

# Synthetic Presence as Moral Perturbation: A Field-Theoretic Model for Evaluating Moral Behaviour in Human–Robot Interaction

Francesco Perrone

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

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



November 2025

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

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

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

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

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

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

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

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

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

*Francesco Perrone*

# Abstract

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

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

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

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

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>Declaration</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 From Research Question to Hypotheses: Framing the Investigative Architecture . . . . .	5
1.2 The Need for a New Theoretical Orientation . . . . .	8
1.3 Structure of the Thesis . . . . .	9
<b>2 Literature Review</b>	<b>12</b>
2.1 Introduction: Scope, Objectives, and Theoretical Commitments . . . . .	12
2.2 The Two Research Projects in Machine Ethics . . . . .	18
2.3 A Clarifying Perspective on Where This Work Belongs—and Where the Field Must Go . . . . .	19
2.4 Moral Psychology and Moral Philosophy: Cognitive–Affective vs. Rationalist–Intuitionist Models . . . . .	21
2.5 Levels of Abstraction and the Failure of Machine Ethics . . . . .	22
2.6 Evaluative Topology, Affective Architecture, and Synthetic Moral Perturbation . . . . .	24
2.6.1 The Evaluative Field . . . . .	24
2.6.2 Moral Behaviour as Trajectory . . . . .	25
2.6.3 Synthetic Presence as Field Operator . . . . .	25
2.6.4 Topology and the Limits of Machine Ethics . . . . .	25
2.6.5 Toward a Unified Framework . . . . .	26
2.7 Integrative Synthesis: Toward a Cognitive–Affective Model of Machine-Mediated Morality . . . . .	26
2.8 Global Synthesis: From Inferential Displacement to Synthetic Moral Topology . . . . .	27
2.8.1 From Question to Framework . . . . .	27
2.8.2 Why a Multi-Hypothesis Framework Was Needed . . . . .	27
2.8.3 What the Literature Alone Establishes . . . . .	28
<b>3 Cognitive–Affective Architecture of Moral Judgment</b>	<b>30</b>
3.1 Descriptive and Normative Domains . . . . .	31
3.1.1 Why Definitions Vary . . . . .	33
3.1.2 Minimal Operational Definition for This Thesis . . . . .	34
3.2 Judgments: Factual and Normative . . . . .	35
3.3 Internal Architecture of Moral Judgment . . . . .	36

3.3.1	Psychological and Neuroscientific Foundations of Moral Decision-Making . . . . .	39
3.4	From Moral Architecture to Perturbation by Synthetic Agents . . . . .	41
3.4.1	Philosophical Synthesis . . . . .	43
3.4.2	Concluding Perspective: Why This Matters for the Thesis	43
<b>4</b>	<b>Measurement Framework and Experimental Design</b>	<b>44</b>
4.1	The Role of Psychometric Tools in the Evaluative–Topological Architecture . . . . .	46
4.2	Why These Tools: Methodological Criteria and Alignment with the Thesis . . . . .	48
4.3	The Empathizing Quotient (EQ): Affective Resonance as a Moral Vector Field . . . . .	50
4.3.1	Historical and Theoretical Foundations . . . . .	50
4.3.2	Psychometric Validation and Cross-Cultural Work . . . . .	51
4.3.3	Empirical Applications Across Disciplines . . . . .	51
4.3.4	Critiques and Methodological Limitations . . . . .	51
4.3.5	Relevance to the Evaluative–Topological Framework . . . . .	52
4.3.6	EQ Within the Evaluative–Topological Framework . . . . .	52
4.3.7	EQ in HRI and Moral Cognition Research . . . . .	53
4.3.8	Why EQ Matters . . . . .	53
4.4	The Systemizing Quotient (SQ): Structural Evaluation and the Precision of Moral Gradients . . . . .	54
4.4.1	Historical Origins and Theoretical Motivation . . . . .	54
4.4.2	Psychometric Validation and Cross-Cultural Findings . . . . .	55
4.4.3	Empirical Uses Across Psychology, Neuroscience, and Behavioural Science . . . . .	55
4.4.4	Critiques and Limitations . . . . .	55
4.4.5	SQ Within the Evaluative–Topological Framework . . . . .	56
4.4.6	SQ, Synthetic Presence, and Behavioural Perturbation . . . . .	56
4.4.7	Why SQ Matters . . . . .	57
4.5	The Big Five Inventory (BFI): Personality Geometry and Moral Topology . . . . .	58
4.5.1	Historical Development and Theoretical Foundations . . . . .	58
4.5.2	Psychometric Strength and Cross-Contextual Validity . . . . .	58
4.5.3	Personality Predictors of Moral Behaviour . . . . .	59
4.5.4	BFI in Social Cognition, SSP, and HRI . . . . .	59
4.5.5	Personality Geometry Within the Evaluative–Topological Framework . . . . .	60
4.5.6	BFI, Perturbation, and the Interpretation of Uniform Attenuation . . . . .	60
4.5.7	Critiques, Limitations, and Relevance to the Thesis . . . . .	61
4.5.8	Cluster Semantics and BFI Geometry . . . . .	61
4.6	The Watching-Eye Paradigm: Moral Salience Amplification and Its Deformation Under Synthetic Presence . . . . .	63
4.6.1	The Watching-Eye Effect as a Topological Amplifier . . . . .	63
4.6.2	Why Child-Poster Eyes Serve as Valid Social Cues . . . . .	64
4.6.3	Why Synthetic Agents May Dilute or Distort the Watching-Eye Effect . . . . .	65

4.6.4	Watching-Eye Under Synthetic Co-Presence: Empirical Findings . . . . .	66
4.6.5	Why the Watching-Eye Paradigm Matters for the Experiment . . . . .	66
4.6.6	Integration With the Donation Paradigm . . . . .	67
4.6.7	Synthesis: The Watching-Eye Paradigm as a Window Into Moral Topology . . . . .	67
4.7	General Conclusion: Tools as the Measurement Logic of Synthetic Moral Perturbation . . . . .	68
<b>5</b>	<b>MORAL DISPLACEMENT: AN EXPERIMENTAL INVESTIGATION</b>	<b>71</b>
5.1	Conceptual Foundations of the Research Question . . . . .	71
5.2	Experimental Design and Behavioural Paradigm . . . . .	73
5.2.1	Why Minimal Presence Matters: Ontological Ambiguity as Experimental Variable . . . . .	73
5.2.2	Levels of Abstraction and the Design Logic of Minimal Robotic Presence . . . . .	75
5.2.3	Experimental design and Preliminary Results . . . . .	76
5.2.4	From Behavioural Setup to Evaluative Structure . . . . .	77
5.3	Conceptualisation of the Experiment: Moral Decision-Making under Synthetic Presence . . . . .	81
5.3.1	Formalisation of Hypothesis and Experimental Logic . . . . .	84
5.3.2	Methodological Design: Inferring Moral Perturbation through Controlled Artificial Co-presence . . . . .	84
5.3.3	Formalisation of the Experimental Logic . . . . .	85
5.3.4	Methodological Architecture: Inferring Moral Perturbation through Structured Artificial Co-presence . . . . .	86
5.3.5	Procedural Architecture of the Experimental Protocol . . . . .	87
5.3.6	Participants as Agents under Constraint . . . . .	89
5.3.7	Experimental Conditions: The Robotic Displacement Hypothesis . . . . .	89
5.3.8	Interim Evaluation of the Hypotheses and Formal Framework . . . . .	91
5.3.9	Interim Conclusion to Question 5.1 . . . . .	93
5.3.10	Preprocessing the Moral Field: Semiotic Modulation and Ontological Symmetry . . . . .	93
5.3.11	Preliminary Descriptive Patterns: Indications of Inferential Displacement . . . . .	97
5.3.12	Inferential Assessment of Attenuation: Behavioural Evidence for Perturbation . . . . .	97
5.3.13	Interim Evaluation of the Hypotheses and Formal Framework . . . . .	99
5.3.14	Interim Conclusion to Question 5.1 . . . . .	102
5.3.15	Quantification of Behavioural Modulation: Parametric and Nonparametric Effect Sizes . . . . .	103
5.4	Dispositional Baseline: Big Five Personality Traits Across Conditions	105
5.4.1	Between-Condition Differences in Big Five Personality Traits	106
5.4.2	Predictive and Moderating Roles of Big Five Traits . . . . .	106
5.4.3	Interpretive Synthesis . . . . .	107
5.4.4	Latent Trait Structures and Individual Modulation of Moral Perturbation . . . . .	108

5.4.5	Psychometric Interpretation and Semantic Labelling of Latent Personality Clusters . . . . .	111
5.4.6	Interim Synthesis: Moral Attenuation, Topological Deformation, and Trait-Contingent Modulation . . . . .	114
5.4.7	The Dilution of the Watching Eye Effect under Robotic Co-Presence . . . . .	117
5.4.8	Cluster-Specific Regression Analysis of Robotic Perturbation	117
5.4.9	Bayesian Estimation and Epistemic Gradient Framing . . .	120
5.4.10	Final Synthesis: Moral Topology, Synthetic Presence, and the Limits of Machine Ethics . . . . .	126
5.4.11	Synthesis: Moral Topology, Synthetic Presence, and the Limits of Machine Ethics . . . . .	128
<b>6</b>	<b>ETHICAL COGNITION AND NORMATIVE FOUNDATIONS</b>	<b>131</b>
6.1	From Moral Cognition to Ethical Theory . . . . .	131
	Bridging Note: From Moral Cognition to Ethical Theory . . . . .	131
6.2	Introduction: Why Ethics Needs Psychology (and Why Computing Science Needs Both) . . . . .	132
6.3	Ethical Theory as Second-Order Analysis . . . . .	134
6.3.1	Ethical Reflection and the Second-Order Stance . . . . .	134
6.3.2	Levels of Abstraction and the Proper Location of Ethical Explanation . . . . .	135
6.3.3	Evaluative Topology as a Bridge Between Orders . . . . .	137
6.4	The Normative Landscape: Structuring Ethical Theories Through LoA and Topology . . . . .	140
6.4.1	The Three Dimensions of Normative Analysis . . . . .	141
6.4.2	Why This Framework Matters for the Experimental Chapter	141
6.5	Deontological Structures: The Architecture of Practical Reason .	142
6.5.1	The Source of Normativity: Rational Agency and the Form of Law . . . . .	143
6.5.2	Mode of Evaluation: Maxims, Duties, and the Structure of Permissibility . . . . .	144
6.5.3	Action-Guidance: How Normative Constraints Influence Behaviour . . . . .	144
6.5.4	Deontological Normativity as Topological Invariance . . . .	145
6.5.5	Why Deontology Matters for the Experimental Logic . . .	145
6.6	Consequentialist Structures: Value Gradients and the Topology of Outcomes . . . . .	148
6.6.1	The Source of Normativity: Welfare, Impartiality, and the Structure of Reasons . . . . .	148
6.6.2	Mode of Evaluation: Consequences, Expected Value, and Scalar Normativity . . . . .	149
6.6.3	Action-Guidance Mechanism: From Value Gradients to Behavioural Pressure . . . . .	150
6.6.4	Consequentialist Topology: Moral Action as Gradient Following . . . . .	151
6.6.5	Why Consequentialism Matters for the Experimental Logic	151
6.7	Virtue-Theoretic Structures: Dispositions, Character Topology, and Moral Sensitivity . . . . .	152

6.7.1	The Source of Normativity: Character, Practical Wisdom, and Moral Perception . . . . .	153
6.7.2	Mode of Evaluation: Dispositions as Topological Structure	153
6.7.3	Action-Guidance Mechanism: Habituation, Stability, and Situated Sensitivity . . . . .	154
6.7.4	Virtue-Theoretic Topology: Stability, Curvature, and Susceptibility to Perturbation . . . . .	154
6.7.5	Why Virtue Ethics Matters for the Experimental Logic . .	155
6.7.6	Virtue-Ethical Interpretation of Latent Ecologies . . . .	157
6.8	Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology . . . . .	159
6.8.1	The Source of Normativity: Sentiment as the Basis of Moral Appraisal . . . . .	160
6.8.2	Mode of Evaluation: Affective Resonance as Moral Metric	160
6.8.3	Action Guidance: Affective Vector Fields and Behavioural Dynamics . . . . .	161
6.8.4	Contrast with Machine Ethics: The Blind Spot of Affective Architecture . . . . .	161
6.8.5	Experimental Realisation: Synthetic Dampening of Empathic Resonance . . . . .	162
6.9	Contractualism, Particularism, and Hybrid Normative Models .	163
6.9.1	Contractualism: Moral Claims as Justification-Equilibria .	164
6.9.2	Moral Particularism: Contextual Salience and the Fragmented Topology of Reasons . . . . .	165
6.9.3	Hybrid and Pluralist Models: Multidimensional Topologies	167
6.9.4	Integrative Ethical Interpretation of the Experimental Findings . . . . .	168
<b>7</b>	<b>General Discussion and Theoretical Integration</b>	<b>171</b>
7.1	Introduction: Why the Experiment Requires a Structural Interpretation . . . . .	171
7.1.1	From Behaviour to Structure: Why a Higher-Level Interpretation is Required . . . . .	173
7.1.2	Why This Chapter Cannot Be Pure “Discussion” in the Conventional Sense . . . . .	176
7.1.3	A Structural Reading of the Core Experimental Result . .	176
7.1.4	Why the Synthetic Presence Effect Matters Beyond the Experiment . . . . .	177
7.2	Cluster-by-Cluster Integrative Interpretation . . . . .	178
7.3	Global Normative–Topological Synthesis . . . . .	181
7.4	From the Failure of Machine Ethics to a Reconstruction of Computational Morality . . . . .	183
7.4.1	Reconstructing Computational Morality: An Empirically Grounded Paradigm . . . . .	184
7.4.2	Computational Morality as a Scientific Research Programme	185
7.5	Thesis-Wide Synthesis and Closing Reflections . . . . .	186
<b>Bibliography</b>		<b>189</b>

