mathbf{A}$ is comparatively weak; affect is not the dominant driver.

In short, synthetic presence modulates the evaluative field by *reducing affective curvature*—a canonical sentimentalist effect.

7.10.4 Machine Ethics and the Blind Spot of Affective Architecture

Classical Machine Ethics is structurally incapable of recognising this mechanism. It replaces:

- empathic resonance with rule sets,
- moral perception with logical inference,
- affective appraisal with propositional justification.

But on a sentimentalist account, affect is not peripheral: it is the *substrate* of moral cognition. Our experiment makes this omission explicit. A silent robot, devoid of speech or action, modifies behaviour not by altering rules or utilities, but by reshaping the affective vectors through which moral cues become behaviourally operative.

Machine Ethics has no representational resources for modelling such perturbations. A sentimentalist topology does.

7.10.5 Experimental Realisation: Synthetic Dampening of Empathic Resonance

The core empirical finding is that robotic co-presence attenuates prosocial donation even in the presence of a strong empathic cue (Watching-Eye stimulus). In sentimentalist terms, this corresponds to:

$$\delta \mathbf{A}(x; \mathscr{R}) < 0,$$

for affectively weighted regions of the evaluative field, where:

- x is the agent's evaluative state;
- $\mathbf{A}(x)$ encodes empathic pull and related affective forces;
- $\mathscr{R}$ denotes robotic co-presence.

This inequality states that $\mathscr{R}$ reduces the strength of affective forces driving prosocial action. The perturbation does not reverse moral direction; it *dampens* the affective momentum that would otherwise support donation.

Two mechanisms are plausible:

1. **Affective dilution**: attention and empathic focus are partially diverted to an ambiguous social other.
2. **Affective deflection**: ontological ambiguity disrupts the clarity of empathic pathways toward the child beneficiary.

Cluster differences appear as natural consequences:

- **Prosocial–Empathic**: attenuation via diluted empathic resonance;
- **Emotionally Reactive**: attenuation via heightened volatility;
- **Analytical–Structured**: weak attenuation because affect is not primary.

Sentimentalism therefore provides the most *mechanistically precise* interpretation of the perturbation: synthetic presence alters the affective landscape that underwrites moral sensitivity.

Interpretive Synthesis: Sentimentalism and Synthetic Moral Perturbation

The attenuation of prosocial behaviour under robotic co-presence is a paradigmatic sentimentalist phenomenon. In affectively driven ecologies, $\delta\mathbf{A}(x;\mathscr{R})$ dampens empathic resonance; in volatile ecologies, it amplifies instability; in structurally dominated ecologies, its influence is limited. These cluster-specific dynamics cannot be captured by rule-based or utility-maximising models. They require a framework in which affective forces are constitutive of moral cognition. Reconstructed as a vector-field theory of affective appraisal, sentimentalism thus offers the most direct normative interpretation of the experiment: synthetic presence deforms the affective topology through which moral salience becomes action.

7.11 Contractualism, Particularism, and Hybrid Normative Models

The preceding sections reconstructed deontological, consequentialist, and virtue-theoretic ethics as topological configurations of the evaluative field. To complete the normative architecture required for interpreting the experimental results, we now introduce three additional frameworks—*contractualism*, *particularism*, and *hybrid or pluralist models*. Each is reconstructed briefly but with conceptual precision, and each is integrated into the LoA discipline and the evaluative-topological model that structures this chapter.

Two motivations justify their inclusion. First, these theories constitute major branches of contemporary ethics. Contractualism foregrounds interpersonal justification and mutual accountability [37]; particularism emphasises contextual moral salience over general principles [250, 228]; and pluralist approaches highlight the multidimensionality of moral reasons [259]. Second, the experimental effects demonstrated in this thesis cannot be interpreted solely through invariants, gradients, or dispositional attractors. Rule-based invariants fail to capture context-dependence [260], outcome-based gradients omit intuitive and affective dynamics [31], and virtue-theoretic attractors do not fully explain global field-level perturbations [259]. Minimal cues of social evaluation—watching eyes, ambiguous agency, or robotic presence—modulate cooperation and prosociality across contexts [51, 261, 262, 34]. These phenomena require frameworks that can model justification pressure, situational salience, and relational moral dynamics.

Their inclusion is therefore methodological rather than ornamental. A thesis that aims to integrate ethical theory with empirical results and computational structure must preserve continuity with the normative canon. Without these frameworks, the chapter would lack both systematic coverage and the conceptual resources needed to situate the experimental findings within the full contemporary landscape of moral theory.

7.11.1 Contractualism: Moral Claims as Justification-Equilibria

Contractualism, classically articulated by Scanlon [37], grounds moral rightness in the requirement that one's actions be justifiable to others on principles that no one could reasonably reject. The *source of normativity* is thus located not in rules, welfare, or character, but in the relational structure of mutual accountability.

In the LoA framework, contractualism occupies the reflective normative LoA: it specifies the standards according to which agents can regard themselves as standing in legitimate moral relations. Yet contractualist justification presupposes cognitive capacities—sensitivity to others' perspectives, empathic uptake, and the perception of oneself as under evaluative regard.

Topological interpretation. Contractualism can be expressed as defining *justificatory equilibria* in the evaluative field: regions where an action can withstand the test of mutual recognisability and reasonable non-rejection. Scanlon emphasises the interpersonal nature of moral motivation [37], while Strawson's analysis of reactive attitudes highlights that accountability presupposes recognition of others as answerable participants [263]. These equilibria remain stable only when agents perceive themselves as situated within a network of evaluative regard.

Synthetic presence interacts with this structure in a distinctive way. Watching-eye cues typically heighten the salience of interpersonal accountability, increasing prosociality by intensifying the sense of being answerable to others [261, 262]. A humanoid robot, however, is perceptually social yet ontologically ambiguous. Empirical work shows that such agents can elicit social facilitation while failing to occupy stable interpersonal roles [51, 34]. The result is a perturbation of the justificatory field: the implicit sense of being under the evaluative regard of others is displaced or diluted.

Relevance to the experimental findings. Contractualism illuminates why the Prosocial–Empathic ecology exhibited strong attenuation under robotic presence. Individuals in this ecology are dispositionally sensitive to accountability cues and interpersonal evaluation [**?**]. Under ordinary conditions, the Watching-Eye cue amplifies mutual recognisability and reinforces justificatory pressure to donate [261, 262]. The robot, however, disrupts this justificatory equilibrium: although it triggers social cognition, it does not reliably anchor interpersonal accountability. Its ambiguous status—neither fully agentic nor normatively irrelevant—diminishes the perceived field of mutual evaluative regard [34, 264]. Donation declines not because duty is overridden, nor because consequences are miscalculated, but because the justificatory landscape loses structural integrity.

Thus contractualism interprets the displacement effect as a *deformation of interpersonal accountability*: a weakening of the conditions under which reasons become mutually recognisable and moral motivations are sustained.

7.11.2 Moral Particularism: Contextual Salience and the Fragmented Topology of Reasons

Moral particularism rejects fixed principles, stable evaluative gradients, and invariant reason-valences. On this view, what counts morally in a situation is entirely context-dependent: a consideration that favours an action in one case may count against it in another [66]. McDowell's perceptual account makes the same point in phenomenological terms: moral salience emerges from the concrete situation rather than from any codifiable rule [250]. Work in moral epistemology reinforces this picture, emphasising that evaluative uptake is governed by context-sensitive attention rather than generalisable principles [228].

In evaluative-topological terms, particularism corresponds to a landscape without global invariants or fixed gradients. Instead, the moral field is composed of *local salience contours* that continually shift with changes in attention, affect, and perceptual framing. Empirical research in moral psychology supports this: intuitive responses, perceptual cues, and distributed cognitive processes dynamically determine which features of a situation are experienced as morally significant [16, 31, 233]. Moral appraisal, on this account, is a matter of context-sensitive responsiveness, not rule-following nor global optimisation.

Synthetic perturbation under particularism. If the evaluative landscape is locally assembled, then synthetic presence need not override a stable map—indeed, there may be no stable map to override. Instead, the robot reshapes the *local salience geometry* through which the situation is initially apprehended.

Watching-eye cues heighten accountability salience almost immediately [261], but the introduction of a humanoid robot modifies attention, affect, and perceived agency in more ambiguous ways [51, 34, 264]. The result is not a shift in principle or outcome assessment, but a reordering of which cues enter the evaluative episode first. Social Signal Processing shows that socially meaningful agents exert bottom-up pressure on attentional allocation [67, 73], and HRI studies demonstrate that even minimal humanoid cues redirect gaze and reorganise the perceptual field [265, 266, 267]. Emotion- and attention-based research similarly shows that agentive or affectively salient stimuli suppress competing cues [268, 47, 49].

In this topological setting, synthetic presence functions as a local perturbator: it alters what becomes salient, how quickly, and for how long. For the Prosocial–Empathic cluster, the Watching-Eye stimulus typically heightens empathic attunement and interpersonal accountability. But the robot's ambiguous interpersonal status—neither fully social nor fully inert—introduces a conflicting source of salience that partially eclipses the eye cue. The result is attenuated empathic uptake and reduced prosocial behaviour. This matches perceptual accounts in which the ordering and persistence of salience are constitutive of the evaluative episode itself [250, 70].

For the Emotionally Reactive cluster, the picture is different. Their evaluative fields are already dominated by situational micro-variability; the robot introduces noise, but not disruption relative to an already-fluid topology. This is exactly what particularism predicts: the more context-sensitive the agent, the weaker the relative effect of an additional perturbation.

7.11.3 Hybrid and Pluralist Models: Multidimensional Evaluative Topologies

Hybrid or pluralist theories—from Ross's irreducible prima facie duties [269] to contemporary value pluralism [110]—hold that normativity arises from multiple independent sources. Moral assessment is shaped by the interplay of constraints, outcomes, character, relationships, and contextual considerations [270, 271, 80, 37]. No single evaluative dimension dominates.

Topologically, pluralism corresponds to a *multi-dimensional evaluative manifold*. Rather than a single axis of moral value, the evaluative field contains intersecting constraints, gradients, attractors, and salience structures. Psychological and neurocognitive research supports this picture: affective intuitions, rule-based processes, and outcome-tracking mechanisms operate semi-independently and interact dynamically in judgment [16, 31, 241]. Moral appraisal is thus the navigation of a field shaped by heterogeneous normative forces.

Why pluralism fits the experimental results. The experimental displacement effect is best understood as a *manifold-level perturbation*. Each normative dimension is involved:

- Watching-eye cues activate deontic expectations (public accountability).
- Donation expresses consequentialist gradients (welfare benefits).
- Cluster-level differences reflect virtue-theoretic dispositions.
- The robot refracts interpersonal meaning (contractualist disruption).
- Salience competition reflects particularist sensitivity to context.

No single theory predicts the uniform attenuation across clusters. Instead, the results indicate that synthetic presence modulates several normative gradients simultaneously. The robot alters empathic resonance, perceived accountability, attentional competition, and expected social payoffs at once [34, 157, 267, 264, 51, 188]. The Watching-Eye effect, ordinarily robust, is dampened by competing social signals—precisely the pattern revealed in studies of accountability cues and attentional capture [261, 262, 47, 49, 67, 73].

The experiment thus provides empirical grounding for the core claim of normative pluralism: moral judgment emerges from the configuration of multiple evaluative dimensions, each susceptible to contextual perturbation [269, 110, 37]. The robot's presence produces a field-level reconfiguration, not merely a shift in a single evaluative axis.

Pluralism and dispositional structure. This field-level displacement does not contradict the stable trait differences revealed by the clustering analysis. The clusters represent distinct *starting positions* within the manifold—different dispositional orientations that shape ordinary evaluative navigation. But the robotic perturbation acts on the *shared topology* of the field itself. This is why all clusters, despite psychological divergence, show a consistent directional attenuation. Dispositions shape baseline trajectories; synthetic presence reshapes the manifold in which those trajectories unfold.