## List of Tables

5.1	Demographic balance tests across experimental conditions. The table reports the original and FDR-corrected p-values for comparisons of gender, age, and educational background. No significant differences were detected after correction, supporting the assumption of demographic equivalence between groups. . . . .	91
5.2	Experimental conditions are behaviorally and procedurally identical, differing only in robotic presence. . . . .	94
5.3	Measured variables and psychometric constructs used in inferential modelling of moral behaviour. . . . .	94
5.4	Summary of central tendencies for key behavioural and psychometric variables. The Robot condition shows numerically lower donation amounts and empathizing scores, suggesting potential attenuation effects of passive robotic presence. . . . .	97
5.5	Inferential comparisons of donation behaviour across conditions. The chi-squared test identifies a significant difference in aggregate donation totals, while the Mann–Whitney U test and bootstrapped mean difference indicate substantial distributional overlap and a diffuse, heterogeneous perturbative effect. . . . .	99

## List of Figures

5.1	Top-down view of experimental vs. control configurations. Both settings retain identical spatial and visual layouts, isolating the variable of robotic presence as the only ontological difference. . . . .	77
5.2	Age distribution across experimental conditions. Histogram representation confirms no major between-group demographic divergence. . . . .	96
5.3	Distribution of donation behaviour by condition. Violin plot representation visualizes the heavier tail in the robot group, supporting the hypothesized interpretive perturbation. . . . .	96
5.4	Mean donation amounts by experimental condition, with 95% bootstrapped confidence intervals. Participants in the Control condition donated more on average than those in the Robot condition, aligning with the hypothesis that robotic presence attenuates the inferential mapping from moral salience to prosocial action. The overlapping confidence intervals highlight substantial individual-level variability and the probabilistic nature of the perturbation. . . . .	100
5.5	Kernel density estimates of donation distributions across conditions. The Control group exhibits higher central mass and a heavier rightward extension relative to the Robot group, consistent with a directional attenuation of high-value prosocial acts in the presence of the synthetic co-presence $\mathcal{R}$ . . . . .	104
5.6	Mean donation amounts with standard error bars by condition. The Control group donates more on average (£1.89) than the Robot group (£1.17), corroborating the hypothesis that robotic presence modulates—rather than eliminates—the evaluative pathway from moral salience to action. . . . .	105
5.7	Kernel density estimates for each Big Five trait across experimental conditions, demonstrating substantial distributional overlap. . . . .	106
5.8	Scatter plots with fitted regression lines for each Big Five trait against donation amount. Each panel displays individual participant scores alongside a smoothed linear trend. No clear predictive relationships emerge, reinforcing the conclusion that the Big Five traits do not meaningfully predict prosocial donation within this experimental context. . . . .	107
5.9	Participants clustered in PCA-reduced psychometric space, coloured by cluster identity and shaped by experimental condition. The clustering reveals three latent personality regimes, each representing a distinct cognitive-affective configuration encoded in $\beta_C$ . . . . .	109

5.10	Elbow plot of within-cluster sum of squares (left axis) and silhouette coefficients (right axis) across candidate values of $k$ . The elbow at $k = 3$ and interpretable silhouette profile support the selection of three clusters as a parsimonious and psychologically meaningful solution. . . . .	110
5.11	Mean donation amount by experimental condition within each personality cluster, derived from $k$ -means analysis on psychometric trait profiles. Error bars represent standard deviation. Cluster 1 shows a marked attenuation of donation under robotic presence, whereas Clusters 0 and 2 exhibit minimal or modest differences. This pattern suggests that the perturbative effect of $\gamma_R$ is contingent upon latent cognitive-affective regimes encoded in $\beta_C$ . . . . .	111
5.12	Comparative radar profiles of the three latent personality ecologies. <b>Emotionally Reactive / Low-Structure Profile</b> (left): elevated Neuroticism with reduced Conscientiousness and Systemizing. <b>Prosocial-Empathic / Warm-Sociable Profile</b> (centre): high Openness, Extraversion, Agreeableness, and Empathizing. <b>Analytical-Structured / High-Systemizing Profile</b> (right): high Systemizing and Conscientiousness with lower Empathizing. . . . .	112
5.13	Regression coefficients for the Robot condition within each personality cluster (95% confidence intervals). The Prosocial-Empathic profile shows a pronounced attenuation effect, while the Emotionally Reactive and Analytical-Structured profiles exhibit negligible or non-significant coefficients. This pattern demonstrates that robotic presence exerts a differentiated moral influence, contingent on latent cognitive-affective ecologies. . . . .	119
5.14	Posterior distribution of the estimated donation difference between the Control and Robot conditions. The density skews toward negative values, indicating directional probabilistic evidence that robotic co-presence attenuates prosocial behaviour. The vertical dashed line denotes the point of no effect. Bayesian inference renders the effect size and its uncertainty as a continuous epistemic field rather than a binary verdict. . . . .	121

## Acknowledgements

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

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

## **Declaration**

I declare that, with the exception of background sections that review existing literature and established theoretical frameworks (chapters 2, ??, and part of chapter 7) all original research presented in this thesis—including the experimental design, data collection, statistical analysis, clustering procedures, and the development of the evaluative-topological model of moral perturbation—was carried out independently by the author, unless otherwise explicitly stated. All sources have been appropriately acknowledged, and no part of this thesis has been submitted for any other degree or qualification.

# 1. Introduction

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.

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 develops within 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  $f : \mathcal{X} \rightarrow \mathcal{A}$  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— $f : \mathcal{X} \rightarrow \mathcal{A}$ —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 deform 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.

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:

*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

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

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 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  $\beta_C$  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 co-presence induces a measurable, structured alteration in the mapping from moral salience to action that cannot be reduced to trait-level variation or noise.

Chapter 6 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 methodolog-

ical 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 7 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.

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 rather a first test looking for 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:

*Whether 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<sup>1</sup>; 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].

---

<sup>1</sup>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 Chapter 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 eval-

uative 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.

## 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 Judgment

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 [299, 200] 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 housekeep-

ing; 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 [93, 94]. Its importance here lies in the fact that:

*synthetic perturbation affects normative judgment, even though the factual perception of the situation might remain 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 [95, 80]. This is more than a semantic contrast. It marks a functional division within the cognitive architecture: judgments about what 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, 96].

Psychological models therefore treat factual information as input to evaluative appraisal [97, 98, 99, 87]. Normative judgment requires an additional mapping: the transformation of descriptive cues into action-guiding evaluations [100, 101, 102]. 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 [103, 104, 105, 106, 107]. 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, 101, 102, 87]. 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 pres-

ence 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.<sup>1</sup> 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?” [108, 19, 101].

**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 [108]. 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 [109, 110, 111]. The amygdala detects threat, intentional aggression, and aversive outcomes [112, 47], while the anterior insula responds to disgust, norm violations, and aversive interoception [113, 114, 115]. Electrophysiological studies indicate that these affective signals often precede conscious deliberation [116, 117], functioning as rapid gating mechanisms for downstream moral appraisal.

**Social-Cognitive and Interpretive Systems.** Moral judgments frequently hinge on beliefs, intentions, and reasons [118, 119]. The temporo-parietal junction (TPJ), medial prefrontal cortex (mPFC), and posterior superior temporal sulcus (pSTS) form a network specialised for mental-state attribution [121, 122, 123, 124]. TPJ activation, for example, is reliably observed when distinguishing intentional from accidental harms or attributing blame or forgiveness [125, 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 [126, 127]. Classic moral dilemmas recruit ACC activity when intuitive emotional responses and reflective considerations collide [10, 128]. This conflict-monitoring function indicates that moral cognition involves arbitration among multiple evaluative forces [129, 130].

---

<sup>1</sup>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 [131, 132]. Disruption of dlPFC activity via TMS alters willingness to endorse instrumental harm [133, 134], demonstrating that this region contributes to structuring action policies that integrate affective, deontic, and goal-directed considerations [96, ?].

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 [135, 136, 137]. Recent accounts describe this network as computing action-guiding commitments rather than abstract evaluations [138, 139].

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, 140, 141]. This integration is fundamentally practical. The vmPFC and OFC compute the affective value of potential actions [142, 143]; the TPJ and mPFC generate intention-sensitive interpretations of agents’ behaviour [122, 125]; the ACC detects conflict between competing behavioural tendencies [126, 127]; and the dlPFC regulates whether intuitive impulses should be suppressed, enacted, or balanced against normative constraints [131, 133]. 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, 115].

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 [101, 144]. 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*.<sup>2</sup>

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 [145, 146]. At our operative LoA, the morally relevant property of a robot is its ability to *warp attentional and affective gradients* that structure human appraisal [147, 148]. A robot can function as a normative

---

<sup>2</sup>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 [149, 150]. 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* [151, 152].

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

$$f : \Sigma \rightarrow \Delta, \quad \mathcal{P}_{\mathcal{R}} : \Sigma \rightarrow \Sigma', \quad f_{\mathcal{R}} = f \circ \mathcal{P}_{\mathcal{R}}$$

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. Measurement Framework and Experimental Design

Empirical work aimed at understanding moral cognition must specify, with some philosophical care, the instruments through which psychological and behavioural structures become accessible to observation. Moral appraisal itself is never directly given; it is inferred from patterned responses—*affective*, *dispositional*, *perceptual*, and *social*—that reflect how evaluative information is encoded in the agent’s cognitive architecture [9, 180, 10, 11, 12, 181]. The tools employed in this thesis therefore function not as neutral measurement devices but as theoretically motivated probes: each instrument targets a specific dimension of the evaluative topology developed in earlier chapters, rendering latent dispositional structure empirically tractable without collapsing its complexity into reductive summary scores.

The methodological commitments of this thesis require a principled account of the instruments through which evaluative behaviour becomes empirically accessible. Work in moral psychology and cognitive science has repeatedly shown that moral appraisal is not directly observable but manifests through structured patterns of affective response, controlled cognition, and social cue integration [9, 180, 10, 11, 12, 181]. 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 [182], and the Big Five Inventory [183, 184] 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 behaviourally respond to contextually salient perturbations.

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

The Empathizing Quotient [7], the Systemizing Quotient [182], and the Big Five Inventory [183, 184] 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.

This chapter therefore serves a conceptual rather than merely procedural purpose. The psychometric instruments and experimental paradigms introduced here are situated explicitly within the evaluative-topological model developed in earlier chapters, in which moral cognition is understood not as a sequence of discrete judgments but as the dynamic evolution of a manifold of interacting evaluative gradients. Contemporary theories of moral psychology emphasise that such gradients integrate affective, social, and contextual inputs in a manner shaped by stable dispositional architecture [9, 180, 10, 11, 12, 181]. Within this framework, the role of each tool is to reveal invariant dispositional structures—the stable dimensions along which individuals differ in how incoming evaluative information is encoded and transformed. These structures correspond to the latent parameters governing how a subject’s evaluative gradients bend, flatten, or intensify as the informational environment is perturbed.

In the experiment motivating this thesis, such perturbation is elicited not through explicit moral dilemmas but through a more subtle and ecologically grounded manipulation: the silent perceptual presence of a humanoid robot. Prior work in human–robot interaction shows that even passively positioned robots can shift perceived social affordances, alter attentional allocation, and modulate expectations concerning norm-relevant behaviour [35, 185, 186]. Their ambiguous ontological status disrupts default social priors and thereby reconfigures the salience landscape within which moral reasons become behaviourally operative. In this respect, robotic presence functions as a controlled perturbation to the evaluative topology itself, enabling the empirical study of how dispositional invariants interact with contextual cues to produce measurable differences in moral behaviour.

The aim of the chapter is thus twofold.

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

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

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

The tools included here—the Empathizing Quotient (EQ), the Systemizing Quotient (SQ), the Big Five Inventory (BFI), and the Watching–Eye paradigm—were selected because they satisfy three stringent criteria grounded in established empirical research. First, each instrument has a well-defined construct lineage supported by extensive psychometric validation. The EQ [7] and SQ [182] constitute the canonical operationalisations of empathizing and systemizing tendencies, with consistent factor structures, cross-cultural robustness, and demonstrable discriminant validity within both clinical and non-clinical populations. The BFI, in its original form [183] and in its widely used short version [184], provides a compact yet psychometrically rigorous assessment of the five broad personality domains that anchor contemporary trait theory.

Second, the Watching–Eye paradigm has developed into a mature experimental framework for probing reputation-sensitive prosocial behaviour. Numerous studies have demonstrated that minimal cues of observation modulate cooperative and charitable actions [1, 2, 4, 6, 5], and the paradigm’s effects have been replicated across diverse contexts, task structures, and elicitation modalities. This makes it uniquely suited for isolating perturbations to social-evaluative processing—a core requirement for the present analysis.

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

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

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

#### 4.1 The Role of Psychometric Tools in the Evaluative–Topological Architecture

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

trajectories arise from the joint interaction of environmental cues, dispositional structure, and perturbational influences [9, 180, 10, 11, 12, 181]. Accordingly, the formal decomposition

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

captures the three principal determinants of evaluative dynamics:

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

The psychometric tools employed in this study belong to the  $\beta_C$  term. They render dispositional structure empirically tractable by quantifying constructs shown to be central in the integration of affective, social, and contextual information. The Empathizing Quotient [7] indexes the **affective bandwidth** through which agents register morally salient others; the Systemizing Quotient [182] captures the **analytical curvature** underlying structural interpretation of social situations; and the Big Five Inventory [183, 184] measures the **personality geometry** shaping attentional allocation, normative sensitivity, and regulatory control [187]. These constructs have well-established roles in models of moral appraisal and behavioural prediction [9, 12].

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

What is theoretically significant, however, is what followed. Despite this structured dispositional diversity, the humanoid robot exerted a *uniform directional effect* on prosocial behaviour across all clusters. No Big Five trait, EQ subscale, SQ dimension, or latent profile moderated the displacement. Prior research in human–robot interaction has shown that even passive robotic agents can shift perceived social affordances, modulate attention, and alter expectations surrounding norm-relevant behaviour [35, 185, 186]. The present findings extend this line of work by demonstrating that robotic presence does not operate through trait-dependent amplification or suppression of behavioural tendencies. Instead, it perturbs the evaluative field itself—its salience structure, affective gradients, and normative attractors—such that all dispositional trajectories are bent in the same behavioural direction.

In this sense, the psychometric tools were indispensable. They allowed the analysis to dissociate the *shape of the dispositional manifold* from the *geometry of the perturbation*. Without psychometric grounding, the attenuation of donation

behaviour might have been misinterpreted as a trait-level effect rather than a field-level displacement. The instruments thereby provided the empirical precision needed to show that the robot acted not upon who the participants were, but upon the evaluative topology within which their moral choices unfolded.

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

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

which measures how the moral transformation function is reshaped by  $\gamma_R$  given a fixed dispositional configuration. This formulation reflects the broader consensus in moral psychology that moral behaviour emerges from the interaction between environmental cues, dispositional structure, and perturbational influences on evaluative processing [9, 180, 10, 11, 12, 181]. The empirical results presented in Chapter ?? showed that although  $\beta_C$  exhibits a structured internal topology—revealed through clustering analyses of EQ, SQ, and BFI scores [7, 182, 183, 184, 187]—the perturbation introduced by the humanoid robot did *not* depend on those dispositional differences. All clusters displayed the same directional attenuation of prosocial behaviour, indicating that  $\gamma_R$  operates primarily at the *field level*, reshaping the evaluative landscape within which dispositional trajectories unfold rather than interacting with trait-specific gradients.

This pattern aligns with established findings in human–robot interaction, where even passive robotic agents have been shown to modulate perceived social affordances, attentional allocation, and norm-relevant expectations irrespective of observer traits [35, 185, 186]. In the present experiment, robotic presence functioned as a global perturbation of the evaluative field rather than as a selective amplifier or suppressor of individual dispositions.

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

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

## 4.2 Why These Tools: Methodological Criteria and Alignment with the Thesis

Given the dual-layer structure revealed by the experiment—stable dispositional variation on the one hand, and a field-level displacement effect induced by robotic presence on the other—the selection of psychometric and experimental tools cannot be arbitrary. The instruments employed here were chosen because they satisfy

three methodological criteria essential for interpreting the attenuation of prosocial behaviour observed in the study.

**(1) Cross-paradigmatic relevance.** The EQ, SQ, BFI, and Watching–Eye paradigm each rest on extensive empirical traditions across several domains of inquiry. In moral and social psychology, these tools have been used to study prosociality, empathic concern, harm aversion, and the integration of affective and cognitive processes in moral judgment [9, 180, 10, 11, 12, 181]. In personality psychology, the Big Five Inventory provides a compact but psychometrically robust measure of trait architecture with well-established predictive value for behavioural outcomes [183, 184, 187]. The Empathizing and Systemizing Quotients offer validated assessments of affective resonance and analytic style [7, 182].

In parallel, the Watching–Eye paradigm constitutes one of the most reliable experimental manipulations of prosocial salience, with repeated demonstrations that subtle cues of observation can modulate cooperative and charitable behaviour [1, 2, 4, 6, 5]. Crucially, these literatures intersect with contemporary Human–Robot Interaction research, where robotic agents are known to shift social affordances, attentional allocation, and normative expectations [35, 185, 186]. Their use therefore positions the present study within a broad empirical landscape while maintaining continuity with the theoretical commitments of the evaluative–topological framework.

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

- **EQ:** the affective attractors that anchor early moral and social appraisal [7];
- **SQ:** the structural curvature associated with analytic or rule-based processing [182];
- **BFI:** the multidimensional geometry of personality traits that modulate salience, attentional uptake, and behavioural regulation [183, 184, 187];
- **Watching–Eye paradigm:** an experimentally validated perturbation that shifts moral salience without instruction or coercion [1, 2, 4, 6, 5].

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

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

- they support clustering of participants within dispositional space,
- they enable regression modelling of trait influences on donation behaviour,

- and they admit interpretation through established normative and metaethical frameworks, including sentimentalism, virtue-theoretic accounts, and pluralist models of moral reasoning.

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

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

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

The Empathizing Quotient (EQ) occupies a central place in the measurement of affective sensitivity within contemporary psychology. Developed by Baron-Cohen and colleagues as part of the broader Empathizing–Systemizing (ES) framework [7, 188, 189], the EQ was originally designed to quantify individual differences in emotional resonance, perspective-taking, and the capacity to infer and respond appropriately to the mental states of others. Its construction reflects two decades of theoretical and empirical work stemming from autism research, sex differences in social cognition, and the development of trait-based accounts of empathic functioning.

#### 4.3.1 Historical and Theoretical Foundations

The EQ emerged against the background of two influential lines of inquiry. The first concerned the cognitive and affective profiles observed in autism spectrum conditions, where empathic difficulties appeared as a core diagnostic dimension. Baron-Cohen's early work on "mindblindness" and the ES theory [188] proposed that empathizing and systemizing represent partially dissociable cognitive styles, with autism characterized by diminished empathizing abilities alongside preserved or enhanced systemizing capacities. The second line of inquiry derived from trait psychology and social cognition, where stable inter-individual differences in emotional attunement, empathic accuracy, and prosocial inclinations were increasingly understood as predictive of moral and social behaviour.

The EQ was designed to operationalise the empathizing construct in a psychometrically rigorous manner. It includes affective items (e.g., sensitivity to distress), cognitive-empathic items (e.g., perspective-taking), and items assessing spontaneous concern for others. Initial investigations [7] demonstrated large group differences between autistic and neurotypical adults, robust sex differences, and high internal reliability. Subsequent factor-analytic studies [190] further clarified the latent structure of the scale, identifying separable components associated with emotional reactivity, cognitive perspective-taking, and social attunement.

### 4.3.2 Psychometric Validation and Cross-Cultural Work

Psychometric validation of the EQ has been extensive. Beyond the initial work in clinical and neurotypical samples, replication studies have demonstrated strong internal consistency, acceptable test-retest reliability, and predictable convergence with related constructs such as empathic concern, emotional intelligence, and social sensitivity [7]. Cross-cultural validations, including Japanese and Western samples, have shown that the EQ maintains its factor structure and predictive value across cultural contexts [191].

These findings situate the EQ within the broader movement toward trait-based quantification of social-cognitive skills. Within personality psychology, empathizing correlates with the Agreeableness and Openness dimensions of the Big Five [187], while remaining psychometrically distinguishable from both. Within social neuroscience, EQ scores have been found to correlate with vmPFC–amygdala coupling and with the strength of activation in neural substrates associated with social pain, affect sharing, and mentalising.

### 4.3.3 Empirical Applications Across Disciplines

The EQ has become a standard instrument in multiple research paradigms. In moral psychology, empathy-related traits are strong predictors of altruistic helping, harm aversion, guilt sensitivity, and responses to moral dilemmas [9, 180, 10, 11, 12, 181]. High EQ scores are consistently associated with stronger prosocial choices in economic games, including the ultimatum, dictator, and trust games. Behavioural economics work shows that individuals with higher empathic sensitivity display increased generosity even when anonymity is preserved, suggesting that empathic traits modulate internalised moral norms beyond external social cues.

In social neuroscience, EQ scores track activation patterns in regions associated with affective resonance, including the anterior insula, temporoparietal junction, and amygdala–vmPFC networks. Oxytocin administration studies further demonstrate selective improvement in empathic accuracy [192], reinforcing the biological plausibility of affective resonance as a trait-like dimension.

The EQ has also gained significance in Human–Robot Interaction (HRI), where empathic predispositions shape attributions of intentionality, perceived moral standing, and expectations regarding robots’ behaviour [185, 186, 35]. Individuals with higher EQ scores tend to ascribe richer mental states to robots, respond more strongly to cues of intentionality, and exhibit greater sensitivity to violations of social or moral norms in robotic agents. In group-based interactions, empathic individuals demonstrate greater behavioural alignment with robots, particularly when robots display subtle affective or communicative signals [33].

### 4.3.4 Critiques and Methodological Limitations

Despite its widespread use, the EQ has faced several critiques. Some researchers argue that its factor structure is not fully stable across populations, with certain studies reporting two or three factors rather than the originally proposed triadic structure. Concerns have also been raised regarding response biases, social desir-

ability, and the possibility that self-report measures may not accurately capture behavioural or neural indices of empathy. Cross-cultural studies have noted differences in average EQ scores, prompting questions about cultural calibration and the extent to which certain items rely on culturally specific norms of emotional expression.

Within experimental psychology, some scholars have argued that empathic responding is situationally variable and cannot be fully reduced to trait-level constructs. Studies demonstrating dissociations between empathic concern and moral behaviour in high-stakes dilemmas [11, 10] raise further questions about the predictive specificity of the EQ. Nevertheless, the scale remains one of the most widely used and empirically grounded measures of individual differences in affective resonance.

#### 4.3.5 Relevance to the Evaluative–Topological Framework

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

In the context of the experiment, the EQ plays a crucial role in modelling the dispositional term  $\beta_C$ . It enables the analysis to determine whether differences in affective sensitivity condition the behavioural response to a perturbation in the evaluative field—namely, the silent presence of a humanoid robot. The finding that EQ did *not* moderate the displacement effect provides strong evidence that the robot acted at the field level rather than through trait-specific amplification or suppression. This result is consistent with HRI studies showing that robotic presence alters normative expectations independently of empathic predispositions [185, 186, 35].

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

#### 4.3.6 EQ Within the Evaluative–Topological Framework

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

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

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

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

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

#### 4.3.7 EQ in HRI and Moral Cognition Research

Studies have shown that high empathizers:

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

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

#### 4.3.8 Why EQ Matters

Although the Empathizing Quotient possesses deep theoretical relevance for modelling the affective attractors that shape trajectories within the evaluative manifold, its primary role in the experiment was methodological. The EQ provides a validated measure of emotional resonance, perspective-taking, and sensitivity to others' mental and affective states [7, 190, 191, 188, 189]. Because empathic capacity is a well-established predictor of altruistic behaviour, harm aversion, and cooperative decision-making [9, 180, 10, 11, 12, 8], failing to quantify it would have introduced a serious confound into the interpretation of the attenuation effect.

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

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

Although the Empathizing Quotient possesses deep theoretical relevance for modelling the affective attractors that shape trajectories within the evaluative manifold, its primary role in the experiment was methodological. The EQ provides a validated measure of emotional resonance, perspective-taking, and sensitivity to others' mental and affective states [7, 190, 191, 188, 189]. Because empathic capacity is a well-established predictor of altruistic behaviour, harm aversion, and

cooperative decision-making [9, 180, 10, 11, 12, 8], failing to quantify it would have introduced a serious confound into the interpretation of the attenuation effect.

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

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

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

Where the Empathizing Quotient (EQ) captures affective resonance, the Systemizing Quotient (SQ) [193, 194, 191] quantifies an individual's propensity for identifying structural regularities, constructing causal models, and applying rule-based inference. Developed in parallel with the ES theory of cognition [188, 189], the SQ was designed to measure the degree to which an agent seeks predictive coherence in complex environments. It therefore provides a natural operationalisation of what, within the evaluative-topological framework of this thesis, we describe as the *analytical curvature* of the evaluative field: the tendency to encode moral situations via structural invariants rather than affective attractors.

##### 4.4.1 Historical Origins and Theoretical Motivation

The origins of the SQ lie in the broader attempt to model cognitive styles that differentiate autistic from neurotypical populations. Baron-Cohen's early work proposed that systemizing reflects a cognitive drive for rule extraction and causal precision, complementing empathizing but operating through distinct computational mechanisms [188]. The Systemizing Quotient was introduced as the psychometric realisation of this construct, with the initial validation study [193] demonstrating its sensitivity to within-group variation as well as to group-level differences between autistic and neurotypical adults.

The theoretical motivation for systemizing has since broadened. While originally embedded in autism research, systemizing has come to be linked with general tendencies toward mechanistic reasoning, causal Bayes nets, and algorithmic-level representations of environmental structure. Psychometric studies have shown that high-SQ individuals exhibit a preference for deterministic rules, hierarchical schemas, and low-noise value comparisons [194]. In the ES theory, empathizing and systemizing jointly define a two-dimensional space in which variation in social

cognition, emotional regulation, and reasoning strategies can be mapped.

#### 4.4.2 Psychometric Validation and Cross-Cultural Findings

Validation studies demonstrate that the SQ has high internal consistency, good test-retest reliability, and predictable correlations with cognitive-style measures, including analytic problem-solving, sensitivity to pattern structure, and preference for system-based explanations. Importantly, cross-cultural validation work [194, 191] has shown that the SQ retains its factor structure and predictive validity across different cultural contexts, supporting the claim that systemizing taps into a cognitive style with cross-cultural generality.

Neurocognitive work complements these findings. High systemizing tendencies correlate with activation in lateral prefrontal and parietal cortices associated with analytic reasoning, causal inference, and top-down attentional control. Conversely, higher systemizing scores are associated with reduced activation in affective salience networks during social evaluation tasks [8], reinforcing the link between SQ and reduced susceptibility to affect-laden cues.

#### 4.4.3 Empirical Uses Across Psychology, Neuroscience, and Behavioural Science

Systemizing has been deployed across a wide range of empirical domains. In moral psychology, systemizing tendencies predict a greater reliance on deliberative processes in dual-process moral judgment models [46]. High-SQ individuals exhibit greater stability in moral evaluations across contexts, a preference for rule-consistency, and an increased likelihood of endorsing principle-based judgments in high-conflict dilemmas [11]. The reduced affective reactivity associated with high systemizing is consistent with findings showing that utilitarian judgments arise under conditions of weaker affective engagement and stronger top-down control [10, 8].