In pluralist terms, the robot perturbs the evaluative manifold, not the individual gradients. The cluster analysis and the displacement effect therefore capture complementary layers of moral cognition: enduring dispositional geometry and context-sensitive field-level modulation.

7.11.4 Integrative Ethical Interpretation of the Experimental Findings

Bringing the reconstructed frameworks together, we can now articulate the ethical significance of the experimental results in a manner that reflects both the normative pluralism developed throughout this chapter and the dual-layer structure of moral cognition revealed empirically. The attenuation of prosocial donation under robotic co-presence does not arise from the weakening of a single moral principle or evaluative dimension. Rather, it reflects a *global perturbation* of the evaluative field—the structured moral ecology in which diverse moral reasons are ordinarily weighted, integrated, and rendered behaviourally operative.

1. **Deontological lens: weakened accountability cues.** The robot diminishes the felt presence of a morally relevant observer, thereby attenuating the duty-oriented accountability that the Watching-Eye cue is designed to amplify. The displacement effect indicates a disruption in the implicit normative expectations that scaffold rightful agency, rather than a violation of explicit moral rules.

2. **Consequentialist lens: flattened outcome gradients.** Synthetic presence alters the perceived payoff structure of helping. Reputational, affective, and interpersonal "returns" become less sharply defined, flattening the gradient that normally favours donation. Altruistic output declines not because agents miscalculate utility, but because the social-evaluative topology itself has shifted.

3. **Virtue-theoretic lens: dispositional curvature under field-level modulation.** The perturbation does not target trait-based motivations directly. Instead, it reveals that even robust dispositional architectures—captured in the psychometric clusters—are expressed *within* an evaluative field susceptible to contextual deformation. The uniform directional shift in donation across clusters demonstrates that character is not a self-contained engine of action but a gradient embedded in a modifiable field.

4. **Contractualist lens: disrupted justificatory equilibrium.** Contractualist motivation depends on recognising the presence of others to whom reasons are owed. The robot introduces ambiguity into this interpersonal field, weakening the sense of mutual answerability. The justificatory landscape becomes noisier and less structured, reducing the force of the requirement to act in ways that others could not reasonably reject.

5. **Particularist lens: reconfigured salience geometry.** The robot alters the fine-grained pattern of contextual salience. The Watching-Eye cue remains physically present, but its normative traction is displaced by a new and ambiguous source of social meaning. What becomes salient first—and for how long—changes, thereby altering the evaluative episode itself.

6. **Pluralist-topological lens: manifold-level displacement.** The findings are precisely what a pluralist model predicts when multiple normative gradients interact with a global perturbation to social meaning. The donation attenuation reflects not the suppression of a single evaluative dimension but a deformation of the multi-dimensional evaluative manifold. This explains both the robustness and the cross-cluster consistency of the effect.

Taken together, these interpretations converge on a unified thesis:

Integrative Conclusion: The Ethical Signature of Moral Displacement

The presence of a humanoid robot reshapes the multi-dimensional evaluative topology through which moral salience becomes action. This perturbation operates at the level of the evaluative field itself, modulating deontic expectations, consequentialist gradients, dispositional attractors, justificatory relations, and contextual salience structures simultaneously. No monolithic ethical framework captures this phenomenon. The experimental results therefore vindicate a pluralist, topological, empirically grounded model of moral cognition—revealing how synthetic agents can globally displace moral evaluation in ways systematically overlooked by classical Machine Ethics.

By reconstructing the major normative theories through Levels-of-Abstraction discipline and embedding them within a topologically structured model of moral cognition, this chapter has provided the conceptual architecture required to understand the ethical significance of synthetic moral perturbation. The experiment demonstrates how such perturbation manifests as a field-level displacement effect, thereby integrating normative theory, cognitive psychology, and computational modelling into a unified account of how artificial agents reshape the evaluative terrain of human moral behaviour.

8. General Discussion and Theoretical Integration

8.1 Introduction: Why the Experiment Requires a Structural Interpretation

The preceding chapters developed three interconnected strands: (i) a cognitive–affective account of moral judgment, (ii) a normative–philosophical reconstruction of ethical theory through the lenses of Level-of-Abstraction discipline and evaluative topology, and (iii) an empirical demonstration that robotic copresence systematically attenuates prosocial donation under morally salient conditions.

Before turning to the integrative task, it is necessary to articulate the higher-order insight guiding the trajectory of this thesis. Situated within the cognitive, philosophical, and formal analyses of the preceding chapters, the empirical study indicates that *moral decision-making is, at root, a practical phenomenon*, grounded in the structures of agency and practical reason [79, 80, 272, 228, 69]. Moral events are not abstract judgements suspended in conceptual space; they are situated transitions from perception to action embedded in a socially organised environment, consistent with empirical models that treat moral cognition as perceptual, affective, and socially modulated [16, 17, 60, 50, 148]. Because such events culminate in observable behavioural outputs, they are empirically tractable and available to systematic measurement and analysis [178, 179, 177]. Their structural and methodological precision is rarely recognised in the prevailing discourse of Machine Ethics and Computational Morality, which has long been criticised for its limited integration of empirical findings [52, 23, 53, 54].

This sequence is methodologically significant. Across both philosophy and moral psychology, ethical inquiry typically proceeds not by legislating the quality of actions from a priori first principles, but by beginning with the existence of *moral events* themselves—episodes in which agents respond to cues, saliences, and social affordances—and then seeking theoretical structures that best explain these patterns of behaviour [69, 228, 273, 82, 45]. This bottom-up orientation stands in sharp contrast to much of the historical trajectory of Machine Ethics, which has principally advanced top-down models that attempt to encode or implement normative theories prior to securing an empirical understanding of how moral cognition unfolds in practice.

A large body of Machine Ethics scholarship exemplifies this top-down, normative-first orientation. Early and influential work sought to engineer explicit ethical rules or principles for artificial agents [52, 23, 20], often drawing upon deontological, utilitarian, or virtue-theoretic frameworks whose normative structure was taken as directly implementable in computational systems [22, 274, 275, 276, 277, 278, 279]. Subsequent developments reinforced this tendency by constructing logical architectures intended to represent moral constraints, permissibility conditions, or value hierarchies independently of empirical models of human moral agency [280, 281, 282, 283, 284, 285]. Even approaches motivated by psychological plausibility, such as computational models of ethical reasoning [231, 24, 286], largely inherit the same structural assumption that normative content can be specified in advance of empirical measurement.

Critiques of this methodological inversion are now widespread. Authors working within both ethics of AI and social-robotics research argue that designing moral agents without grounding in empirical evidence about cognition, affect, social interaction, or developmental patterns of moral behaviour is epistemically unstable and risks constructing systems whose 'moral' outputs lack psychological validity [53, 287, 288, 289, 54, 255]. On these accounts, moral behaviour cannot be treated as an externally specifiable target for implementation; rather, it emerges from structured interactions among cognitive, affective, embodied, and social-signalling processes [18, 50, 290, 67, 148]. These processes must therefore be empirically characterised before any attempt at normative codification. Only through such empirically informed grounding can normative theory enter the analysis in a methodologically stable and scientifically responsible manner.

The present work therefore advances a methodological reversal. It shows that moral salience, moral displacement, and the perturbation of prosocial behaviour are empirically measurable phenomena that *must* be mapped before being codified, an approach supported by behavioural studies of attentional and prosocial modulation [65, 2, 5, 31, 17, 48]. Because these phenomena are embedded within attentional, affective, and dispositional architectures, they admit rigorous experimental design, statistical modelling, and formal reconstruction [178, 179, ?]. Accordingly, the experimental study is not an auxiliary illustration but the epistemic anchor of the thesis. Only once the structure of moral events is empirically established can normative theory enter the analysis—precisely the reverse of the methodological sequence characteristic of Machine Ethics, normative-first LLM evaluation, and much of Affective Computing [52, 23, 55, 56, 57, 291, 292].

The task of the present chapter is not to repeat these analyses, but to integrate them. It offers a theoretical synthesis that explains *why* the experimental effect occurs, *what* its ethical significance is, and *how* it reshapes the methodological landscape for research in Human–Robot Interaction, moral psychology, and the emerging field of Computational Morality.

In this sense, the experiment is not an isolated behavioural result but a *probe* into the architecture of moral cognition. The observed attenuation of prosocial behaviour is theoretically meaningful only when interpreted through the structures developed earlier: dual-process architectures, the Social Intuitionist Model, evaluative topology, and the reconstructed normative frameworks of deontology, consequentialism, virtue ethics, sentimentalism, contractualism, and particularism. The present chapter therefore provides a synoptic interpretation in which the behavioural signature revealed by the data becomes a lens through which the nature of moral cognition—and its vulnerability to perturbation—is rendered theoretically transparent.

8.1.1 From Behaviour to Structure: Why a Higher-Level Interpretation is Required

The experimental paradigm—Watching-Eye moral cue embedded within a silent synthetic presence—does not merely generate a difference in donation behaviour; it reveals a deformation of the evaluative field that links moral salience to action. Classical interpretations of donation differences (e.g., generosity, altruism, compliance) lack the conceptual resources to capture this phenomenon. A purely behavioural description would record that participants donated less in the Robot condition, with the Prosocial–Empathic cluster showing the numerically steepest decline. But such a description omits the structural logic that makes the result scientifically and philosophically significant.

The central claim developed throughout the thesis is that *moral behaviour is not invariant under changes to the perceptual–social environment.* The robot's presence does not overwrite moral norms nor impose new ones; instead, it modifies the cognitive–affective conditions under which evaluative forces act. It shifts attentional allocation, alters affective resonance, and modifies the perceived sociality of the space. In topological terms, the robot introduces a perturbation γ_R that deforms the curvature of the evaluative manifold, thereby weakening the salience gradient induced by the Watching-Eye stimulus.

A simple behavioural difference thus reflects a deeper structural transformation in the evaluative field. As demonstrated by the regression models and Bayesian estimation, the attenuation effect was uniform in direction across participants, indicating that the perturbation introduced by the robot operates at the field level rather than through trait-specific pathways. Yet this uniformity does not imply psychological homogeneity.The PCA–k-means clustering revealed three coherent dispositional ecologies—distinct configurations of empathic resonance, affective volatility, and structural–analytical processing. These ecologies are consistent with the established dimensions of empathizing and systemizing [85], personality variation captured by the BFI-10 [153], and broader accounts of moral-psychological "ecologies" that organise evaluative processing [148, 16]:

- the **Emotionally Reactive / Low-Structure Profile**,
- the **Prosocial–Empathic / Warm–Sociable Profile**,
- the **Analytical–Structured / High-Systemizing Profile**.

These clusters instantiate different evaluative topologies—distinct attractor formations, sensitivities to perceptual and affective salience, and pathways of modulation—consistent with multidimensional models of affective valuation and moral cognition [241, 83, 10, 16]. Within this framework, the Prosocial–Empathic cluster exhibits the steepest affective gradients and the strongest baseline responsiveness to Watching-Eye cues. This ecological structure aligns with theoretical expectations: Watching-Eyes primes amplify empathic accountability [261, 262], and empathic resonance is known to be highly sensitive to contextual modulation [49].

That this cluster nevertheless showed the same directional attenuation as the others is therefore theoretically significant. Rather than reflecting a trait-dependent shift, the humanoid robot's ambiguous social presence perturbs the salience structure itself, weakening the amplification mechanisms on which empathic ecologies depend [34]. In other words, the perturbation operates *upstream* of individual dispositional pathways: it modifies the evaluative field within which those pathways are embedded. The displacement observed in the experiment is thus best understood as a *field-level suppression of moral salience*, overriding the ordinarily divergent dispositional trajectories that shape prosocial behaviour.

Ethical Interpretation: Why the Attenuation Matters Normatively. The ethical significance of this finding becomes visible only when the result is interpreted through the reconstructed normative frameworks developed in Chapter 7. Each theory identifies a different locus of normative structure, and each provides a distinct—yet convergent—reading of the deformation caused by $\mathscr{R}$:

- *Deontological perspective.* The Watching-Eye cue implicitly invokes deontic expectations of reciprocity, fairness, and beneficence. The robot's presence attenuates donation precisely by dulling this sensitivity. Normatively, this appears as a disruption of the agent's capacity to track *ought-constraints* in the environment—an interference with the cognitive substrate on which deontic responsiveness relies.
- *Consequentialist perspective.* The moral field includes gradients of anticipated social evaluation. Watching-Eye cues steepen these gradients; synthetic presence flattens them. The robot therefore functions as a *gradient-suppressor*, reducing the perceived payoff of prosocial action. In topological terms: it alters the vector field governing welfare-oriented trajectories.
- *Virtue-ethical perspective.* The three clusters correspond to differing dispositional configurations. The strongest attenuation occurring within the Prosocial–Empathic cluster implies that the robot disrupts precisely those virtues—empathy, warmth, prosocial orientation—that ordinarily stabilise prosocial attractors. The perturbation thus interacts with *character topology* rather than bypassing it.
- *Sentimentalist (Humean) perspective.* The attenuation reflects a dampening of empathic vector fields: $\delta\mathbf{A}(x;\mathscr{R}) < 0$. The robot selectively reduces affective resonance with the Watching-Eye cue. Normatively, this implies that the moral valence of the situation is felt less intensely, weakening the motivational energy required for prosocial action.
- *Contractualist perspective.* The moral event of donation under observation involves tacit justifiability relations: "What could reasonably be expected of me in the eyes of others?" The ambiguous presence of a synthetic observer destabilises this justificatory equilibrium. The subject no longer clearly apprehends *to whom* justifiability is owed.
- *Particularist perspective.* Moral appraisal depends on local saliences. The robot modifies the salience landscape: the morally relevant cue (the child in need) becomes less perceptually dominant. Thus, the attenuation is interpreted as a shift in the pattern of reasons that obtain in this particular context.

LoA Interpretation: Why the Perturbation Occurs at the Wrong Level for Machine Ethics. Floridi's Level-of-Abstraction analysis clarifies the structural error revealed by the experiment. The attenuation does *not* occur at the normative LoA (where duties, values, or justifiability live), but at the cognitive–affective LoA (where salience, resonance, and attention are regulated). Machine Ethics traditionally operates at the wrong LoA: it attempts to implement high-level normative constructs while ignoring the low-level substrates on which moral responsiveness depends.

The experiment shows why this is untenable. Ethical responsiveness is mediated by:

- attentional allocation (Who or what do I notice?)
- affective resonance (What emotional weight does this carry?)
- perceived social ontology (Who counts as the observer?)
- dispositional pathways (How does my cognitive ecology integrate this cue?)

Synthetic presence perturbs all of these upstream mechanisms. Thus, even perfect normative reasoning at a reflective LoA cannot salvage moral action when the lower-level architecture of moral cognition has been deformed. In Floridi's terms:

> *Normative correctness is orthogonal to causal efficacy. A system may know what is right and yet fail to act rightly if the cognitive LoA is perturbed.*

Integrative Insight. The field-level suppression observed in the experiment therefore reveals a principle of broad ethical and psychological importance:

> *Moral failure under synthetic presence is not a failure of principle but a failure of salience. Ethical norms lose their grip not because agents reject them, but because the evaluative machinery that normally brings them to bear is disrupted.*

This insight is the conceptual hinge on which the whole thesis turns. It unifies:

- the cognitive architecture (moral judgments arise from salience → appraisal → integration),
- the topological formalism (moral cues define gradients and attractors),
- the normative frameworks (moral theories describe different structural aspects of the evaluative field),
- and the empirical results (synthetic presence suppresses these structures at the field level).

With these interpretive tools in place, we can now proceed to the cluster-by-cluster integrative analysis that further refines the ethical and cognitive significance of the experimental findings.

8.1.2 Why This Chapter Cannot Be Pure "Discussion" in the Conventional Sense

Traditional discussion chapters in empirical theses typically emphasise methodological limitations, alternative interpretations, and directions for future work. While such elements remain relevant here, they are insufficient for the present project. The experiment developed in this thesis sits at the intersection of cognitive science, social robotics, computational modelling, and normative ethics. The behavioural effect it reveals—reduced prosocial donation under synthetic co-presence—is only the observable trace of a deeper structural transformation: a perturbation of the evaluative machinery through which agents convert moral salience into action. Because this transformation engages multiple theoretical layers—cognitive–affective processing, dispositional topology, normative interpretation, and Level-of-Abstraction analysis—a standard discussion section cannot capture its full conceptual significance. What is needed instead is a structural synthesis that explains not merely *what* happened, but *why* it happened and *what it reveals* about the nature of moral cognition and its vulnerability to synthetic perturbation.

To articulate this phenomenon requires a conceptual integration that cannot be confined to standard "discussion" categories. Instead, the chapter must synthesise:

1. the **cognitive architecture** (dual-process, SIM, dynamic integration);
2. the **evaluative geometry** (topology, curvature, gradient flow);
3. the **normative reconstruction** (deontic invariants, consequentialist gradients, dispositional attractors, sentimentalist vector fields, contractualist justificatory structure, and particularist salience responsiveness);
4. and the **empirical structure** of the data (cluster-specific susceptibility, Bayesian attenuation, topological deformation of the Watching-Eye effect).

The present chapter therefore functions as an *interpretive pivot*: it translates the empirical findings into philosophical insight, and reinterprets philosophical frameworks in light of empirical constraints.

8.1.3 A Structural Reading of the Core Experimental Result

The empirical pattern can be summarised as follows:

- The humanoid robot NAO is perceptually salient but ontologically ambiguous.
- The Watching-Eye cue ordinarily induces an empathic salience gradient that increases donation.
- The robot introduces a perturbation γ_R that competes with, and partially overrides, this empathic amplification.
- Attenuation is strongest in the Prosocial–Empathic cluster, weaker in the Analytical–Structured cluster, and statistically negligible in the Emotionally Reactive cluster.

Interpreted through the cognitive framework developed earlier, this pattern shows that moral appraisal begins with intuitive and affective resonance [16, 31]. Synthetic presence disrupts this resonance by altering attention, salience, and perceived sociality [47, 49, 34, 51]. Different dispositional structures absorb this disruption in systematically different ways, consistent with established dimensions of empathizing, systemizing, and moral-schema variability [85, 233]. The resulting behavioural output reflects not a change in moral principle, but a deformation of the evaluative field.

Interpreted through the normative framework, the same pattern yields multiple structurally coherent readings:

- a **deontological reading**: synthetic presence weakens the implicit deontic expectations cued by the Watching-Eye stimulus [15];
- a **consequentialist reading**: synthetic presence flattens the perceived payoff gradient of helping behaviour [**?**];
- a **virtue-ethical reading**: synthetic presence suppresses prosocial attractors associated with empathic or cooperative dispositions [69];
- a **sentimentalist reading**: synthetic presence dampens empathic vector fields that ordinarily drive prosocial action [70];
- a **contractualist reading**: synthetic presence destabilises the justificatory relations normally activated by social observation [37];
- a **particularist reading**: synthetic presence alters the salience pattern such that the Watching-Eye cue no longer carries the same moral significance [66, 250].

Thus, each normative theory yields a structurally distinct but empirically convergent interpretation. The ethical significance of the experiment lies not in any single framework, but in the *coherent intersection* of all of them: a field-level suppression of moral salience, a deformation of the evaluative topology through which moral meaning becomes action.

8.1.4 Why the Synthetic Presence Effect Matters Beyond the Experiment

The attenuation of moral action under synthetic presence is not merely an interesting behavioural anomaly; it demonstrates a deeper principle: *moral cognition is structurally permeable.* It is sensitive to perturbations that operate below the level of explicit reasoning. It is vulnerable to shifts in perceived social ontology. And it is modulated by affectively weighted cues whose influence is seldom acknowledged in normative theory and almost never incorporated in classical Machine Ethics.

This has far-reaching implications:

1. It challenges the assumption that artificial agents can be designed according to purely deliberative ethical frameworks.

2. It shows that synthetic presence modulates moral behaviour even without action, speech, intent, or agency.
3. It reveals that human–robot environments are *ethically loaded* by virtue of perceptual and affective structure alone.
4. It demands a reconsideration of how artificial systems are situated within the moral ecology of human decision-making.

In short, the experiment demonstrates a fact of philosophical significance: *synthetic agents are not normatively inert.* Their presence, even in silent passivity, can deform the evaluative pathways through which moral salience becomes action.

The remainder of this chapter builds on this foundation. Subsequent sections provide:

- a cluster-by-cluster integrative interpretation,
- a cross-framework normative synthesis,
- a critique of monolithic Machine Ethics,
- a reconstruction of Computational Morality grounded in empirical structure,
- and a final consolidation of the thesis' theoretical contributions.

The goal is not only to interpret the experiment, but to show how the experiment reconfigures the conceptual terrain on which research in moral psychology, HRI, and Machine Ethics must proceed.

8.2 Cluster-by-Cluster Integrative Interpretation

Broken Link

The experimental results demonstrate that robotic co-presence $\mathscr{R}$ induces a uniform directional attenuation of prosocial donation across participants, yet the *structure* of this attenuation differs meaningfully across the three latent cognitive–affective ecologies uncovered in Chapter ??. Because these clusters instantiate distinct evaluative topologies—different attractor formations, salience gradients, affective vector fields, and pathways of regulatory modulation—their differential perturbation under $\mathscr{R}$ offers insight into the architecture of moral cognition and the ethical significance of synthetic presence. What follows is an integrative interpretation weaving together the cognitive, topological, normative, and Level-of-Abstraction (LoA) analyses developed across the thesis.

Emotionally Reactive / Low-Structure Ecology

This ecology exhibits high affective volatility, shallow structural integration, and weak systemizing constraints, consistent with established empathizing–systemizing variability [85]. Its evaluative topology is characterised by *broad, low-gradient attractors*: intuitive responses are strong but unstable; attentional salience fluctuates; and the transition from perception to action is mediated by short-lived affective surges rather than sustained deliberative integration.

To avoid terminological ambiguity, it is useful to clarify what is meant here by *broad, low-gradient attractors* in the evaluative–topological framework. In dynamical-systems terms, an attractor represents a region of the evaluative field $\mathcal{E}$ toward which the system's state x naturally converges [293, 294]. A *broad* attractor denotes a basin of attraction with wide boundaries and weak curvature, meaning that many initial states can enter it but none are strongly pulled toward a particular behavioural endpoint. A *low-gradient* attractor is one in which the magnitude of the evaluative gradient $\|\nabla\mathcal{E}(x)\|$ is small across the basin, implying that movement toward prosocial or antisocial trajectories is governed by shallow motivational forces [295, 296].

In psychological terms, this configuration corresponds to intuitive reactions that are easily triggered yet weakly stabilised: the agent may experience transient affective spikes (e.g., momentary empathy, irritation, or ambivalence) without these signals generating a consistent or directed behavioural tendency. This interpretation is consistent with empirical models of low-coherence affect, affective lability, and unstable salience allocation [297, 298, 299]. Because the evaluative landscape lacks sharply defined slopes, small perturbations—including those introduced by environmental ambiguity—tend not to produce substantial directional change. This explains why the Emotionally Reactive / Low-Structure ecology exhibited behavioural invariance in the experiment: the moral field was already characterised by diffuse attractors and unstable salience dynamics, leaving little structured curvature for $\mathscr{R}$ to deform.