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

In Human–Robot Interaction, systemizing tendencies strongly modulate expectations regarding synthetic agents. High-SQ participants are more likely to attribute competence, reliability, and causal predictability to robots, and less likely to respond to anthropomorphic cues [35, 185, 186]. This makes SQ particularly relevant for the experimental context of this thesis: systemizing provides a dispositional anchor for understanding how agents interpret the structural affordances introduced by a humanoid robot, especially in settings where affective cues are minimal and norm-relevant structure must be inferred.

#### 4.4.4 Critiques and Limitations

Although widely used, the SQ is not without critique. Some studies report that its factor structure is more heterogeneous than originally proposed, with potential

subfactors corresponding to mechanical reasoning, abstract pattern detection, and rule-based inference. There are also concerns about cultural calibration, particularly regarding item content related to technical interests, which may vary across populations.

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

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

#### 4.4.5 SQ Within the Evaluative–Topological Framework

Within the evaluative–topological model, SQ modulates the *second derivative* of the evaluative potential function: it influences the *rigidity*, *smoothness*, and *predictability* of evaluative gradients. High-SQ agents encode situations as stable causal schemas rather than affective landscapes. Consequently, their evaluative fields resist deformation under purely affective perturbations and favour top–down interpretive stability. This interpretation aligns with theoretical work emphasising the role of structural representations in moral reasoning [46, 11].

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

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

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

#### 4.4.6 SQ, Synthetic Presence, and Behavioural Perturbation

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

Consistent with research on deliberative dominance [195, 96], high-SQ individuals exhibit reduced coupling between affective cues and evaluative processing, as well as diminished susceptibility to emotionally charged primes. In the context of the present experiment, these individuals correspond most closely to the *Analytical–Structured* dispositional cluster revealed through the combination of EQ, SQ, and BFI scores.

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

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

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

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

### Why SQ Matters for the Experiment

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

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

#### 4.4.7 Why SQ Matters

The inclusion of the Systemizing Quotient provides a unified means of capturing several features of the evaluative landscape that would otherwise remain theoretically disjoint. Within the ES framework [188, 193, 189], systemizing reflects a cognitive style centred on structural analysis, causal precision, and rule-based inference. In topological terms, it indexes the *deliberative curvature* of the moral field: the degree to which evaluative trajectories are shaped by stable causal schemas rather than affective attractors [194, 191].

High-SQ individuals are reliably characterised by reduced affective reactivity and a stronger reliance on deliberative pathways, as shown in both moral psychology [46, 11] and affective neuroscience [8]. This makes the SQ theoretically informative for distinguishing sentimentalist mechanisms of moral judgment—which emphasise affective resonance and intuitive appraisal [9, 12]—from structuralist mechanisms that privilege rule-coherent, model-based evaluation [195, 96].

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

In this light, the SQ serves two essential empirical functions within the study. First, it rules out systemizing tendencies as a proximate explanation for the

behavioural shift, ensuring that the attenuation cannot be attributed to cognitive-analytical style. Second, it supplies the structural dimension of  $\beta_C$  necessary to model how dispositional architecture relates to global field deformation. The SQ therefore anchors the analytical side of the dispositional manifold without implying differential behavioural susceptibility to synthetic presence.

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

Among the major instruments of differential psychology, the Big Five Inventory (BFI) occupies a uniquely robust position. Originating from decades of lexical, psychometric, and theoretical research [196, 197], the Big Five model offers a parsimonious description of personality variation along five orthogonal axes: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These traits have repeatedly demonstrated high stability, cross-cultural generality, and considerable predictive power across behavioural domains [187, 198, 184]. As a measurement instrument, the BFI therefore provides an empirically grounded coordinate system for mapping the dispositional substrate ( $\beta_C$ ) posited within the evaluative-topological framework developed in the present thesis and subsequently used to interpret the uniform displacement effect observed in the experiment.

### 4.5.1 Historical Development and Theoretical Foundations

The Big Five model traces its roots to the lexical tradition in personality research, in which factor-analytic investigations of trait-descriptive adjectives yielded a consistent five-factor structure across languages and populations. Building on this foundation, John and Srivastava [196] formalised the BFI as a psychometrically concise yet highly reliable instrument for assessing the cardinal trait dimensions. Parallel work by McCrae and Costa [197] provided a broader theoretical synthesis, linking the Big Five to a hierarchical model of personality structure and embedding them within a developmental and biological framework.

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

### 4.5.2 Psychometric Strength and Cross-Contextual Validity

The BFI is among the most validated instruments in modern psychology. Its dimensional structure has been replicated across diverse populations, and its items exhibit strong internal consistency and temporal stability [196, 198]. Short-form adaptations such as the BFI-10 [184] preserve much of this reliability while enabling efficient deployment in time-constrained experimental contexts such as the present one.

Importantly, Big Five traits predict consequential real-world outcomes across domains including job performance [187], relationship quality, subjective well-being, and health behaviours. These predictive successes justify the use of BFI metrics as indices of theoretically meaningful dispositions that shape evaluative and behavioural tendencies. The cross-cultural robustness of the Big Five further supports their role as part of a generalisable dispositional architecture that can be integrated into computational and topological models. In our experiment, this stability ensured that personality variation could be meaningfully mapped onto the dispositional manifold used to test whether the robotic perturbation exerted trait-dependent or trait-independent effects.

#### 4.5.3 Personality Predictors of Moral Behaviour

A substantial body of work has examined the relation between Big Five traits and prosocial or moral behaviour. Agreeableness is the most consistent predictor of helping, cooperation, and empathic concern [199, 200]. Individuals high in Agreeableness are more responsive to others' needs, more sensitive to interpersonal harm, and more disposed toward altruistic action even in anonymous or low-reciprocity contexts.

Conscientiousness has been linked to moral rule adherence, planning, and long-horizon evaluative stability. Individuals high in Conscientiousness exhibit greater behavioural regularity and stronger alignment with internalised norms, qualities that translate into reduced noise in moral decision-making. Neuroticism predicts greater affective volatility, heightened sensitivity to social threat, and increased susceptibility to contextual perturbation [201]. Extraversion amplifies responsiveness to social presence and increases the weighting of socially salient cues. Openness broadens receptivity to contextual novelty, increases tolerance of ambiguity, and enhances exploratory behaviour in moral and social domains.

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

#### 4.5.4 BFI in Social Cognition, SSP, and HRI

Beyond moral psychology, the BFI plays a central role in research on social cognition and nonverbal behaviour. Personality traits influence expressive dynamics, gaze patterns, vocal modulation, and gesture production—behaviours that constitute the core observational cues in Social Signal Processing (SSP). Vinciarelli et al.'s foundational survey [63] highlights how personality traits can be inferred from multimodal behavioural signatures (speech prosody, movement patterns, attention allocation), and how these traits modulate social engagement, turn-taking, and responsiveness to social cues.

These insights reinforce the relevance of the BFI in contexts involving robotic presence. Traits such as Extraversion and Agreeableness shape social approach

tendencies, sensitivity to perceived agency, and responsiveness to social affordances—all properties critical in HRI scenarios. Banks [202] demonstrates that personality interacts with perceptions of robot trustworthiness, sociality, and intentionality, thereby linking the Big Five to the cognitive mechanisms underlying moral or cooperative evaluation of artificial agents.

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

#### 4.5.5 Personality Geometry Within the Evaluative–Topological Framework

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

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

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

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

#### 4.5.6 BFI, Perturbation, and the Interpretation of Uniform Attenuation

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

This finding is consistent with evidence from HRI and social cognition that suggests robotic presence can shift social affordances at a level that bypasses trait-level predispositions, acting instead through global modifications of perceived agency, social monitoring, or norm salience [185, 186, 35]. The BFI was crucial in establishing this. By providing a structured mapping of personality geometry, the instrument made it possible to dissociate trait-level variation from field-level displacement. Without this differentiation, the attenuation could have been misattributed to personality differences rather than to the synthetic perturbation introduced by the robot.

#### 4.5.7 Critiques, Limitations, and Relevance to the Thesis

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

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

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

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

#### 4.5.8 Cluster Semantics and BFI Geometry

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

1. **Prosocial–Empathic:** high Agreeableness, high Openness, high EQ; steep affective attractors; strong dispositional orientation toward altruistic trajectories (not reflected in significant behavioural differences in the experiment).
- .

2. **Emotionally Reactive:** high Neuroticism, mixed EQ; unstable gradients; heightened susceptibility to contextual variation.
3. **Analytical–Structured:** high Conscientiousness and SQ; rigid gradients; evaluative trajectories shaped by rule-based stability, though not resistant to the field-level perturbation observed in the experiment.

Including the BFI provides several methodological and theoretical advantages that are essential for interpreting the experimental findings. First, it offers a validated and fine-grained quantification of dispositional topology, grounding the latent structure of  $\beta_C$  in a trait framework with well-established psychometric credentials [196, 197, 183, 198, 184]. Second, it enables the clustering of heterogeneous evaluative architectures, allowing the analysis to detect meaningful personality configurations rather than assuming psychological homogeneity across participants [187]. Third, it anchors normative interpretation within empirically instantiated personality space, thereby linking moral behaviour to stable and theoretically interpretable dispositional dimensions.

Including the BFI allows:

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

Most critically, the inclusion of the BFI makes it possible to demonstrate that the synthetic perturbation introduced by the humanoid robot operates at the *field level* rather than through trait-specific pathways. Agreement, Extraversion, and related traits are well-known predictors of prosociality and interpersonal sensitivity [199, 200, 201], yet none moderated the attenuation in donation behaviour observed in the experiment. By ruling out personality-dependent explanations, the BFI prevents misinterpretation of the results as mere reflections of Agreeableness or Extraversion and instead supports the conclusion—also consistent with findings in HRI [35, 185, 186]—that robotic presence reshapes the evaluative field itself.

Without the BFI, the experiment would lack both the dimensional granularity required to distinguish dispositional variation from field-level displacement and the evidential basis for excluding personality traits as the proximate cause of the attenuation effect. The instrument therefore secures two levels of inference: the micro-level behavioural interpretation (donation behaviour was not simply a function of trait prosociality) and the macro-level topological interpretation (the robot perturbed the evaluative field rather than trait-dependent gradients). In this way, the BFI plays a decisive role in demonstrating that the displacement

effect arises from a global modification of evaluative geometry rather than from personality-based modulation.

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

One of the most robust findings in behavioural ethics, social psychology, and field-based prosociality research is the *watching-eye effect*. Minimal cues of observation—such as stylised eye images, schematic-gaze primes, or human-like monitoring cues—reliably increase prosocial behaviour, generosity, charitable giving, and norm-compliant action [1, 2, 4, 6]. Early interpretations framed this effect in terms of reputational vigilance, proposing that even minimal perceptual cues can activate implicit monitoring systems and thereby increase norm adherence [1]. Subsequent work expanded this view, demonstrating that watching-eye stimuli operate through distributed attentional, affective, and interpretive mechanisms [2, 4, 6].

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

### 4.6.1 The Watching-Eye Effect as a Topological Amplifier

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

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

Formally, the cue introduces:

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

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

**Reputational Mechanisms.** Classical studies show that minimal observation cues trigger reputational vigilance, increasing the perceived costs of norm violation and shifting behaviour toward compliance and fairness [1, 2]. In topological terms, such cues steepen deontic and prosocial attractors within the evaluative field, making cooperative trajectories more gravitationally dominant.

**Attentional Mechanisms.** Watching-eye stimuli also function as attentional amplifiers. Experimental evidence demonstrates that observation cues draw perceptual resources toward socially and normatively relevant features, modulating early-stage appraisal processes [6, 4, 5]. This attentional reweighting alters the initial intuitive gradients that guide moral evaluation, consistent with both dual-process accounts and the topological formalism developed earlier.

**Affective and Self-Conscious Emotions.** Observation cues can induce mild affective arousal, activating self-conscious emotions such as guilt, embarrassment, or pride. Pfattheicher and Keller [203] show that such cues increase prosocial tendencies by elevating somatic markers associated with cooperative action. In topological terms, this creates a local rise in affective curvature, making prosocial trajectories more energetically accessible.

**Context Sensitivity.** The watching-eye effect is not universal. Its magnitude varies with local normative expectations, cultural context, cue ambiguity, and ecological validity [204, 4]. This context dependence is crucial for understanding how synthetic presence may interact with, dilute, or override the effect, particularly when robotic agents introduce novel or ambiguous social affordances.

#### 4.6.2 Why Child-Poster Eyes Serve as Valid Social Cues

Child-poster eyes have become a widely adopted tool in prosociality and donation-based paradigms because they represent a minimal, reliable, and theoretically interpretable form of social cueing. Extensive research across behavioural ethics, evolutionary psychology, and field experimentation demonstrates that stylised eyes—particularly child-like or infantile forms—robustly increase cooperation, charitable giving, and norm compliance in both laboratory and ecologically naturalistic settings [1, 2, 3, 72, 4, 5]. Survey and mechanistic studies on gaze perception further show that eye-like stimuli heighten implicit monitoring, attentional engagement, and the salience of norm-relevant behaviour [6].

**Perceptual Sociality Without Agentic Commitment.** One key advantage of child-eye posters is that they elicit social attentiveness without invoking full-fledged agency, intention, or belief ascription. Findings from developmental social cognition show that infant-like eyes are powerful communicative signals capable of triggering gaze-following, social vigilance, and context-oriented attention [205]. These cues therefore allow the experiment to modulate the environmental input term  $\alpha_E$  without introducing confounds related to mental-state attribution or anthropomorphic inference.

**Care-Related Affective Resonance.** Infant and child imagery reliably evoke *empathic concern and affiliative motivation*, an effect well established in social neuroscience and affective psychology [206, 8]. Within the evaluative-topological model, these stimuli steepen prosocial attractors by amplifying affective resonance, thereby increasing the affective curvature associated with cooperative trajectories.

**Methodological Control.** Child-poster eyes are low-dimensional, easily standardisable stimuli. Unlike dynamic or agentive observers, they minimise interpretive ambiguity while producing replicable increases in prosocial behaviour [1, 2, 3, 72]. Their methodological reliability makes them especially suitable for controlled perturbation of moral salience—precisely the role played by the environmental input term  $\alpha_E$  in the present experimental design.

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

A central theoretical insight of this thesis is that humanoid robots—although perceptually social—possess what can be described as an *unstable social ontology*. Prior work in human–robot interaction shows that robots occupy an ambiguous position within the space of social agents: they can trigger attentional and interactive responses, yet they do not clearly instantiate the mental, moral, or evaluative capacities normally associated with observers [35, 185, 186]. This ambiguity directly disrupts the mechanisms that underpin the watching-eye effect.

In empirical and theoretical accounts of the watching-eye effect, behavioural modulation arises from an automatic inference that an entity with perceptual access also possesses the evaluative and sanctioning capacities required to make one’s behaviour socially consequential. Classical studies show that minimal eye cues activate reputational vigilance, triggering heuristic assumptions about observers who can form impressions, update reputational standings, and administer rewards or punishments [1, 2, 4, 5]. These effects depend on deep-seated attentional and affective systems specialised for detecting evaluative observers, particularly those capable of moral appraisal [6].

Synthetic agents disrupt this chain of inferences. Although their perceptual and bodily cues can signal social presence, their ambiguous social ontology undermines the attribution of intentionality, evaluative capacity, and normative authority. Findings in human–robot interaction show that humanoid robots frequently elicit perceptual but not moral-evaluative attributions [35, 185, 186]. As a result, the observer-detection heuristic receives conflicting inputs: the perceptual system flags an agentive presence, while higher-order cognitive systems register the absence of genuine mental states or sanctioning power. This dissonance weakens reputational motivation, destabilises expectations of being judged, and attenuates the motivational architecture that produces the watching-eye effect in human–human contexts.

**Perceptual Sociality Without Clear Ontology.** Humanoid robots signal presence but lack a stable set of agentic, intentional, or moral attributes. Because reputational vigilance depends on attributing evaluative capacities to an observer, this ontological instability weakens the mapping from perceived observation to expectations of norm compliance [185]. In topological terms, the robot introduces perceptual salience without the deontic structure that normally steepens prosocial attractors.

**Disrupted Affective and Attentional Gradients.** Although robots reliably elicit gaze, they do not consistently activate the affective and evaluative systems

associated with being judged by another mind [6, 35]. The result is a diminished perturbational signal: the effective increase in environmental input,  $\delta\alpha_{\text{eye}}$ , is attenuated, and the prosocial attractors in the evaluative field remain comparatively flat.

**Interpretive Uncertainty.** Participants may attribute perceptual sensitivity to robots (“it sees me”) without extending moral-evaluative capacities (“it judges me”). This asymmetry creates a fractured evaluative field: social presence is registered, but the normative or reputational meaning of that presence is ambiguous or absent [186]. Because the watching-eye effect depends on a coherent mapping between observation and evaluation, this interpretive uncertainty dilutes the salience amplification typically produced by eye cues.

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

#### 4.6.4 Watching-Eye Under Synthetic Co-Presence: Empirical Findings

The experiment yielded a consistent and theoretically revealing result:

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

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

Formally, the robot introduces:

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

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

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

#### 4.6.5 Why the Watching-Eye Paradigm Matters for the Experiment

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

- **A standardised probe of moral salience:** it provides a reproducible baseline against which perturbations can be detected.

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

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

#### 4.6.6 Integration With the Donation Paradigm

Donation tasks provide a measure of real behavioural commitment rather than hypothetical endorsement, capturing the downstream consequences of evaluative processing in action [10, 11, 12, 19, 18]. Integrating watching-eye cues with a cost-bearing prosocial choice allows the experiment to:

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

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

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

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

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

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

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

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

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

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

where  $\alpha_E$  denotes environmental moral cues,  $\beta_C$  the dispositional manifold, and  $\gamma_R$  the perturbational operator introduced by synthetic presence. Each tool corresponds directly to one or more components of this formalism:

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

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

#### Dispositional Mapping and the Rejection of Trait-Based Explanations

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

The empirical findings clearly support the latter. No Big Five trait, no empathizing or systemizing tendency, and no dispositional cluster moderated the effect: the robot's presence produced a *uniform directional attenuation* in prosocial behaviour across all profiles. Without trait-level measurement, this displacement could have been misinterpreted as a consequence of stable personality differences rather than as evidence of structured perturbation.

## Watching-Eye as a Diagnostic Amplifier

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

## Ethical and Theoretical Implications

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

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

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

## Transition to the Experimental Methods

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

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

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

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

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

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

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

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

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

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

## 5. MORAL DISPLACEMENT: AN EXPERIMENTAL INVESTIGATION

### 5.1 Conceptual Foundations of the Research Question

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

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

Within the domains of social signal processing and artificial intelligence, the transformation of subtle environmental cues into behavioural outputs is treated as a mapping from informational stimuli to structured responses [63]. By embedding a humanoid robot—ontologically ambiguous, semantically potent, yet behaviourally inert—into a morality-salient environment, this experiment asks whether such synthetic presences perturb not the content of deliberation, but the signal-to-inference architecture through which salience becomes action.

#### Question 5.1: Inferential Displacement

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

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

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

cues elicit prosocial action.

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

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

Framing the investigation as a *question* (Question 5.1 p. 71) rather than a hypothesis is deliberate. It preserves the conceptual openness required at this stage of the analysis, foregrounding inquiry over prediction. Within interdisciplinary research—spanning moral psychology, social signal processing, and human–robot interaction—prematurely imposing a directional hypothesis risks presupposing the very moral effects that the experiment is designed to probe. By articulating a guiding research question rather than an asserted claim, we allow the empirical structure of the data to shape the inferential trajectory rather than constraining it in advance. This is consistent with both the methodological caution urged in philosophy of science and the epistemic humility appropriate when dealing with morally charged, psychologically subtle, and technologically novel forms of social influence.

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

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

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

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

## 5.2 Experimental Design and Behavioural Paradigm

To investigate this Question 5.1 (see p. 71), we implemented a controlled behavioural experiment [207, 208, 209] derived from the classical *Watching Eye* paradigm [65, 2, 210, 211, 203, 212, 213], in which prosocial behaviour is modulated by the perceived presence of observation. Seventy-three participants were invited individually into a room under the pretext of completing a battery of personality questionnaires in exchange for monetary compensation. Embedded in the experimental space was a morally salient cue: a charity brochure prominently featuring the photograph of a child requiring medical assistance. Decades of empirical work show that such stimuli reliably activate prosocial dispositions through mechanisms of implicit monitoring and empathetic engagement [1, 6].

The robotic manipulation was then introduced as the sole experimental variable. In the control condition, participants completed the task alone. In the experimental condition, a humanoid robot—NAO [214]—remained silently present in *autonomous life mode*, exhibiting only the minimal embodied cues characteristic of that configuration: simulated breathing, subtle postural adjustments, and reactive head orientation triggered exclusively by eye contact.

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

### 5.2.1 Why Minimal Presence Matters: Ontological Ambiguity as Experimental Variable

Much of the literature on moral decision-making in human–robot interaction (HRI) and human–machine interaction (HMI) locates moral modulation in the interactive capacities of artificial agents. Studies routinely foreground expressive behaviour, ostensive cues, adaptive responsiveness, displays of accountability, or anthropomorphic signalling as the levers through which machines influence human judgment and behaviour [34, 215, 216, 217, 218]. These approaches implicitly assume that moral impact requires action: verbal behaviour, communicative intent, social reciprocity, or strategically framed moral cues.

*The present experimental design intentionally refuses this assumption.*

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

This methodological divergence is conceptually foundational. It allows us to target an aspect of moral cognition that is often overlooked: its *pre-reflective permeability* (for a similar use of the term refer to [219, 220, 221, 222]) to agent-like cues even when those cues lack *intentional content* [223, 224, 225]. The question is not whether robots can engage in moral exchange, but whether their presence, by virtue of their bodily form and minimal behavioural affordances, reshapes the inferential scaffolding that mediates between perceiving a moral cue and acting upon it.

This problem is particularly salient in domains such as Social Signal Processing and computational social cognition, where synthetic agents routinely evoke social and moral reactions that exceed the informational complexity of their behaviour [63, 51]. By removing dialogue, task-relevance, and overt interaction while maintaining the perceptual markers of potential agency (eyes, posture, orientation, micro-motion), the experiment isolates **presence itself** as the epistemic variable to be tested.

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

This intuition resonates with the hyperactive intentional stance described by Guthrie [226], Waytz et al. [227], and Dennett [228], according to which humans routinely over-asccribe agency in uncertain environments. By positioning the robot in the liminal space between objecthood and agenthood, the experiment isolates not action, but anticipation—the silent priors that precede full agentive recognition.

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

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

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

### 5.2.2 Levels of Abstraction and the Design Logic of Minimal Robotic Presence

The decision to deploy a humanoid robot in silent autonomous life mode—exhibiting only simulated breathing, subtle postural adjustments, and eye-contact-contingent head orientation—is not a matter of convenience or technological limitation. It is a philosophical and methodological choice grounded in Floridi’s theory of *Levels of Abstraction* (LoA) [25, 229, 230]. To appreciate this decision, the core function of LoAs must be understood with conceptual precision.

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

Applied to the present experiment, LoAs allow us to describe moral influence without relying on metaphysical accounts of robot agency. At the LoA operative for a participant alone in a room, moral relevance does not depend on the robot’s internal states but on its semantic affordances: its posture, its eyes, the symmetry of its body, the direction of its face, its quiet imitation of biological rhythms [231, 232, 233, 234, 235, 236, 237, 238].

These features are perceptually encoded as possible indicators of being watched [231, 239, 240, 232, 234, ?, 241, 242], evaluated, or accompanied—precisely the conditions under which the Watching Eye effect operates. Thus, the robot’s moral relevance emerges not from consciousness, autonomy, or interactive capacity, but from its informational presentation within the participant’s operative LoA.

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

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

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

associated with potential observation but insufficient to produce interactive interpretation. In this capacity, NAO aligns with Floridi and Sanders' notion of an *artefactual moral agent* [151, 230]: a non-sentient entity whose moral relevance arises not from autonomy but from the role it plays within an informationally structured environment.

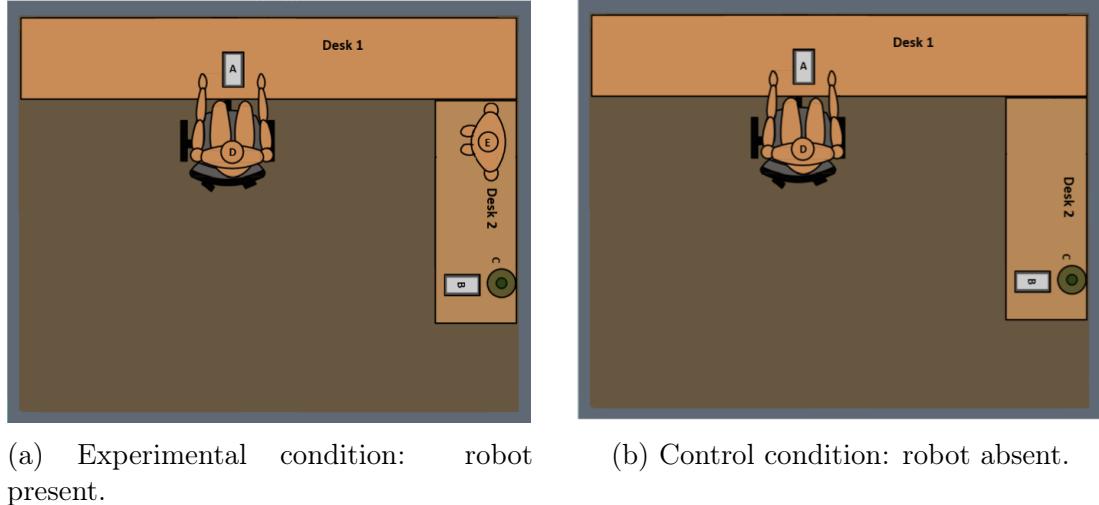
**FP:** This is more a conclusion.

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

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

### 5.2.3 Experimental design and Preliminary Results

To investigate Question 5.1, we implemented a controlled behavioural experiment [207, 208, 209] derived from the classical *Watching Eye* paradigm [65, 2, 210, 211, 203, 212, 213], in which prosocial behaviour is modulated by implicit cues of observation. Each participant was invited individually into a room under the pretext of completing a personality-study session in exchange for monetary compensation. Unbeknownst to them, the experimental environment contained a morally salient stimulus: a charity brochure displaying the photograph of a child requiring medical care. Decades of empirical work demonstrate that such stimuli reliably trigger prosocial dispositions by activating implicit monitoring and empathetic engagement [1, 6].



(a) Experimental condition: robot present. (b) Control condition: robot absent.

Figure 5.1: Top-down view of experimental vs. control configurations. Both settings retain identical spatial and visual layouts, isolating the variable of robotic presence as the only ontological difference.

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

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

Initial results revealed a robust directional pattern: participants in the Robot condition donated substantially less than those in the Control condition. Furthermore, no meaningful between-group differences were found in personality profiles (Empathizing Quotient [7], Systemizing Quotient [182], Big Five Inventory [183]), ruling out trait-based confounds and strengthening the inference that robotic presence itself modulated the evaluative pathway underlying prosocial action.

#### 5.2.4 From Behavioural Setup to Evaluative Structure

In moral philosophy, action is frequently treated as the terminus of deliberation [13, 15, 79]. Yet the present study concerns not the deliberative endpoint but the evaluative transformation that precedes it: the internal process by which morally salient cues are converted into behavioural output [83, 80]. The experimental design above provides the behavioural substrate; what remains is to articulate the evaluative architecture through which robotic presence might exert

its influence.