Within such a landscape, the experimentally observed pattern—minimal or noisy attenuation—is theoretically revealing. The Watching-Eye stimulus σ_{WE} generates only a modest prosocial gradient for this cluster [261, 262], and the robot-induced perturbation γ_R cannot significantly deform a field that already lacks curvature:

$$|\nabla\mathcal{E}_{\text{baseline}}| \approx 0 \quad \Rightarrow \quad |\nabla\mathcal{E}_{\text{perturbed}}| \approx 0.$$

At the cognitive LoA, this ecology functions as a near-critical system: its evaluative machinery exhibits little stability and thus provides minimal structural leverage for $\mathscr{R}$ to disrupt. Normatively, this implies that deontic, sentimental, or virtue-theoretic structures exert limited behavioural influence because the underlying evaluative field lacks the curvature to sustain them.

Prosocial–Empathic / Warm–Sociable Ecology

This cluster displays high empathic resonance, strong sensitivity to social cues, and rich affective attractors. Psychological models of empathic processing support this heightened salience responsiveness [47, 49]. Its evaluative topology is steeply sloped: the Watching-Eye cue generates strong upward gradients toward prosocial action [261], mediated by interpersonal appraisal and affective amplification.

The robot's ontological ambiguity [34, 33, 51] perturbs precisely this amplification mechanism. As demonstrated in Chapter **??**, the perturbation $\delta\mathcal{E}(x;\mathscr{R})$ acts *upstream*, modifying the salience structure itself:

$$\delta\mathcal{E}(x;\mathscr{R}) < 0, \qquad \delta\mathbf{A}(x;\mathscr{R}) < 0.$$

Because the empathic system depends on affective curvature, flattening the field produces the *largest attenuation* in this ecology despite its strong baseline gradients.

Normatively, this yields a convergent interpretation: deontology registers weakened duty-tracking; consequentialism observes a flattened payoff gradient; virtue ethics identifies destabilised prosocial dispositions; sentimentalism finds dampened empathic force-fields; contractualism diagnoses disrupted justificatory orientation; and particularism detects a shift in which contextual features count as reasons.

Analytical–Structured / High-Systemizing Ecology

This ecology exhibits strong systemizing tendencies and comparatively lower empathizing [85]. Its evaluative topology is governed by structural coherence rather than affective curvature. Here, prosocial action arises from rule-consistency, interpretive stability, and contextually well-defined cues.

The experiment reveals only mild attenuation. The Watching-Eye cue produces modest gradients, while $\mathscr{R}$ introduces representational and social-ontological ambiguity [157], subtly undermining the interpretive regularities on which this ecology relies. The perturbation operates primarily on semantic and predictive structure:

$$\delta\mathcal{E}(x;\mathscr{R}) \approx 0^{-}, \qquad \delta\mathbf{A}(x;\mathscr{R}) \approx 0.$$

At the LoA level, this ecology demonstrates that perturbation need not be affective: synthetic presence also functions as a *semantic disruptor*, altering the representational substrate needed for structured evaluative computation. Normatively, this corresponds to weakened rule-clarity (deontology), distorted outcome-modelling (consequentialism), and destabilised interpretive virtues such as discernment and practical wisdom (virtue ethics).

Integrative Synthesis

Across all three ecologies, a unified conclusion emerges: the humanoid robot operates not through communication, norm expression, or explicit social signalling, but through *topological reconfiguration*. It introduces a perturbation γ_R at the cognitive LoA that:

- suppresses affective gradients in empathic ecologies,
- introduces semantic and predictive ambiguity in analytical ecologies,
- and interacts minimally with shallow attractor fields in reactive ecologies.

Normatively, the attenuation is not a failure of duty, utility estimation, virtue, empathy, or justificatory reasoning. Instead, it represents a *structural displacement of moral salience*. This displacement is invisible to explicit reasoning yet measurable in behaviour and interpretable through evaluative topology.

In this sense, the humanoid robot reveals a property of moral cognition that classical ethical theory and classical Machine Ethics could not predict: *moral responsiveness is field-sensitive*. Normativity becomes action only when the evaluative field retains its curvature. Perturb the field, and even well-formed dispositions cannot operate normally.

This insight forms the conceptual hinge for the remainder of the General Discussion.

8.3 Global Normative–Topological Synthesis

The final integrative step requires bringing together the three interpretive lenses that structure this thesis: (i) the *topology* of moral cognition, (ii) the *normative frameworks* reconstructed in the Ethical Cognition chapter, and (iii) the *empirical perturbation* revealed by the experiment. The aim is not to select a single normative theory that "best explains" the data, nor to impose a moral verdict on participants' behaviour. Rather, the task is to demonstrate how the experimental findings become theoretically intelligible *only* when analysed at the correct Level of Abstraction (LoA), through a structure-sensitive account of evaluative dynamics.

Moral Behaviour as a Field-Level Phenomenon

Across deontological, consequentialist, virtue-theoretic, sentimentalist, and contractualist frameworks, one structural insight remains invariant: **moral action does not arise from isolated psychological modules or explicit rule execution**. Instead, it emerges from the configuration of the evaluative field—a relational structure shaped by perception, affect, social meaning, habituation, and normative commitments.

The experiment demonstrates that this field is *globally deformable*: a silent humanoid robot, devoid of agency, instruction, or communication, attenuates prosocial behaviour across all dispositional ecologies. This uniform directionality, combined with cluster-specific differences in amplitude, reveals a core computational insight:

The presence of $\mathscr{R}$ acts as a field-level perturbation, not a trait-level driver.

In topological terms, the robot introduces a deformation operator

$$\gamma_R : \mathcal{E} \to \mathcal{E}',$$

which modifies the curvature of the evaluative manifold such that moral salience diffuses more weakly toward prosocial attractors. This accounts for both the global donation reduction and the heterogeneous susceptibility across ecologies.

Deontological, Consequentialist, and Virtue-Ethical Readings of the Perturbation

The experiment's ethical significance becomes transparent when interpreted through the normative frameworks reconstructed earlier:

- **Deontological interpretation:** The Watching-Eye cue implicitly invokes deontic norms of accountability and interpersonal respect. The attenuation of donation under $\mathscr{R}$ is thus intelligible as a deformation of the agent's sensitivity to these constraints. The robot does not induce norm violation; it *weakens the agent's access* to deontic salience by altering the perceived sociality of the environment.

- **Consequentialist interpretation:** Watching-Eye cues are known to reshape the perceived consequence structure of prosocial acts. The robot's ambiguous presence disrupts this gradient, flattening reputational and affective payoff structures. Donation decreases because the local value landscape is deformed, not because agents become less "ethical."
- **Virtue-ethical interpretation:** The dispositional ecologies uncovered in the clustering analysis map directly onto virtue-ethical accounts of character as a structured, learned sensitivity to moral salience. $\mathscr{R}$ perturbs the field *upstream* of these dispositions, weakening the operative mechanisms of moral perception, especially in the Prosocial–Empathic / Warm–Sociable profile.

Each framework thus provides a different interpretive contour of the same phenomenon. But they converge on one central point: **the perturbation acts on the evaluative field, not on the moral principles themselves**. The agents' normative commitments remain intact; what changes is the salience structure through which those commitments become behaviourally operative.

Sentimentalist, Contractualist, and Particularist Convergence

Sentimentalist theories construe moral judgment as an affective vector field. Under this lens, the robot acts as a dampening force on empathic resonance, decreasing the magnitude of affective gradients required to activate prosocial behaviour. Cluster-specific differences in attenuation severity become intelligible as differences in affective sensitivity and evaluative slope.

Contractualist and justificatory theories interpret the perturbation as a shift in the perceived interpersonal structure of the environment. When $\mathscr{R}$ is present, participants implicitly alter their model of who counts as a moral interlocutor—a phenomenon well-documented in human–robot interaction literature. This recategorisation subtly modifies the justificatory landscape in which prosocial acts acquire meaning.

Particularist and perceptualist theories emphasise moral *attention*. On this view, the robot acts as a competing centre of salience, pulling attentional weight away from the Watching-Eye cue and thereby diluting the moral percept. This aligns precisely with the empirical finding of attenuated donation despite a strong moral prime.

Floridi's Level-of-Abstraction Reading

Floridi's LoA discipline allows us to state the integrative conclusion succinctly:

- At the **cognitive LoA**, the robot perturbs perceptual-affective mechanisms (attention, salience, resonance).

- At the **behavioural LoA**, this perturbation manifests as reduced prosocial action.
- At the **normative LoA**, the agent's ethical commitments remain unchanged, but the pathway by which they become operative is deformed.

This avoids the two characteristic errors of Machine Ethics:

1. treating normative principles as if they were generative psychological operators;
2. treating behavioural shifts as if they were moral judgments.

Integrative Conclusion: Moral Salience, Synthetic Presence, and the Architecture of Agency

Integrative Conclusion: The Ethical Significance of Synthetic Perturbation

The experiment demonstrates that synthetic presence can alter moral action not by introducing new norms or violating existing ones, but by reshaping the evaluative topology through which moral salience acquires behavioural force. Deontological constraints, consequentialist gradients, virtue-theoretic dispositions, sentimental vector fields, and contractualist justificatory demands all converge on the same structural insight: the moral field is deformable. The humanoid robot acts as a perturbation operator γ_R on this field, weakening the pathways that normally lead from moral perception to prosocial action. This field-level deformation explains both the global attenuation effect and the cluster-specific signatures discovered in the experiment. It also reveals a fundamental limitation of classical Machine Ethics: normative content cannot be operationalised without an empirically grounded account of how moral cognition functions within its situational topology. The thesis therefore establishes a new methodological foundation for Computational Morality: synthetic agents must be analysed not merely as potential moral reasoners, but as operators on the moral ecology in which human agency unfolds.

8.4 From the Failure of Machine Ethics to a Reconstruction of Computational Morality

The preceding analyses show that robotic co-presence $\mathscr{R}$ induces a deformation of the evaluative field within which moral salience becomes action. This has direct implications for artificial moral agency and exposes a structural flaw in classical Machine Ethics. Since its inception, Machine Ethics has assumed that moral behaviour can be engineered by encoding ethical principles inside an artificial system—a view explicit in rule-based architectures [52, 21], utilitarian optimisation frameworks [22], virtue-based computational agents [23], and logic-driven decision systems [20, 280]. These approaches presuppose that normative theories function as *implementable specifications*. However, as Floridi's Levels of Abstraction make clear [25, 300], this constitutes a category mistake: normative theories belong to a reflective LoA, whereas moral behaviour emerges at the cognitive LoA through complex interactions of salience, affect, social signalling, and controlled appraisal.

Moral psychology and cognitive science provide a clear counterpoint to the Machine Ethics assumption. Decades of research show that moral behaviour is not generated by rule execution but by intuitive-affective processes [16], conflict-sensitive valuation systems [31], affective-perceptual mappings [241], and schema-based social cognition [233]. Moral appraisal begins with rapid, pre-reflective resonance shaped by perceptual salience [47], empathic responsiveness [49], and contextual cues. The empirical results of this thesis reinforce these findings: robotic presence modifies salience structures upstream of conscious evaluation, consistent with work showing that synthetic agents alter social perception and norm-related behaviour even in minimal-interaction contexts [51, 34, 33].

Machine Ethics models fail to capture these mechanisms. Deontic architectures presuppose invariant constraints, yet even deontic cues—such as Watching-Eye effects [261, 262]—can be attenuated by the mere presence of a humanoid robot. Utilitarian architectures assume stable value gradients, yet the data show that gradients of perceived social consequence are flattened by ontological ambiguity [51]. Virtue-based systems assume globally stable traits, yet situationist critiques [259] and schema ecologies [233] reveal substantial dispositional heterogeneity; the experiment confirms that dispositional structure alone cannot explain behavioural attenuation. Sentimentalist architectures—which would predict affective resonance as a core driver of moral action—are almost entirely absent from Machine Ethics, despite overwhelming evidence that empathy and affective salience strongly modulate moral behaviour [49, 16].