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

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

Empirically, the experiment transposes the Watching Eye paradigm into a minimal social environment co-inhabited by a humanoid robot. Prior variants of the paradigm have relied on stylised pictorial stimuli or supernatural primes [2, 243]. Our design replaces these with an embodied artificial presence whose ontological ambiguity is semantically potent while remaining behaviourally minimal.

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

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

Where:

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

Read aloud, this expresses the hypothesis that:

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

#### Hypothesis 1: *Evaluative Deformation Hypothesis*

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

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

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

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

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

where:

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

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

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

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

To understand the stakes of this perturbation, we must clarify what is meant by *moral salience*. Across philosophical and psychological literatures, moral salience refers to the capacity of a situation, object, or agent to present itself as morally significant—i.e., to become an object of evaluative attention prior to explicit deliberation [80, 83, 31, 45, 113]. It functions as a phenomenological filter: before the agent reasons, before the agent chooses, certain features of the environment appear as normatively charged. Within this framework, synthetic entities may perturb moral salience not by issuing commands or engaging in dialogue, but by reconfiguring what is foregrounded, what is suppressed, and what is affectively or normatively “seen” in the first place.

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

**Hypothesis 2: Synthetic Normativity of Moral Displacement**

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

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

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

This framing rejects simplified accounts that treat moral behaviour as transparent readouts of internal disposition. Instead, it positions moral action as the contingent result of cognitive-affective systems operating under topological deformation [180, 16, 181]. Robotic presence, by virtue of its ontological ambiguity, functions as a refractive moral affordance: a structural condition that may attenuate or redirect the transformation of moral salience into action.

**FP:** old content begins

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

**FP:** old content ends

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

normative cognition in post-human social configurations.

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

This framing **rejects** any simplistic account of moral behaviour as noise-free reflection of trait. Instead, we position moral action as the contingent result of *cognitive-affective systems* operating under *topological deformation* [180, 16, 181]. In this view, robotic presence is not merely a contextual feature, but a morally refractive affordance that alters the mapping between cue and action.

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

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

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

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

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

Each participant entered the experimental room alone and completed a series of psychometric measures—the Empathizing Quotient [188], Systemizing Quotient [193], and the Big Five Inventory [184]. Completion of the questionnaires served a dual methodological purpose. First, it provided data for assessing whether trait dispositions modulated sensitivity to robotic presence, thereby enabling the analysis of trait–context interactions central to this chapter. Second,

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

Upon completion, participants received £10 in £1 coins as compensation. Before exiting, they encountered a latent moral choice: an opaque green charity box placed beside a prominently displayed image of a child requiring medical care. The image served as a *Watching Eye* stimulus—an established elicitor of prosocial behaviour through mechanisms of implicit monitoring, empathy, and reputation sensitivity [1, 6]. Donation behaviour, performed privately and unobserved, constituted the operational measure of moral action.

The central experimental manipulation divided participants into two conditions:

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

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

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

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

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

To understand why this effect is theoretically significant, we must clarify the status of *moral decision-making* within this experimental architecture. Contrary to utilitarian models that construe donation as a form of preference optimisation (see chapter 6), our framing treats the decision to donate as an instantiation

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

The Watching Eye stimulus plays a critical role in this machinery. Anthropological and psychological research shows that images of eyes or children reliably elicit third-party moral concern via affective engagement and implicit audience effects [2, 210, 212]. Our design extends this paradigm by placing, alongside the Watching Eye cue, a humanoid robot whose ontological status is neither human nor ethically inert. NAO thus becomes an *ontological anomalous agent*: a presence that possesses the perceptual affordances of agenthood without the behavioural or normative commitments of actual agency.

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

### Hypothesis 3: *Synthetic Perturbation of Moral Inference*

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

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

where:

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

In control conditions, the transition  $\Sigma \rightarrow \mathcal{D}$  proceeds without interference: the affective weight of moral cues is preserved and expressed through prosocial giving [1, 243]. In robotic conditions, by contrast,  $\mathcal{R}$  deforms this mapping. It may displace empathic identification, dilute the salience of the Watching Eye cue, reshape the normative topology of the environment, or function as a cognitive decoy [35]. Each interpretation bears distinct implications for the design of ethical robots and for understanding how humans recalibrate moral behaviour in the presence of synthetic others.

### 5.3.1 Formalisation of Hypothesis and Experimental Logic

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

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

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

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

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

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

$$\mathcal{R} \in \Sigma \Rightarrow \mathbb{E}[f(\Sigma \cup \mathcal{R})] = D_{\text{attenuated}}$$

where:

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

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

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

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

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

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

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

where:

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

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

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

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

### 5.3.3 Formalisation of the Experimental Logic

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

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

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

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

To make this structure explicit, let us denote:

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

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

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

with the expected attenuation constraint:

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

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

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

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

At the heart of this design lies the formal proposition:

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

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

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

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

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

### 5.3.5 Procedural Architecture of the Experimental Protocol

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

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

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

Participants were recruited through two parallel channels: internal advertisements within the School of Computing Science at the University of Glasgow and via the Psychology subject pool. Eligibility criteria included (i) a minimum age of 17 years, (ii) British nationality, verified upon arrival, and (iii) where applicable, exclusion of Computing Science students from the Psychology pool to prevent sampling overlap (see section 5.3.6 for full demographic detail).

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

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

##### Stage 1: Arrival and Initial Framing

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

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

#### Stage 2: Environmental Exposure and Moral-Salience Priming

Participants entered an isolated experimental room configured according to their assigned condition. In both conditions, a large poster depicting a child beneficiary from a medical charity (*Operation Smile*) was affixed to the wall directly facing the participant. This image served as the Watching Eye stimulus ( $\alpha_E$ ), providing a latent reputational cue that has been shown to activate prosocial tendencies under minimal prompting.

In the *Robot Condition*, a SoftBank Robotics **NAO** robot was placed passively in the room, configured in *autonomous life mode*. In this mode, NAO exhibits subtle embodied cues: simulated breathing, minimal postural adjustments, and reactive head orientation triggered *only* upon direct eye contact. These micro-movements instantiate the perturbative variable  $\gamma_R$ , furnishing a perceptually salient but behaviourally minimal form of co-presence.

#### Stage 3: Completion of Psychometric Instruments

Participants completed three psychometric questionnaires:

- **Empathizing Quotient (EQ)** [188], indexing affective resonance.
- **Systemizing Quotient (SQ)** [193], indexing rule-based cognitive preference.
- **Big Five Inventory-10 (BFI-10)** [184], capturing broad personality traits.

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

#### Stage 4: Monetary Compensation and Moral Decision Opportunity

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

#### Stage 5: Exit and Data Collection

Participants exited the room individually. The experimenter then recorded the amount donated, retrieved completed questionnaires, and anonymised all identifiers for analysis.

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

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

### 5.3.6 Participants as Agents under Constraint

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

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

Demographically, participants were drawn from two sources:

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

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

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

### 5.3.7 Experimental Conditions: The Robotic Displacement Hypothesis

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

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

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

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

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

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

The configuration employed here is deliberately poised between these extremes. NAO is activated enough to be *socially legible*, yet withdrawn enough to remain *epistemically opaque*. In Floridi's terminology, the robot is an artefact whose *LoA-encoded features* (face, posture, micro-movement) render it morally salient despite the absence of moral agency [230, 151]. At this operative LoA, its status is neither neutral nor agentive but semiotically charged: a presence that presents itself as potentially intentional, without fulfilling the criteria for genuine agency.

Within this framework, NAO occupies the role of what Coeckelbergh [145] and Złotowski et al. [35] describe as a *moral appearance operator*: an entity whose embodied features trigger interpersonal expectations even in the absence of genuine communicative exchange. In our design, the robot becomes a **norm deflector**: it does not issue commands, but it may reconfigure the evaluative bandwidth through which the Watching-Eye stimulus is interpreted.

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

#### *Demographic Equivalence and Inferential Symmetry*

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

The results were unequivocal:

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

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

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

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

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

Table 5.1: Demographic balance tests across experimental conditions. The table reports the original and FDR-corrected p-values for comparisons of gender, age, and educational background. No significant differences were detected after correction, supporting the assumption of demographic equivalence between groups.

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

### 5.3.8 Interim Evaluation of the Hypotheses and Formal Framework

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

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

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

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

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

#### *Status of the Mathematical Formalism*

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

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

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

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

thus supporting the retention of the Evaluative Deformation Hypothesis.

#### **(c) The tripartite decomposition**

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

This expression separates environmental cues ( $\alpha_E$ ), dispositional factors ( $\beta_C$ ), and robotic presence ( $\gamma_R$ ). **Contribution so far:** it justifies the inclusion of

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

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

### 5.3.9 Interim Conclusion to Question 5.1

#### Partial Conclusion to Question 5.1

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

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

### 5.3.10 Preprocessing the Moral Field: Semiotic Modulation and Ontological Symmetry

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

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

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

Table 5.2: Experimental conditions are behaviorally and procedurally identical, differing only in robotic presence.

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

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

Table 5.3: Measured variables and psychometric constructs used in inferential modelling of moral behaviour.

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

#### Question 4: Ontological Integrity of the Dataset

##### Question 5.2: *Data structuring*

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

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

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

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

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

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

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

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

Figures 5.2 and 5.3 visually corroborate this reading: age distributions show no demographic divergence, while donation distributions reveal the predicted attenuation under robotic co-presence. The unified visual palette of the plots maintains stylistic continuity with the thesis's typographic aesthetic, reinforcing the epistemic unity of the chapter's representational forms.

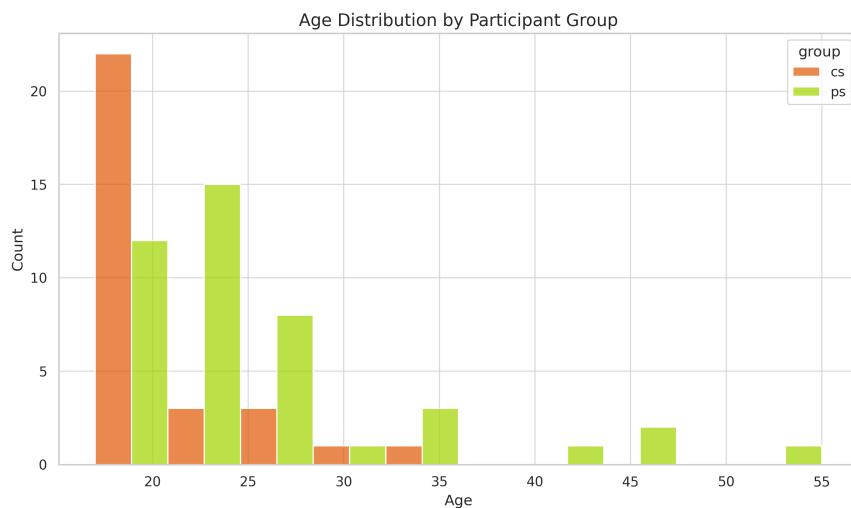


Figure 5.2: Age distribution across experimental conditions. Histogram representation confirms no major between-group demographic divergence.

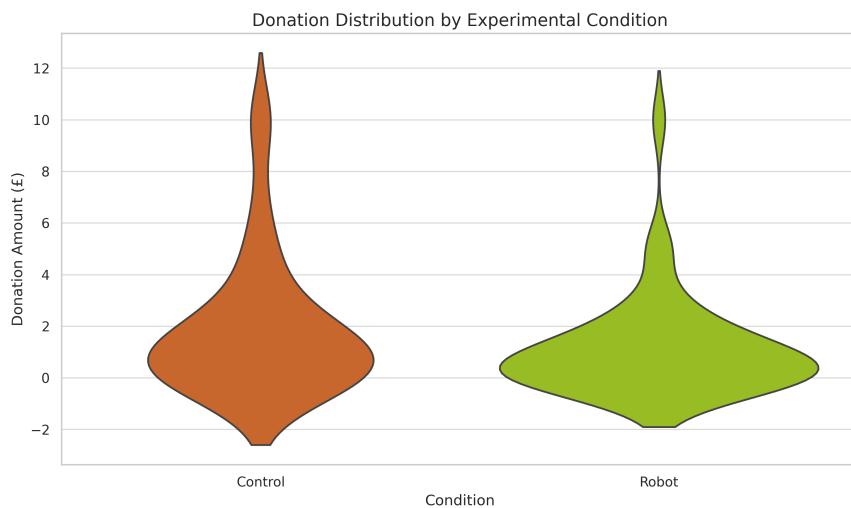


Figure 5.3: Distribution of donation behaviour by condition. Violin plot representation visualizes the heavier tail in the robot group, supporting the hypothesized interpretive perturbation.

### 5.3.11 Preliminary Descriptive Patterns: Indications of Inferential Displacement

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

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

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

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

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

Table 5.4: Summary of central tendencies for key behavioural and psychometric variables. The Robot condition shows numerically lower donation amounts and empathizing scores, suggesting potential attenuation effects of passive robotic presence.

### 5.3.12 Inferential Assessment of Attenuation: Behavioural Evidence for Perturbation

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

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

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

#### Conclusion: Aggregate Attenuation of Prosocial Output

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

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

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

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

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

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

Table 5.5: Inferential comparisons of donation behaviour across conditions. The chi-squared test identifies a significant difference in aggregate donation totals, while the Mann–Whitney U test and bootstrapped mean difference indicate substantial distributional overlap and a diffuse, heterogeneous perturbative effect.