The methodological failure is thus profound. Classical Machine Ethics implicitly assumes:

$$\text{Normative authority} \;\Rightarrow\; \text{Behavioural generation.}$$

This implication is falsified both empirically and theoretically. Normative principles—deontic, consequentialist, virtue-theoretic—do not by themselves generate behaviour, even in humans. Behaviour arises from the evaluative topology within which norms are interpreted. Watching-Eye cues generate deontic *expectations*, but the behavioural manifestation of these expectations is perturbed by γ_R at the level of attention, salience, and affective resonance. A normative rule cannot be enacted when the cognitive–affective substrate enabling its enactment is disrupted.

For these reasons, monolithic Machine Ethics fails. It collapses reflective and cognitive LoAs, ignores the topological structure linking salience to action, neglects the role of affect and social signal processing in moral cognition, and treats moral behaviour as rule-following rather than field-sensitive, dynamically realised evaluation.

8.4.1 Reconstructing Computational Morality: An Empirically Grounded Paradigm

If Machine Ethics fails because it begins with normative theory, the alternative must begin with *empirical structure*. The present thesis advances a methodological reversal:

> **Computational Morality** begins not by encoding principles, but by modelling the cognitive–affective architecture through which moral behaviour is produced and perturbed.

(1) Evaluative Topology as Generative Substrate Moral behaviour emerges from an evaluative manifold shaped by gradients of salience, attractor basins of affective resonance, normative invariants, and dispositional curvature [241]. Robotic presence is formalised as a perturbation operator:

$$\gamma_R : \mathcal{E} \to \mathcal{E}',$$

modifying attentional and affective weights and thereby predicting attenuation of prosocial behaviour without invoking rule-based computation.

(2) Level-of-Abstraction Discipline Normative theories enter as reflective structures operating at the normative LoA [25]. Deontology provides invariants, consequentialism gradients, virtue ethics dispositional metrics, sentimentalism affective vectors, contractualism justificatory equilibria, and particularism context-sensitive modulations. These structures constrain interpretation, not execution.

(3) Dispositional Ecologies as Moral Topologies The PCA–k-means clusters define dispositional geometries that shape evaluative trajectories: *Emotionally Reactive* (broad, shallow attractors), *Prosocial–Empathic* (steep affective gradients), and *Analytical–Structured* (narrow, stable valleys). Synthetic presence perturbs the field upstream of these differences [51, 34], revealing that moral behaviour is topologically sensitive rather than trait-determined.

8.4.2 Computational Morality as a Scientific Research Programme

The reconstructed paradigm transforms the methodological landscape of moral AI. Rather than engineering moral behaviour by encoding principles, *Computational Morality* aims to:

1. model the evaluative field governing moral behaviour;
2. identify perturbation operators introduced by artificial agents;
3. integrate normative theory as reflective constraint rather than behavioural generator;
4. and design artificial systems that stabilise, rather than distort, the evaluative field.

This paradigm extends Social Signal Processing [67, 73] and Affective Computing [68] by adding a normative dimension grounded not in abstract prescription but in empirically measurable topological structure.

In this sense, the robot in the experiment is not an ethical agent but an *evaluative perturbation device*. Its presence reveals the structural sensitivity of human moral cognition. A scientifically responsible programme of moral AI must begin from this insight: artificial agents shape the moral environment long before they act within it.

The next section consolidates these findings into a global synthesis, showing how the normative, cognitive, and topological architectures developed across the thesis converge in a unified model of moral perturbation and ethical interpretation.

8.5 Thesis-Wide Synthesis and Closing Reflections

Across its full argumentative trajectory, this thesis has advanced a single, unified claim: *human moral behaviour is structurally sensitive to the architecture of the perceptual–social environment, and synthetic presence—even when silent and non-sentient—is sufficient to reshape that structure.* This concluding section synthesises the theoretical, empirical, and normative strands developed throughout the work and articulates the implications for computation, moral psychology, and the ethics of artificial agents.

1. Moral Cognition is Field-Sensitive and Structurally Rich

The *Morality Primer* established that moral cognition is a distributed, multi-level, dynamically integrated system. Dual-process models, the Social Intuitionist Model, and empirical findings from social neuroscience converge on a view in which moral appraisal emerges from:

- rapid affective and attentional processes,
- controlled interpretive regulation,
- and an evaluative topology shaped by salience, affective resonance, and contextual cues.

This architecture is not neutral with respect to environmental perturbation. The field in which moral appraisal unfolds has curvature, gradients, attractors, and deformation potentials—all empirically traceable, neurocognitively plausible, and behaviourally measurable.

2. Levels of Abstraction and the Limits of Purely Normative Models

The *Ethical Cognition and Normative Foundations* chapter showed that ethical theory and moral psychology occupy distinct Levels of Abstraction. Normative theories do not function as generative behavioural models; their role is to articulate invariant, justificatory, or virtue-theoretic structures that constrain or interpret behaviour at a reflective LoA.

Machine Ethics has historically collapsed these orders, implementing deontic rules, utility functions, or evaluative labels as if they were cognitive operators. This thesis rejects that methodological inversion. Normative content becomes intelligible only when anchored in empirical structure; without such anchoring, computational morality risks degenerating into symbolic simulation devoid of psychological traction.

3. Empirical Evidence for Synthetic Moral Perturbation

Within this framework, the experiment plays a decisive role. It demonstrates that the presence of a humanoid robot:

- attenuates prosocial donation in a statistically supported manner,
- does so even under a strong moral cue (the Watching-Eye prime),
- and produces a uniform directional displacement across dispositional ecologies, albeit with variation in magnitude.

This attenuation is not reducible to personality differences, response bias, or explicit moral reasoning. The analysis shows that the robot functions as a perturbation operator γ_R that modifies the evaluative field *upstream* of trait-specific and deliberative processes. It acts on the conditions under which moral appraisal acquires behavioural force.

4. Dispositional Ecologies Reveal Structural, Not Idiosyncratic, Perturbation

The clustering analysis established three coherent dispositional ecologies:

- **Emotionally Reactive / Low-Structure**, exhibiting broad low-gradient attractors and high affective volatility;
- **Prosocial–Empathic / Warm–Sociable**, with steep empathic gradients and strong responsiveness to social cues;
- **Analytical–Structured / High-Systemizing**, with narrow, stable attractors shaped by deliberative integration.

Despite their divergent evaluative geometries, all clusters showed the same *direction* of moral displacement. This finding is decisive: it shows that the perturbation is field-level, not agent-level. The robot reshapes the evaluative manifold within which trajectories unfold, rather than interacting with any single cognitive disposition. This is the empirical signature of a *structural perturbator*.

5. Normative Interpretation of Structural Perturbation

The reconstructed normative frameworks illuminate the ethical significance of this empirical result:

- deontologically, γ_R disrupts the recognition of accountability cues implicit in the Watching-Eye stimulus;
- consequentially, it flattens the perceived payoff gradient of beneficence;
- virtuously, it weakens the stabilising force of prosocial dispositions;
- sentimentally, it dampens empathic vector fields that anchor reactive moral emotions;
- contractually, it disrupts justificatory visibility between moral agents;
- particularistically, it shifts the situational salience profile.

These converging interpretations reveal the central structural insight of the thesis: *the robot does not add a new norm; it shifts the evaluative conditions under which norms become behaviourally operative.*

6. Final Position of the Thesis

We may now return to the guiding hypotheses:

H1 — Evaluative Deformation *Confirmed.* The evaluative process f linking perception to action is systematically altered by synthetic presence.

H2 — Synthetic Normativity *Confirmed.* Synthetic agents acquire derivative normative force by altering the field of salience and accountability.

H3 — Synthetic Perturbation of Moral Inference *Confirmed.* The robot refracts the transition from moral appraisal to prosocial behaviour, attenuating the expressive force of the Watching-Eye cue.

Accordingly, the thesis takes the following stand:

Final Thesis Position (Definitive)

Human moral agency is not internally autonomous. It is structurally coupled to the perceptual–social field in which it is embedded. Synthetic agents, even when lacking sentience, intentionality, or communicative acts, act as ***modulators of that field****. They reshape attentional gradients, dampen empathic resonance, and deform the topological structures through which moral appraisal acquires behavioural expression. Moral displacement under synthetic presence is therefore not a behavioural curiosity, but a structural fact about the architecture of moral cognition.*

7. Implications for the Future of Computational Morality

This final insight reshapes the methodological landscape. Artificial agents cannot be treated as moral subjects but must be understood as **moral modifiers**: entities whose design implicitly reconfigures the evaluative field. Future research in computational morality must therefore move beyond rule encoding and value annotation toward a structural science of moral environments, moral salience, and field-sensitive interaction.

In this sense, the thesis does not simply present an experimental result; it offers a new conceptual foundation for the empirical and ethical study of artificial agents. It reorients the field toward a *topological*, *empirically grounded*, and *LoA-disciplined* understanding of moral cognition—one capable of addressing the forms of synthetic presence that will increasingly populate human social life.

With this synthesis, the thesis closes. Its central claim is now complete: moral behaviour is field-dependent, and synthetic presence reshapes that field.

9. Conclusion

9.1 Returning to the Question: What This Thesis Has Shown

The Introduction framed this work around a deceptively simple research problem: *does the mere presence of a humanoid robot alter the transformation of morally salient cues into moral action?* Formally, this was articulated in Question 5.1 and operationalised through three hypotheses (H1–H3) anchored in the evaluative mapping

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

which models how environmental cues, dispositional structure, and synthetic presence jointly shape behavioural output. The core task of the thesis was to determine whether this mapping is susceptible to perturbation, and what such susceptibility reveals about the architecture of human moral cognition.

Across the chapters that followed, this problem—posed abstractly in the Introduction—developed into an empirical and conceptual investigation of how humans register moral salience in the presence of non-human bodies. Drawing on contemporary models of moral psychology that emphasise intuitive, affect-laden appraisals over reflective deliberation [16, 31, 147, 17], the experimental work examined whether NAO influences the early evaluative stages that precede conscious reasoning. The design therefore situated participants within a minimally structured moral environment, introducing γ_R as a silent, embodied presence whose influence could arise only through its perceptual affordances.

With this concluding chapter, we return to the commitments articulated at the start. The results now allow us to say, with the empirical grounding that the Introduction could only anticipate, that the evaluative process is indeed sensitive to synthetic presence.

H1 (Evaluative Deformation) is supported: the expected behavioural output under $\Sigma \cup \mathscr{R}$ diverges from that under Σ alone.

H2 (Synthetic Normativity) is refined: NAO does not generate new normative affordances but modulates the salience of existing ones through its informational and embodied profile.

H3 (Synthetic Perturbation of Moral Inference) is supported: the influence of γ_R manifests in the perceptual–affective transition from moral cue to action, not in explicit deliberation.

Moreover, the evaluative mapping $f(\alpha_E, \beta_C, \gamma_R)$ predicted a structured form of dispositional heterogeneity—an expectation borne out in the latent trait ecologies recovered via clustering, where perturbation effects concentrated within the Prosocial–Empathic regime and remained muted or absent elsewhere.

Crucially, the thesis has treated these findings with the epistemic caution appropriate for a subtle effect in a modest dataset. Bayesian estimation explicitly accommodates uncertainty arising from zero-inflated donation data and uneven cluster sizes: the posterior for the donation difference is skewed toward attenuation, yet reflects the heterogeneity of cognitive–affective architectures rather than compressing evidence into a binary verdict. This probabilistic contour strengthens the interpretive claim: when moral cognition operates intuitively, ambiguity in the environment should appear as ambiguity in the data.

This leads to the broader significance of the work. The results suggest that:

> *Human moral appraisal is structurally sensitive to the perceptual ecology in which it unfolds.*

Moral behaviour emerges from a sequence of cognitive–affective transitions shaped by what is noticed, foregrounded, or rendered ambiguous. Synthetic presence makes this sensitivity visible: even a minimally expressive artificial body can redirect the flow of evaluative attention, thereby altering the probability that moral salience becomes action.

In bringing these strands together, this chapter begins where the Introduction left off: with the claim that artificial systems participate in human moral life long before they reason, decide, or speak. What the experiment now shows is that their influence arises from how they reconfigure the perceptual and affective scaffolds through which moral meaning is formed.