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

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

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

### 5.3.13 Interim Evaluation of the Hypotheses and Formal Framework

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

The **Evaluative Deformation Hypothesis** (Hypothesis 1 p. 78) asserts that the expected outcome of moral behaviour, as computed by the evaluative trans-

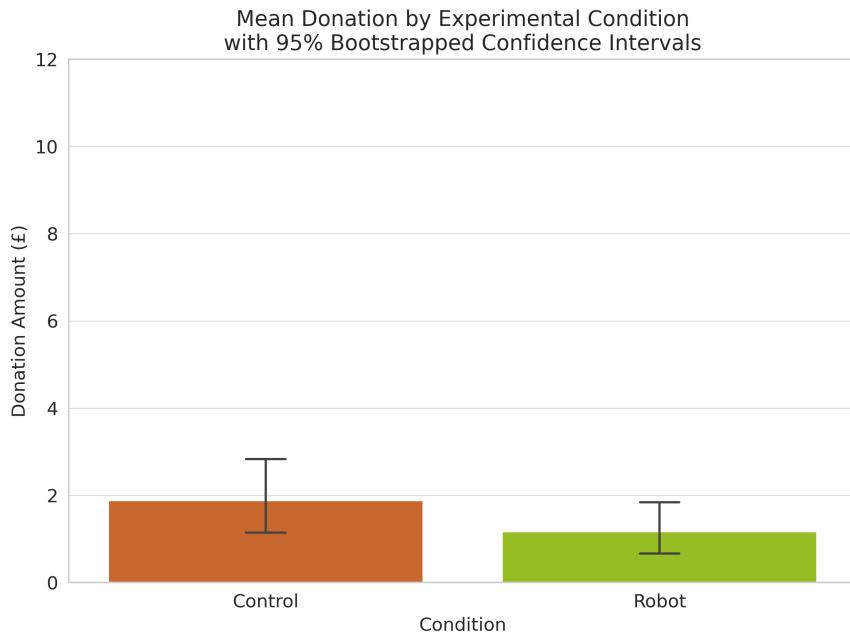


Figure 5.4: Mean donation amounts by experimental condition, with 95% bootstrapped confidence intervals. Participants in the Control condition donated more on average than those in the Robot condition, aligning with the hypothesis that robotic presence attenuates the inferential mapping from moral salience to prosocial action. The overlapping confidence intervals highlight substantial individual-level variability and the probabilistic nature of the perturbation.

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

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

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

The **Synthetic Normativity of Moral Displacement** hypothesis (Hypothesis 2, p. 80) provides the ontological and conceptual groundwork for interpreting this deformation. It claims that synthetic presences, though devoid of sentience, may acquire normative affordances by virtue of their perceived ontology. The present evidence neither confirms nor disconfirms this hypothesis in a narrow statistical sense; rather, it shows that a non-interactive yet semantically rich artefact, **positioned at the appropriate Level of Abstraction**, can exert measurable influence on prosocial behaviour without issuing commands, arguments, or reasons. This is exactly the pattern one would expect if normative affordances were grounded in informational presentation at a given LoA, rather than in intrinsic moral status. The hypothesis is thus **retained** as the principal conceptual lens through which the behavioural results are interpreted.

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

p. 83) specifies the mechanism connecting the previous two: the robot does not merely co-occur with lowered donations; it perturbs the inferential transition from moral salience to prosocial action by refracting the affective–empathic cues that would otherwise support donation behaviour. The combined pattern of (i) significant aggregate attenuation, (ii) overlapping individual-level distributions, and (iii) non-trivial yet fragile effect sizes is coherent with this mechanistic reading: the evaluative mapping is not destroyed, but **its topology is altered**. This hypothesis is therefore **retained** as a working account of how the deformation is instantiated at the level of moral inference. In sum, all three hypotheses remain live and mutually reinforcing:

- Hypothesis 1, (p. 78) identifies the *empirical signature* of deformation at the level of expected behaviour.
- Hypothesis 2, (p. 80) explains the *ontological possibility* of such deformation within Floridi’s informationalist framework and its Levels of Abstraction.
- Hypothesis 3, (p. 83) articulates the *inferential pathway* through which robotic presence reshapes the transition from moral salience to action.

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

#### *Status of the Mathematical Formalism*

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

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

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

#### **(c) The tripartite decomposition**

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

This expression disaggregates environmental cues ( $\alpha_E$ ), dispositional factors ( $\beta_C$ ), and robotic presence ( $\gamma_R$ ). **Contribution so far:** it justifies the joint consideration of (i) the Watching Eye stimulus, (ii) psychometric traits and demographics,

and (iii) robotic co-presence as distinct yet interacting contributors to moral behaviour. The current behavioural results speak primarily to the  $\gamma_R$  component, while leaving open the possibility that its effect is modulated by structured configurations of  $\beta_C$ —a possibility that will be examined through regression and interaction models in the analyses that follow.

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

### 5.3.14 Interim Conclusion to Question 5.1

#### Partial Conclusion to Question 5.1

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

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

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

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

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

### 5.3.15 Quantification of Behavioural Modulation: Parametric and Nonparametric Effect Sizes

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

Two complementary measures were selected:

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

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

**Cohen's  $d$ :**

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

Where:

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

**Cliff's Delta  $\Delta$ :**

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

Where:

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

The empirical results yield:

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

Both indices fall within the range typically interpreted as *small to modest* behavioural modulation. Yet as argued earlier, the theoretical significance of these

values does not lie in their magnitude alone, but in the fact that they instantiate a reproducible *directional deformation* of the evaluative transformation  $f(\cdot)$  under controlled manipulation of  $\mathcal{R}$ .

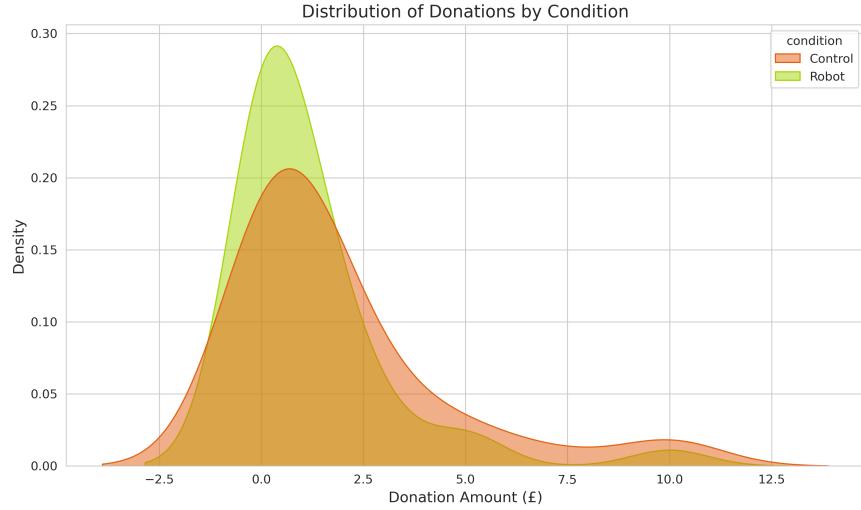


Figure 5.5: Kernel density estimates of donation distributions across conditions. The Control group exhibits higher central mass and a heavier rightward extension relative to the Robot group, consistent with a directional attenuation of high-value prosocial acts in the presence of the synthetic co-presence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

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

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

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

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

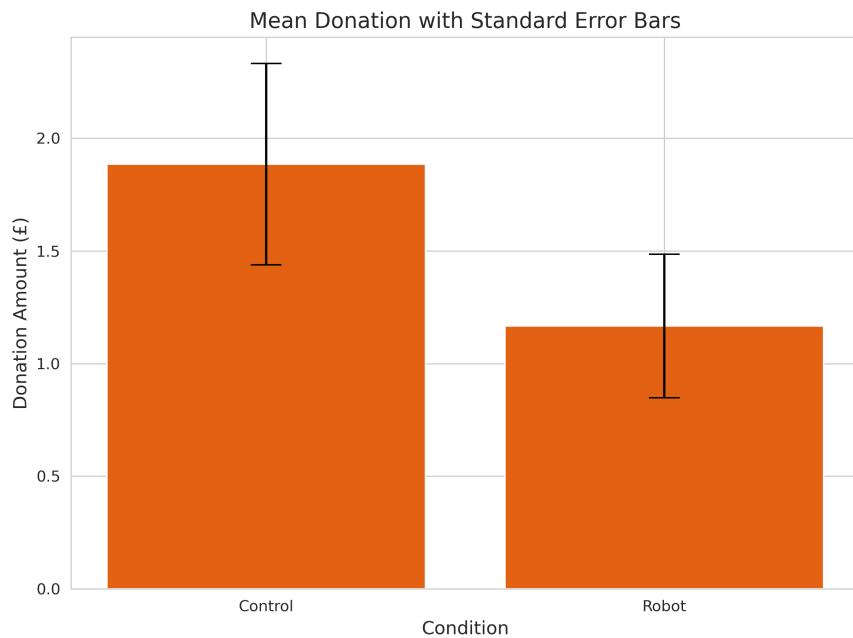


Figure 5.6: Mean donation amounts with standard error bars by condition. The Control group donates more on average (£1.89) than the Robot group (£1.17), corroborating the hypothesis that robotic presence modulates—rather than eliminates—the evaluative pathway from moral salience to action.

#### Conclusion: Amplitude of Moral Refraction

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

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

#### 5.4 Dispositional Baseline: Big Five Personality Traits Across Conditions

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

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

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

#### 5.4.1 Between-Condition Differences in Big Five Personality Traits

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

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

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

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

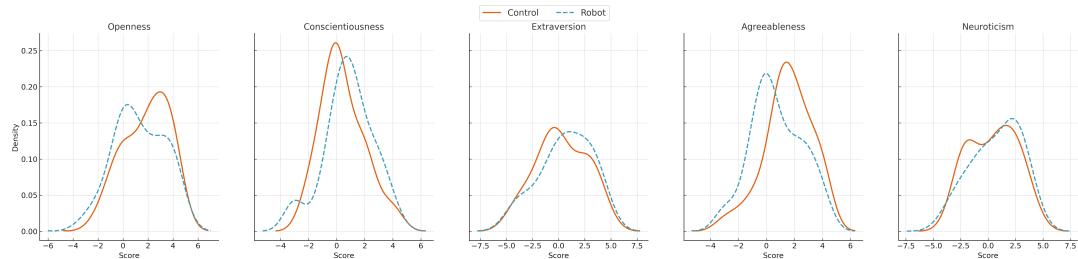


Figure 5.7: Kernel density estimates for each Big Five trait across experimental conditions, demonstrating substantial distributional overlap.

#### 5.4.2 Predictive and Moderating Roles of Big Five Traits

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

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

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

For the moderation question, interaction models of the form

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

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

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

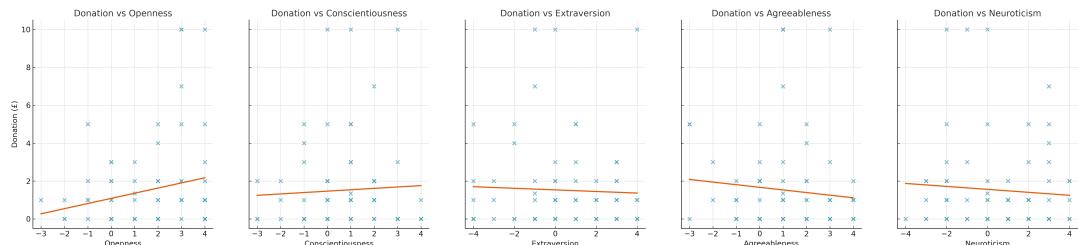


Figure 5.8: Scatter plots with fitted regression lines for each Big Five trait against donation amount. Each panel displays individual participant scores alongside a smoothed linear trend. No clear predictive relationships emerge, reinforcing the conclusion that the Big Five traits do not meaningfully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

#### 5.4.3 Interpretive Synthesis

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

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

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

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

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

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

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

#### 5.4.4 Latent Trait Structures and Individual Modulation of Moral Perturbation

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

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

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

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

from a mere vector of trait scores into a set of structurally defined personality constellations.

Seven variables were included in the initial psychometric space: Empathizing, Systemizing, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each participant's score vector was  $z$ -standardised and submitted to Principal Component Analysis (PCA). Two orthogonal principal components were retained, capturing the most informative axes of variance in the trait space while reducing dimensionality and mitigating redundancy among correlated measures.

The resulting two-dimensional representation was then subjected to  $k$ -means clustering with  $k = 3$ , yielding three psychologically interpretable personality clusters. These clusters were visualised in the reduced PCA space to assess structural separability and interpretative coherence (Figure 5.9).

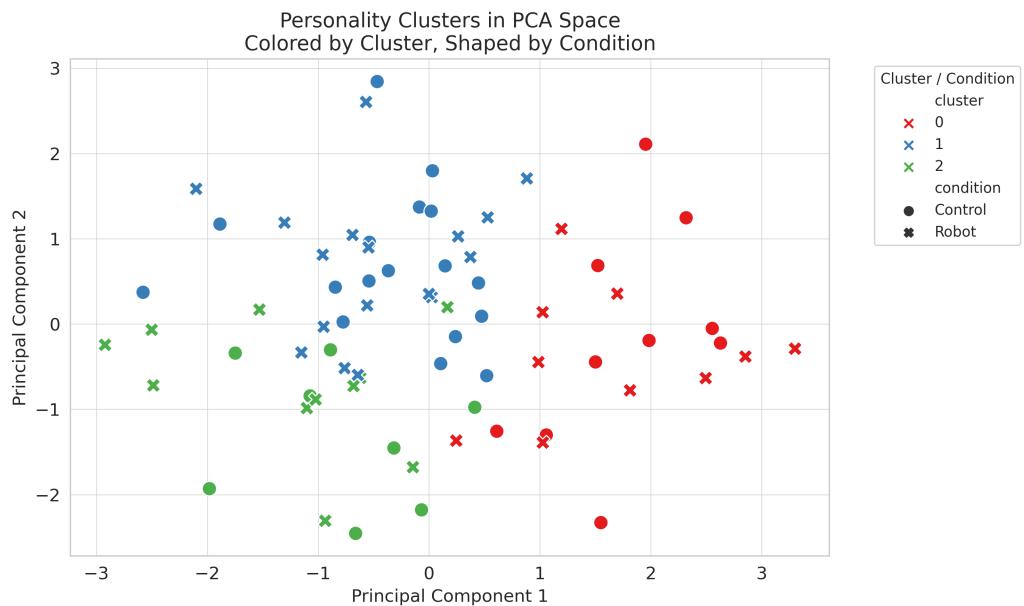


Figure 5.9: Participants clustered in PCA-reduced psychometric space, coloured by cluster identity and shaped by experimental condition. The clustering reveals three latent personality regimes, each representing a distinct cognitive-affective configuration encoded in  $\beta_C$ .