The remainder of the Conclusion develops this insight, situating it within social robotics, affective computing, and the philosophy of moral cognition, and drawing the thesis toward its final synthesis.

9.2 Contributions to Human–Robot Interaction, Affective Computing, and Moral Cognition

The empirical and conceptual results developed throughout this thesis amount to a set of contributions that cut across three adjacent domains: human–robot interaction (HRI), affective computing, and the study of moral cognition.

While each field approaches social behaviour from a distinct methodological tradition, the present work shows that the phenomenon under investigation—the attenuation of prosocial action under synthetic co-presence—sits precisely at their intersection. The following synthesis articulates these contributions in a form that reflects the structural unity of this research.

9.2.1 Contribution to Human–Robot Interaction

Within HRI, the dominant research programmes have traditionally focused on interactional behaviours: communication, signalling, collaboration, engagement,trust, and alignment [301, 223, 224, 302, 225]. The present study contributes a different insight:

> *Interaction is not required for a humanoid robot to exert a measurable influence on human behaviour*

NAO's silent presence reshaped the evaluative conditions under which participants interpreted a morally salient cue in experimental setting. This demonstrates that robots participate in human moral environments not only through explicit social action but also through their perceptual affordances, spatial occupation, and ontological ambiguity.

This finding positions synthetic presence as an *environmental factor* in HRI—a contributor to the structure of the evaluative field rather than a node in an interactional sequence. The notion of robotic co-presence as a "semiotic operator" opens conceptual space for a new class of HRI effects: those that emerge upstream of explicit social behaviour, and which operate through shifts in attention, salience, and interpretive framing.

9.2.2 Synthetic Presence and Floridi's Account of Moral Agency

A central question naturally arises from the empirical findings of this thesis:

> *if NAO reshapes the evaluative conditions under which morally salient cues are interpreted under experimental settings, what does this imply for its moral status within Floridi's framework of agency, patiency, and informational relevance?*

Addressing this question clarifies both the conceptual footing of the results and their broader implications for social robotics and machine ethics.

Floridi's theory of moral agency distinguishes between three classes of entities: (i) *moral agents*, capable of performing morally qualifiable actions; (ii) *moral patients*, capable of moral harm or benefit; and (iii) a broader class of *morally relevant informational objects*, whose properties modulate the normative landscape without possessing agency or interests [145, 198].

Within this taxonomy, NAO is neither an agent nor a patient: it issues no commands, performs no autonomous decisions, and has no capacity for norm-responsive deliberation or moral vulnerability.

Yet the empirical results show that NAO is not normatively inert. The robot exerts a measurable influence on the evaluative transformation

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

not by acting or reasoning, but by altering the *informational environment* in which human agents interpret moral cues. In Floridi's terms, NAO functions as a *morally relevant artefact*: its perceptual affordances—embodied form, gaze, spatial presence, and subtle motion—shape the *potential* salience landscape at the operative Level of Abstraction [26, 198].

At this LoA, participants do not encounter the robot as source code or computation. They encounter it as a *semiotic body* [139, 146]: an entity that appears socially expressive while lacking the behavioural depth of a human agent. This ontological ambiguity, we claim, is precisely what enables the perturbation observed in the experiment. The robot does not generate new norms (refining H2), nor does it provide explicit reasons for action; instead, it *modulates the salience of existing cues*, bending the intuitive trajectory through which the Watching-Eye stimulus gains behavioural force [6, 5].

Situating NAO as a morally relevant informational object strengthens the central conclusion. It shows why synthetic presence can reshape moral behaviour without approaching the threshold of moral agency, and it clarifies a conceptual point often missed in Machine Ethics:

> *artificial systems can exert morally significant influence prior to, and independent of, any capacity for ethical reasoning.* [52, 23]

Their normative impact arises from the environmental and semiotic roles they occupy within human cognitive ecologies [198, 146]. In this sense, Floridi's framework does not merely accommodate the findings; it provides the conceptual architecture that renders them intelligible. NAO's influence is not paradoxical but the predictable consequence of how informational artefacts participate in—and subtly deform—the evaluative landscapes through which human moral cognition operates.

9.2.3 Contribution to Affective Computing and Social Signal Processing

In affective computing and social signal processing, a central ambition is to infer latent cognitive–affective processes from observable behavioural traces. The experiment presented in this thesis provides an empirical demonstration that moral behaviour—here, prosocial donation—can serve as such a behavioural indicator. The attenuation observed under synthetic presence suggests that moral evaluations, although not directly observable, leave systematic behavioural traces when the evaluative topology is perturbed. This is consistent with the foundational premise of affective computing that latent states become inferable through patterned interactions between behaviour and context [68, 303, 63, 304].

A methodological insight follows:

> *latent dispositional architecture can be recovered from psychometric and behavioural data in a way that reveals differential susceptibility to contextual perturbation.*

The three dispositional ecologies identified through clustering (Emotionally Reactive, Prosocial–Empathic, Analytical–Structured) illustrate how personality configurations form distinct regions in evaluative space. These ecologies show that affective modulation by artificial systems is a *structure-sensitive* process, linking affective computing's interest in latent state inference with HRI's concern for the social affordances of synthetic presence. Comparable patterns in social signal processing show that stable dispositional profiles shape responsiveness to contextual cues and interpersonal signals [67, 169, 86].

Finally, the Bayesian modelling framework demonstrates how uncertainty—a constitutive property of both affective states and behavioural data—can be represented as a structured epistemic gradient rather than statistical noise. This aligns Bayesian inference with the broader aims of affective computing: to quantify latent evaluative processes under contextual variability, in continuity with hierarchical Bayesian approaches in computational cognitive science [305, 306].

9.2.4 Contribution to Moral Cognition Research

Within moral psychology, the present findings contribute evidence for an intuitionist and ecological interpretation of moral judgement. Consistent with leading models of intuitive moral appraisal and dual-process architectures [16, 31, 147, 46, 17], the study shows that moral behaviour is shaped by the perceptual and affective scaffolds through which salient cues are first registered. The attenuation effect emerges *prior* to deliberation, within the cognitive–affective space where intuitive appraisals and affect-driven evaluations are formed. This aligns with work demonstrating that moral cognition is sensitive to subtle environmental modulation—implicit monitoring, gaze cues, ambient social presence [1, 2, 6, 5]—and extends these findings to *synthetic* social affordances [35, 188, 34].

A central contribution lies in the demonstration that moral susceptibility is *trait-contingent*. The Prosocial–Empathic profile, characterised by high interpersonal attunement and affective resonance, exhibited the clearest attenuation under NAO's presence, while the Analytical–Structured and Emotionally Reactive profiles did not. This pattern empirically supports the prediction derived from the formalism $f(\alpha_E, \beta_C, \gamma_R)$: that dispositional architecture modulates how moral salience undergoes evaluative transformation. The finding is consistent with research showing that empathy, agreeableness, and prosocial orientation shape sensitivity to social cues and moral demands [86, 307, 18], while extending this literature by demonstrating that such sensitivity generalises to *synthetic* embodied presence.

Taken together, the results show that moral cognition is both intuitive and ecologically embedded: structured by the interplay between perceptual environments and latent evaluative topologies. Synthetic presence provides a novel perturbation of this evaluative field, revealing the dispositional contours through which moral meaning becomes behaviour.

9.2.5 Integrative Contribution: A Unified Field-Theoretic Approach

The broader contribution of this research lies in unifying these three domains within a single explanatory framework. By treating moral behaviour as the output of an evaluative function influenced by environmental cues, dispositional architecture, and synthetic presence, the thesis provides a consistent conceptual vocabulary for explaining effects that span psychological, computational, and robotic contexts. This framework is field-theoretic in a strict sense: it arises directly from the formal decomposition $f(\alpha_E, \beta_C, \gamma_R)$ and models moral appraisal as movement through a structured evaluative landscape.

This account aligns with long-standing findings in moral psychology that behaviour emerges from affective appraisal and intuitive processing [16, 31, 18], with Social Signal Processing research showing that social cues reorganise attentional and evaluative structures [67, 63, 6], and with HRI evidence that robotic presence modulates human social and moral behaviour even in minimal-interaction settings [33, 34, 51, 35]. Crucially, Floridi's Levels of Abstraction supply the conceptual hinge: synthetic presence exerts influence at the perceptual LoA, independent of its internal architecture.

Across HRI, affective computing, and moral cognition, the central insight is the same:

> *synthetic presence need not act, speak, or decide to shape moral behaviour; it need only reorganise the evaluative conditions under which moral salience is processed.*

This insight establishes a theoretical bridge between empirical measurement, computational modelling, and philosophical analysis. It connects affective-computational approaches concerned with latent evaluative states [68], SSP models of multimodal social inference [63], and ecological accounts of context-sensitive moral judgement [16, 46].

Taken together, these strands converge on a unified field-theoretic perspective in which artificial agents function not as loci of ethical reasoning but as operators on the evaluative topology through which moral meaning becomes behavioural output.

9.2.6 A Unified Explanatory Structure Rather Than Three Independent Literatures

One might ask whether the thesis merely juxtaposes contributions from three separate domains—moral psychology, Floridian information ethics, and social robotics—or whether these strands genuinely converge into a single explanatory framework. The answer, made clear only at the end of the research journey, is that the thesis was never structured as three parallel literatures. It was structured around a single phenomenon that demands all three.

The core explanatory object is the evaluative transformation

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

which specifies how environmental cues, dispositional architectures, and synthetic presence jointly shape moral behaviour. Each literature supplies a different layer of this model: moral psychology identifies the intuitive and affective mechanisms that generate evaluative trajectories and explains why trait–contingent modulation should occur; Floridi's information ethics provides the ontological and epistemic conditions under which synthetic entities acquire moral relevance at a Level of Abstraction; and social robotics and Social Signal Processing provide the empirical and computational domains in which such perturbations become observable, measurable, and theoretically tractable.

Seen from this vantage point, the "field-theoretic" perspective is not a metaphor but the necessary generalisation of the mapping f. Moral cognition unfolds in a structured space of salience and attention, and synthetic presence acts as an operator that subtly reconfigures the geometry of that space. None of the three literatures could explain this phenomenon alone: moral psychology without LoA lacks an ontology of synthetic presence; information ethics without empirical grounding lacks a mechanism of perturbation; and HRI without a field-theoretic model lacks an account of why minimal presence should matter at all.

In this respect, the thesis does not merely connect disparate literatures; it reorganises them around a common evaluative architecture. The experimental results show that artificial agents participate in human moral environments not through ethical reasoning, but through the way their perceptual and ontological profiles modulate the topology through which moral meaning becomes action. This unification is the conceptual kernel that ties the entire research programme together.

9.3 Final Synthesis and Closing Reflections

At the end of this research journey, the central phenomenon around which the thesis was constructed can be stated with clarity: *synthetic presence alters the conditions under which human beings register, evaluate, and act upon moral salience.* What began as a question about a quiet, non-interactive humanoid robot has unfolded into a broader insight about the structure of moral cognition and the ethical texture of technologically saturated environments.

USA LA PAROLA "SHOW" PIUTTOSTO CHE "DEMONSTRATES". LA STATISTICA CHE HAI USATO NON È STRUMENTO PER DIMOSTRAZIONE

The empirical work demonstrated that NAO's presence does not introduce new moral reasons, nor does it engage participants in explicit social interaction. Instead, it reshapes the evaluative conditions under which moral cues acquire behavioural force. This perturbation is subtle, probabilistic, and contingent upon latent dispositional architecture—yet empirically robust. Across inferential contrasts, cluster-specific analyses, and Bayesian estimation, the evidence converges on the same structural interpretation: artificial systems participate in human moral situations by modulating the perceptual and affective scaffolds from which intuitive moral appraisals develop.

What makes this more than an isolated result is the conceptual synthesis that underwrites it. The thesis has argued that moral behaviour is not a direct output of rules or principles; it is the trajectory of a cognitive–affective system shaped by attention, affect, expectation, and salience. Contemporary models of moral psychology have long emphasised the primacy of intuitive evaluation [16, 31, 17]. Floridi's Levels of Abstraction provide the correct ontological lens for specifying where synthetic entities exert influence. HRI and Affective Computing supply the methodological tools through which such influence can be observed and modelled. When joined together, these literatures do not merely complement one another—they reveal that the phenomenon under investigation requires all of them to be intelligible [308, 309].

A key outcome of this integration is the field-theoretic interpretation articulated throughout the thesis. The evaluative mapping

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

is not a formal artefact appended to the data; it is the conceptual structure that makes sense of how environmental cues, dispositional architectures, and artificial systems cooperate to shape moral behaviour. Moral appraisal unfolds within a dynamic field of salience. Synthetic presence acts as an operator on that field—not by issuing commands but by shifting the local geometry through which moral meaning is formed. The experimental results show that even minimal artificial bodies can introduce such a shift. ~~This finding is modest in magnitude but profound in implication.~~

SEMBRA UN PO' PRETENZIOSO

The ethical considerations developed in Chapter 8 sharpen this point. Traditional Machine Ethics asks how to embed principles into machines. Yet the empirical evidence here shows that artificial systems shape moral life not through principles but through presence: through their semiotic affordances, their perceived ontology, and the attentional demands they impose on human cognitive ecologies. In Floridi's framework, NAO is not a moral agent; it is a morally relevant informational object. Its influence is neither agential nor normative in itself, but environmental: it alters the field in which human agents construct their moral understanding. This is not a limit case or a marginal curiosity. It is a demonstration that ethical relevance emerges from the interaction between artefacts and the conditions of human moral cognition, long before questions of machine agency arise.

From this vantage point, the contribution of the thesis becomes unmistakable. The work provides an empirically grounded demonstration that artificial systems—humanoid robots today, pervasive AI systems tomorrow—can reshape moral cognition through their informational presence alone. It offers a formal architecture for modelling such influence, a set of empirical methods for detecting it, and a conceptual framework for understanding why it occurs. It also clarifies that the normative stakes of synthetic presence lie not merely in what machines may eventually *do*, but in how they silently reorganise the evaluative landscape in which human moral behaviour unfolds.

LA PAROLA DIMOSTRAZIONE E' TROPPO FORTE. SUGGERISCO INDICAZIONE.

This brings us back, finally, to the thesis's animating question. Yes—synthetic presence can perturb the evaluative transformation through which moral salience becomes moral action. It can do so without interacting, without expressing norms, without engaging in dialogue, and without crossing the threshold of moral agency. It is enough that it is *perceived*. This insight opens the way for a new research programme in moral AI: one that takes seriously the structure of moral cognition, the ecological nature of moral environments, and the field-level effects of artificial systems embedded within them.

If this thesis has shown anything, it is that the future of moral AI will not begin with building artificial moral agents. It will begin with understanding how artificial systems already shape human moral life. The rest—ethical design, moral alignment, normative governance—must follow from that foundational fact. The task now is no longer to ask whether such influence exists, but to learn how to measure it, model it, anticipate it, and ultimately, to govern the moral topologies we are already co-constructing with our synthetic companions.

This is the horizon toward which the work now points. The thesis closes here, but the evaluative field it uncovers is only beginning to take shape.

Bibliography

[1] 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.

[2] 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.

[3] M. Ernest-Jones, D. Nettle, and M. Bateson, "Effects of eye images on everyday cooperative behavior: A field experiment," *Evolution and Human Behavior*, vol. 32, no. 3, pp. 172–178, 2011.

[4] 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.

[5] G. E. Dear, K. Dutton, and E. Fox, "The watching-eyes effect in the dictator game: A meta-analysis," *Evolution and Human Behavior*, vol. 40, no. 3, pp. 271–284, 2019.

[6] 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.

[7] 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.

[8] E. Gleichgerrcht and L. Young, "Low empathic concern predicts utilitarian moral judgment," *Cognition*, vol. 126, no. 3, pp. 364–372, 2013.

[9] J. Haidt, "The emotional dog and its rational tail: A social intuitionist approach to moral judgment," *Psychological Review*, vol. 108, no. 4, pp. 814–834, 2001.

[10] 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.

[11] 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.

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

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

[14] P. Foot, *Natural Goodness.* Oxford: Oxford University Press, 2001.

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

[16] 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.

[17] 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.

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

[19] 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.

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

[21] 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.

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

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

[24] M. Guarini, "Computational neural modeling and the philosophy of ethics: Reflections on the particularism-generalism debate," *Cambridge University Press*, 2006.

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

[26] L. Floridi, *The Philosophy of Information.* Oxford: Oxford University Press, 2011.

[27] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and Information Technology*, vol. 7, no. 3, pp. 149–155, 2005.

[28] K. Arkoudas and S. Bringsjord, "Toward ethical robots via mechanized deontic logic," in *Machine Ethics: AAAI Fall Symposium*, (Menlo Park, CA), pp. 17–23, AAAI Press, 2005.

[29] A. F. T. Winfield, M. Ortega, and R. Harper, "The ethical black box: An ai safety concept to facilitate ethics review and accountability," *IEEE Technology and Society Magazine*, vol. 38, no. 3, pp. 62–69, 2019.

[30] M. Anderson and S. L. Anderson, "Robot be good: A call for ethical autonomous machines," *Scientific American*, vol. 303, no. 4, pp. 72–77, 2010.

[31] 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.

[32] J. Prinz, *The Emotional Construction of Morals*. Oxford: Oxford University Press, 2007.

[33] D. Kuchenbrandt, F. Eyssel, S. Bobinger, and M. Neufeld, "Minimal group-maximal effect? evaluation and anthropomorphization of the humanoid robot nao," in *International conference on social robotics*, pp. 104–113, Springer, 2011.

[34] 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.

[35] 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.

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

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

[38] Z. Jin, H. Zhang, T. Ge, and M. Zeng, "Moral foundations of large language models," *arXiv preprint arXiv:2205.12329*, 2022.

[39] N. Scherrer, E. Clark, and N. A. Smith, "Evaluating moral reasoning in large language models," *arXiv preprint arXiv:2306.00030*, 2023.

[40] A. Nguyen *et al.*, "Moral self-correction for large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[41] G. Aher and R. Arriaga, "Using large language models to simulate human moral decision-making," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

[42] P. Charlton and D. Danks, "Large language models show human-like moral dynamics," *arXiv preprint arXiv:2308.13129*, 2023.

[43] D. Hendrycks *et al.*, "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2021.

[44] D. Emelin *et al.*, "Moral foundations in large language models: A case study on moralclassification," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–16, 2023.

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

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

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

[48] J. Decety and M. Meyer, "From emotion resonance to empathic understanding: A social developmental neuroscience account," *Development and psychopathology*, vol. 20, no. 4, pp. 1053–1080, 2008.

[49] J. Zaki and K. N. Ochsner, "The neuroscience of empathy: Progress, pitfalls, and promise," *Nature Neuroscience*, vol. 15, no. 5, pp. 675–680, 2012.

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

[51] 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.

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

[53] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos, "Artificial moral agents: A survey of the current status," *Science and Engineering Ethics*, vol. 26, no. 2, pp. 501–532, 2020.

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

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

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

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

[58] M. Andrus, M. Spitzer, *et al.*, "What do models know about morality? a review of ethical reasoning in ai," *arXiv preprint arXiv:2305.15765*, 2023.

[59] A. Kasirzadeh and I. Gabriel, "The mirage of moral agency in large language models," *Philosophy & Technology*, vol. 37, no. 1, pp. 1–26, 2024.

[60] 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.

[61] J. Gardner and et al., "Models that write like moral agents are not moral agents," *AI & Society*, 2024. Forthcoming.

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

[63] 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.

[64] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The effect of robot personality on human-robot interaction," in *Proceedings of the 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 141–142, ACM, 2011.

[65] 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.

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

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

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

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

[70] M. Slote, *Moral Sentimentalism*. Oxford: Oxford University Press, 2010.

[71] 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.

[72] M. Ekström, "Do watching eyes affect charitable giving? evidence from a field experiment," *Experimental Economics*, vol. 15, no. 3, pp. 530–546, 2012.

[73] 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.

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

[75] 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.

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

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

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

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

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

[81] J. Annas, *Intelligent Virtue.* Oxford: Oxford University Press, 2011.

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

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

[84] C. Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms.* New York: Oxford University Press, 2016.

[85] S. Baron-Cohen, *The Essential Difference: The Truth about the Male and Female Brain.* London: Penguin, 2003.

[86] M. M. Habashi, W. G. Graziano, and A. E. Hoover, "Searching for the prosocial personality: A big five approach to linking personality and prosocial behavior," *Personality and Social Psychology Bulletin*, vol. 42, no. 9, pp. 1177–1192, 2016.

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

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

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

[90] A. Shenhav, S. Musslick, F. Lieder, W. Kool, T. L. Griffiths, J. D. Cohen, and M. M. Botvinick, "Toward a rational and mechanistic account of mental effort," *Annual Review of Neuroscience*, vol. 40, pp. 99–124, 2017.

[91] 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.

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

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

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

[95] 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.

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

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

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

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

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

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

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

[103] 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.

[104] 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.

[105] 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.

[106] 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.

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

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

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

[114] 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.

[115] 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.

[116] 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.

[117] 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.

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

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

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

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

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

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

[135] 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.

[136] 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.

[137] 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.

[138] 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.

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

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

[141] 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.

[142] 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.

[143] 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.

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

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

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

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

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

[149] 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.

[150] S. Baron-Cohen, "Autism: The empathizing?systemizing (e?s) theory," *Trends in Cognitive Sciences*, vol. 13, no. 6, pp. 274–280, 2009.

[151] O. P. John, E. M. Donahue, and R. L. Kentle, "The big five inventory?versions 4a and 54," tech. rep., University of California, Berkeley, Institute of Personality and Social Research, 1991.

[152] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research* (L. A. Pervin and O. P. John, eds.), pp. 102–138, Guilford Press, 1999.

[153] 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.

[154] 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.

[155] 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.

[156] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many? people apply different moral norms to human and robot agents," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 117–124, IEEE, 2015.

[157] T. Komatsu, "Japanese students apply same moral norms to humans and robot agents: Considering a moral hri in terms of different cultural and academic backgrounds," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 457–458, IEEE, 2016.

[158] A. Wakabayashi, S. Baron-Cohen, S. Wheelwright, N. Goldenfeld, J. Delaney, D. Fine, and R. Smith, "Development of short forms of the empathy quotient (eq-short) and the systemizing quotient (sq-short)," *Personality and Individual Differences*, vol. 41, no. 5, pp. 929–940, 2006.

[159] N. Goldenfeld, S. Baron-Cohen, and S. Wheelwright, "Empathizing and systemizing: A cross-cultural investigation," *Personality and Individual Differences*, vol. 39, no. 1, pp. 173–183, 2005.

[160] J. Lawson, S. Baron-Cohen, and S. Wheelwright, "Empathising and systemising in adults with and without asperger syndrome: A factor analysis," *Journal of Autism and Developmental Disorders*, vol. 34, no. 3, pp. 301–310, 2004.

[161] A. Konovalov and I. Krajbich, "Revealed prioritization using a novel economic task," *Journal of Experimental Psychology: General*, vol. 145, no. 6, pp. 802–825, 2016.

[162] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: a meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

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

[164] S. Baron-Cohen, "Autism and the empathizing-systemizing (es) theory," *Developmental social cognitive neuroscience*, pp. 125–138, 2009.

[165] 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.

[166] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research* (L. A. Pervin and O. P. John, eds.), pp. 102–138, New York: Guilford Press, 1999.

[167] R. R. McCrae and P. T. Costa, "The five-factor theory of personality," *Handbook of Personality: Theory and Research*, pp. 159–181, 2008.

[168] W. G. Graziano, N. Eisenberg, and R. M. Tobin, "Agreeableness and helping behavior: A meta-analysis," *Psychological Bulletin*, vol. 119, no. 3, pp. 371–394, 1996.

[169] J. Banks, "Theory of mind in social robots: replication of five established human tests," *International Journal of Social Robotics*, vol. 12, no. 2, pp. 403–414, 2020.

[170] C. Thompson-Booth, E. Viding, L. C. Mayes, and H. J. V. Rutherford, "Here's looking at you: Emotional faces predict eye-gaze behaviors in parents and non-parents," *Social Neuroscience*, vol. 9, no. 6, pp. 605–613, 2014.

[171] C. D. Batson, *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

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

[173] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, pp. 785–791, 2003.

[174] F. Warneken and M. Tomasello, "Altruistic helping in human infants and young chimpanzees," *Science*, vol. 311, no. 5765, pp. 1301–1303, 2006.

[175] J. Andreoni, "Impure altruism and donations to public goods: A theory of warm-glow giving," *The Economic Journal*, vol. 100, no. 401, pp. 464–477, 1990.

[176] H. Gintis, "Strong reciprocity and human sociality," *Journal of Theoretical Biology*, vol. 206, no. 2, pp. 169–179, 2000.

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

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

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

[180] 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.

[181] 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.

[182] 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.

[183] 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.

[184] 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.

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

[186] 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.

[187] 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.

[188] 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.

[189] 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.

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

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

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

[193] 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.

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

[195] 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.

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

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

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

[199] 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.

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

[201] 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.

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

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

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

[205] 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.

[206] 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.

[207] D. J. Gunkel, *The Machine Question: Critical Perspectives on AI, Robots, and Ethics.* Cambridge, MA: MIT Press, 2012.

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

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

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

[211] 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.

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

[213] 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.

[214] A. Vinciarelli, A. Esposito, M. Tayarani, G. Roffo, F. Scibelli, F. Perrone, and D.-B. Vo, "We are less free than how we think: Regular patterns in nonverbal communication," in *Multimodal Behavior Analysis in the Wild*, pp. 269–288, Elsevier, 2019.

[215] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411–437, 2020.

[216] E. Hutchins, *Cognition in the Wild.* MIT Press, 1995.

[217] A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension.* Oxford University Press, 2008.

[218] L. Floridi, "The ethics of artificial intelligence: Large language models and the new semantic environment," *Philosophy & Technology*, 2024.

[219] L. Speri and M. Alfano, "Moral ecologies and the ethics of artificial agents," *Ethics and Information Technology*, 2023.

[220] R. E. Nisbett and T. D. Wilson, "Telling more than we can know: Verbal reports on mental processes," *Psychological Review*, vol. 84, no. 3, pp. 231–259, 1977.

[221] T. D. Wilson, *Strangers to Ourselves: Discovering the Adaptive Unconscious.* Harvard University Press, 2002.

[222] J. Decety and J. M. Cowell, "The complex relation between morality and empathy," *Trends in Cognitive Sciences*, vol. 17, no. 7, pp. 337–339, 2013.

[223] C. Breazeal, "Toward sociable robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 167–175, 2003.

[224] J. S. Lee and S. Kiesler, "Human mental models of humanoid robots," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 586–593, 2010.

[225] K. Fischer, "Interpersonal alignment in human–robot interaction," in *Proceedings of the 7th Annual ACM/IEEE International Conference on Human–Robot Interaction*, pp. 1–2, 2011.

[226] G. Gigerenzer, *Rationality for Mortals.* Oxford University Press, 2008.

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

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

[229] J. Knobe, "Person as scientist, person as moralist," in *Explanation and Cognition* (P. A. Machamer and M. Thagard, eds.), pp. 119–136, MIT Press, 2010.

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

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

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

[233] 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.

[234] 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.

[235] N. Lemos, *An Introduction to the Theory of Knowledge.* Cambridge University Press, 2nd ed., 2020.

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

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

[238] 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.

[239] 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.

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

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

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

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

[244] I. Kant, *Groundwork of the Metaphysics of Morals.* Cambridge University Press, 1785. Modern edition.

[245] Aristotle, *Nicomachean Ethics.* Oxford University Press, 2000. Trans. Terence Irwin.

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

[247] A. Smith, *The Theory of Moral Sentiments.* Cambridge University Press, 2002. Originally 1759.

[248] R. J. Wallace, *The View From Here: On Affirmation, Attachment, and the Limits of Regret.* Oxford University Press, 2012.

[249] P. Foot, "The problem of abortion and the doctrine of double effect', in her virtues and vices," *Berkeley and Los Angeles: University ofCalifornia Press. FootThe Problem of Abortion and the Doctrine of the Double Effect19Virtues and Vices1978*, pp. 19–32, 1978.

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

[251] P. Railton, "Moral realism," *Philosophical Review*, vol. 95, no. 2, pp. 163–207, 1984.

[252] J. Prinz, *Gut Reactions: A Perceptual Theory of Emotion.* Oxford University Press, 2004.

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

[254] 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.

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

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

[257] A. Smith, *The Theory of Moral Sentiments*. Cambridge: Cambridge University Press, 1759. Edited by D. D. Raphael and A. L. Macfie (1976 edition).

[258] S. Nichols, *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press, 2004.

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

[260] E. Morscher, "The definition of moral dilemmas: A logical confusion and a clarification," *Ethical theory and moral practice*, vol. 5, no. 4, pp. 485–491, 2002.

[261] D. Francey and R. Bergmüller, "Images of eyes enhance investments in a real-life public good," *PLoS One*, vol. 7, no. 5, p. e37397, 2012.

[262] 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.

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

[264] J. Carpenter, M. Davis, S. Erwin, and J. E. Young, "Functional and social roles in human–robot interaction: Exploring the effects of robot appearance and task," *Journal of Human-Robot Interaction*, vol. 5, no. 2, pp. 25–49, 2016.

[265] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *Proceedings of the 4th ACM/IEEE International Conference on Human?Robot Interaction*, pp. 69–76, ACM, 2009.

[266] H. Admoni and B. Scassellati, "Social eye gaze in human?robot interaction: A review," *Journal of Human?Robot Interaction*, vol. 6, no. 1, pp. 25–63, 2017.

[267] S. Krach, F. Hegel, B. Wrede, G. Sagerer, G. Bente, and T. Kircher, "Can machines think? interaction and perspective taking with robots investigated via fmri," *PLoS ONE*, vol. 3, no. 7, p. e2597, 2008.

[268] J. A. Bargh and T. L. Chartrand, "The unbearable automaticity of being.," *American psychologist*, vol. 54, no. 7, p. 462, 1999.

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

[270] J. Griffin, *Well-Being.* Oxford: Oxford University Press, 1986.

[271] M. Stocker, *Plural and Conflicting Values.* Oxford: Oxford University Press, 1990.

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

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

[274] P. Lin, G. Bekey, and K. Abney, "Autonomous military robotics: Risk, ethics, and design," tech. rep., California Polytechnic State Univ San Luis Obispo, 2008.

[275] K. Atkinson and T. Bench-Capon, "Action-based alternating transition systems for arguments about action," in *AAAI*, vol. 7, pp. 24–29, 2007.

[276] K. Atkinson and T. Bench-Capon, "Addressing moral problems through practical reasoning," in *International workshop on deontic logic and artificial normative systems*, pp. 8–23, 2006.

[277] M. Hjelmblom, *Deontic action-logic multi-agent systems in Prolog.* Högskolan i Gävle, 2008.

[278] A. Horn, "On sentences which are true of direct unions of algebras," *The Journal of Symbolic Logic*, vol. 16, no. 1, pp. 14–21, 1951.

[279] M. H. Van Emden and R. A. Kowalski, "The semantics of predicate logic as a programming language," *Journal of the ACM (JACM)*, vol. 23, no. 4, pp. 733–742, 1976.

[280] A. Saptawijaya and L. M. Pereira, "Towards modeling morality computationally with logic programming," in *International Symposium on Practical Aspects of Declarative Languages*, pp. 104–119, Springer, 2014.

[281] A. R. Honarvar and N. Ghasem-Aghaee, "An artificial neural network approach for creating an ethical artificial agent," in *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA)*, pp. 290–295, 2009.

[282] C. Battaglino, R. Damiano, and L. Lesmo, "Emotional range in value-sensitive deliberation," in *AAMAS International conference on Autonomous Agents and Multi-Agent Systems*, vol. 2, pp. 769–776, 2013.

[283] M. Sergot, "Action and agency in norm-governed multi-agent systems," in *International Workshop on Engineering Societies in the Agents World*, pp. 1–54, Springer, 2007.

[284] R. Montague and R. H. Thomason, "Formal philosophy. selected papers of richard montague," *Erkenntnis*, vol. 9, no. 2, 1975.

[285] R. Carnap, *Introduction to symbolic logic and its applications.* Courier Corporation, 2012.

[286] L. M. Pereira and A. Saptawijaya, "Modeling morality with prospective logic," *Cambridge University Press*, 2007.

[287] "The problem of machine ethics in artificial intelligence," *AI and SOCIETY*, vol. 35, no. 1, pp. 103–111, 2020.

[288] J. McDermid, V. C. Muller, T. Pipe, Z. Porter, and A. Winfield, "Ethical issues for robotics and autonomous systems," 2019.

[289] D. Howard and I. Muntean, "Artificial moral cognition: moral functionalism and autonomous moral agency," in *Philosophy and computing*, pp. 121–159, Springer, 2017.

[290] M. Pantic and A. Vinciarelli, "Social signal processing," *The Oxford handbook of affective computing*, p. 84, 2014.

[291] R. W. Picard, *Affective computing.* MIT press, 2000.

[292] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford handbook of affective computing.* Oxford Library of Psychology, 2015.

[293] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* Cambridge, MA: Perseus Books, 1994.

[294] R. D. Beer, "A dynamical systems perspective on agent–environment interaction," *Artificial Intelligence*, vol. 72, no. 1–2, pp. 173–215, 1995.

[295] L. B. Smith and E. Thelen, "Development as a dynamic system," *Trends in Cognitive Sciences*, vol. 7, no. 8, pp. 343–348, 2003.

[296] K. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[297] P. Kuppens, F. Tuerlinckx, P. K. Y. de Roover, and I. V. Mechelen, "Emotional inertia: A longitudinal study of individual differences in emotion dynamics," *Emotion*, vol. 10, no. 1, pp. 92–100, 2010.

[298] R. J. Larsen and E. Diener, "Affect intensity as an individual difference characteristic: A review," *Journal of Research in Personality*, vol. 21, no. 1, pp. 1–39, 1987.

[299] T. Hollenstein, *State Space Grids: Depicting Dynamics Across Development.* New York: Springer, 2015.

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

[301] M. A. Goodrich and A. C. Schultz, "Human–robot interaction: A survey," *Foundations and Trends in Human–Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[302] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human–robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[303] R. A. Calvo and S. D'Mello, *Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications*, vol. 1. IEEE Transactions on Affective Computing, 2010.

[304] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[305] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[306] J. K. Kruschke, *Doing Bayesian Data Analysis*. Academic Press, 2nd ed., 2014.

[307] B. E. Hilbig, A. Glöckner, and I. Zettler, "Personality and prosocial behavior: Linking basic traits and social value orientations," *Journal of Personality and Social Psychology*, vol. 105, no. 3, pp. 469–484, 2013.

[308] J. Katsyri, K. Forger, M. Makarainen, and T. Takala, "A review and meta-analysis of the uncanny valley: Toward a quantified model," *Frontiers in Psychology*, vol. 6, p. 390, 2015.

[309] K. Fischer, "How social is a robot? minimal social cues and their cognitive consequences," *Interaction Studies*, vol. 24, no. 1, pp. 1–29, 2023.