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

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

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

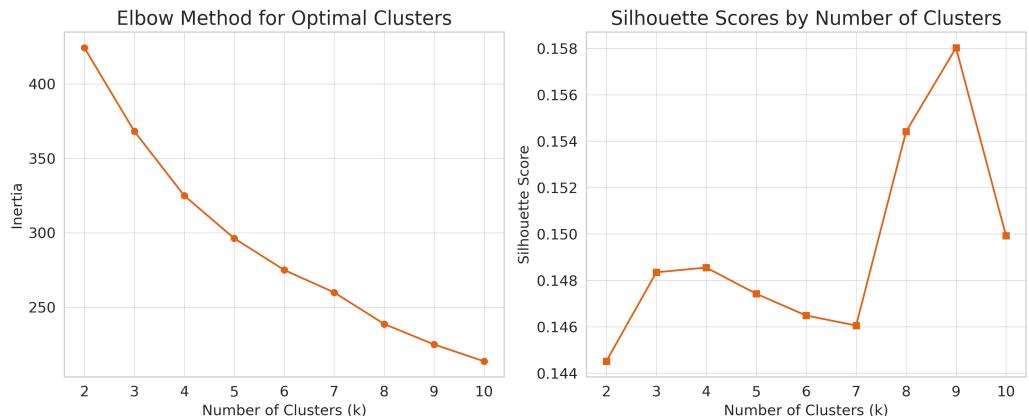


Figure 5.10: Elbow plot of within-cluster sum of squares (left axis) and silhouette coefficients (right axis) across candidate values of  $k$ . The elbow at  $k = 3$  and interpretable silhouette profile support the selection of three clusters as a parsimonious and psychologically meaningful solution.

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

Cluster-specific analyses of donation behaviour reveal heterogeneous responses to moral cues across these latent regimes (Figure 5.11). In one cluster (Cluster 1), the presence of the robot appears to strongly attenuate donation amounts,

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

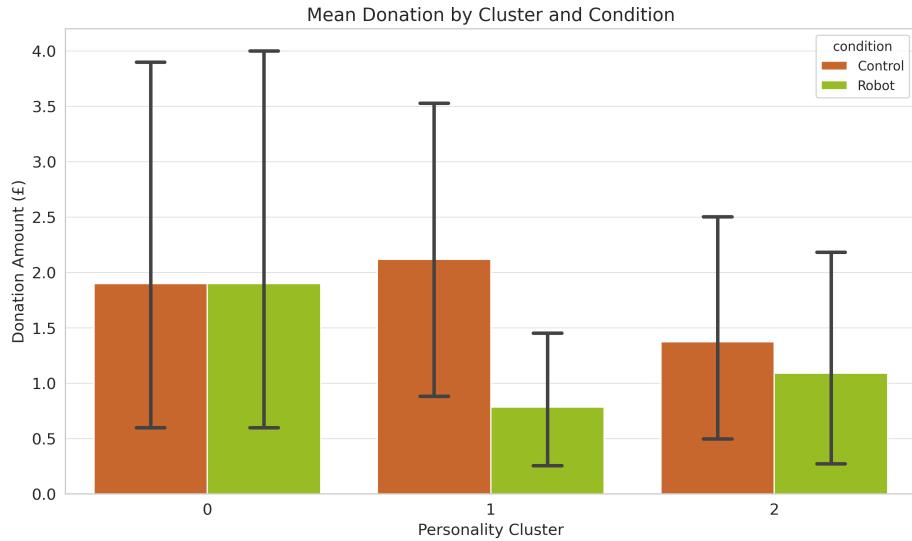


Figure 5.11: Mean donation amount by experimental condition within each personality cluster, derived from  $k$ -means analysis on psychometric trait profiles. Error bars represent standard deviation. Cluster 1 shows a marked attenuation of donation under robotic presence, whereas Clusters 0 and 2 exhibit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

#### 5.4.5 Psychometric Interpretation and Semantic Labelling of Latent Personality Clusters

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

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

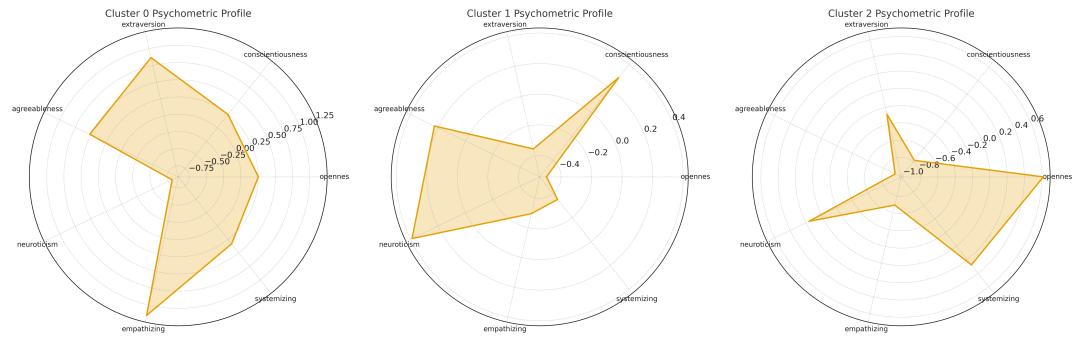


Figure 5.12: Comparative radar profiles of the three latent personality ecologies. **Emotionally Reactive / Low-Structure Profile** (left): elevated Neuroticism with reduced Conscientiousness and Systemizing. **Prosocial–Empathic / Warm–Sociable Profile** (centre): high Openness, Extraversion, Agreeableness,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

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

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

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

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

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

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

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

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

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

### Conclusion: Contingent Structure of Cognitive Modulation

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

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

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

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

- (1) A measurable attenuation of prosocial donation under robotic co-presence (section 5.3.15);
- (2) A structurally heterogeneous, cluster-contingent modulation of this attenuation (section 5.4.5).

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

*Status of the Hypotheses*

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

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

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

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

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

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

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

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

*Condensed Status of the Formal Framework*

The mathematical apparatus introduced earlier has now been substantively activated:

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

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

has proven essential: –  $\alpha_E$  held constant (Watching Eye), –  $\beta_C$  refined via PCA and clustering, –  $\gamma_R$  isolated as the only experimental manipulation.

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

### *Topological and Ontological Interpretation*

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

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

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

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

#### Interim Conclusion: Topological and Trait-Dependent Moral Modulation

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

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

Within the present experimental design, the morally salient cue was instantiated through the photograph of an infant in need, prominently displayed on the Operation Smile brochure. As established earlier (see ??), such pictorial stimuli operate as *implicit moral surveillance cues*: they trigger affective empathy, reputational sensitivity, and the pre-reflective sense of “being observed” that underlies the classical Watching Eye effect [2, 210, 212].

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

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

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

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

where  $\Sigma_{\text{eye}}$  denotes the moral-affective perceptual space dominated by the infant’s image. Under robotic co-presence, this mapping becomes

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

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

Thus, the Watching Eye stimulus does not lose its moral force; rather, its *evaluative amplitude* is refracted by the semiotic presence of the robot, producing a diluted conversion of moral salience into prosocial action. This interpretation coheres with both the ‘Amplitude of Moral Refraction’ conclusion (section 5.3.15) and the ‘Contingent Structure of Cognitive Modulation’ conclusion (section 5.4.5), and it reinforces the central claim of this chapter: synthetic presences modulate moral cognition by altering the topology through which normative cues are interpreted, not by erasing those cues.

#### 5.4.8 Cluster-Specific Regression Analysis of Robotic Perturbation

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

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

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

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

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

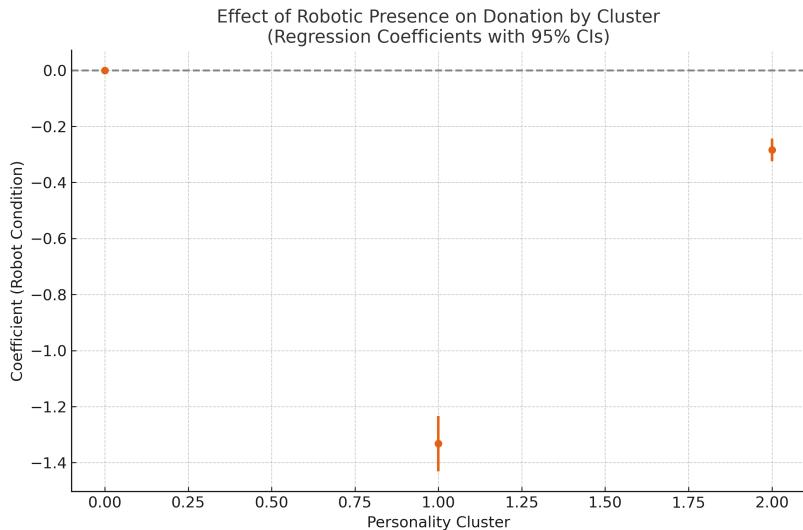


Figure 5.13: Regression coefficients for the Robot condition within each personality cluster (95% confidence intervals). The Prosocial–Empathic profile shows a pronounced attenuation effect, while the Emotionally Reactive and Analytical–Structured profiles exhibit negligible or non-significant coefficients. This pattern demonstrates that robotic presence exerts a differentiated moral influence, contingent on latent cognitive–affective ecologies.

#### Conclusion: Differentiated Moral Sensitivity to Robotic Presence

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

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

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

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

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

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

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

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

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

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

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

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

#### 5.4.9 Bayesian Estimation and Epistemic Gradient Framing

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

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

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

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

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

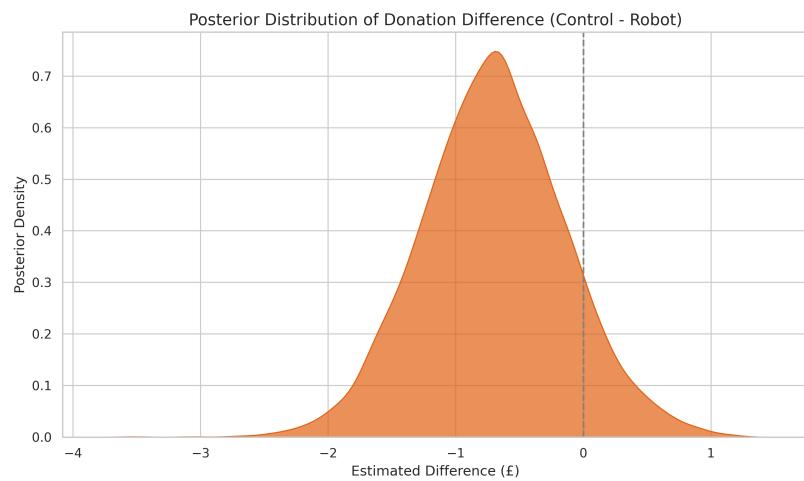


Figure 5.14: Posterior distribution of the estimated donation difference between the Control and Robot conditions. The density skews toward negative values, indicating directional probabilistic evidence that robotic co-presence attenuates prosocial behaviour. The vertical dashed line denotes the point of no effect. Bayesian inference renders the effect size and its uncertainty as a continuous epistemic field rather than a binary verdict.

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

1. **Uncertainty as structure, not noise.** The posterior distribution reflects graded belief over effect magnitudes, aligning with the chapter's emphasis on moral topologies rather than discrete behavioural outputs.
2. **Compatibility with ontological ambiguity.** Robotic presence operates as a *semiotic perturbator* whose influence is subtle, non-deterministic, and context-dependent. Bayesian inference accommodates such phenomena by modelling effect strength as a distribution across epistemic space.
3. **Hierarchical alignment with trait-dependent regimes.** The differential sensitivities observed in the Prosocial–Empathic versus Analytical–Structured profiles, and the near-invariance of the Emotionally Reactive profile, are naturally represented within a Bayesian hierarchical model.

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

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

### *Epistemic Interpretation of the Bayesian Results*

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

In practical terms, the posterior distribution shown in Figure 5.14 does **not** claim that robotic presence definitely reduces donation behaviour. Instead, it says that—given the observed data—the reduction is *more likely than not*. The most plausible magnitude of this attenuation is located around £0.70, but with substantial uncertainty surrounding it. This uncertainty is not a flaw; it is a feature of the Bayesian framework, which makes visible the epistemic limits of the evidence rather than compressing them into a single thresholded output.

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

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

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

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

#### Conclusion: Gradient of the Impact of Moral Refraction

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

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

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

The empirical and probabilistic results obtained thus far permit the first integrated assessment of Question 5.1. Taken together, the behavioural attenuation, the cluster-specific regression patterns, and the Bayesian posterior distribution converge on a coherent interpretative claim: **the silent co-presence of a humanoid robot reshapes the evaluative topology through which morally salient cues become actionable for human agents**. This reshaping is neither universal nor deterministic; it is a graded, structure-dependent perturbation whose amplitude and direction emerge from the interplay of ontological ambiguity, individual trait configuration, and the Level of Abstraction at which the robot is cognitively encountered.

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

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

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

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

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

#### Provisional Answer to Question 5.1

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

Thus, the NAO robot's presence in the room—silent, minimally animated, ontologically ambiguous—modulates moral action not by interrupting reflective deliberation, but by reconfiguring the *interpretive topology* within which morally salient cues acquire behavioural force. The charity poster depicting a child beneficiary of medical aid—our operationalisation of the Watching Eye stimulus—normally functions as an affectively loaded reputational cue, activating empathic concern and third-party moral vigilance [65, 2, 210, 212]. In the Robot condition, however, this prime is **perceptually and semantically diluted**: attentional and inferential resources are partially displaced from the poster toward the robot's embodied but ontologically indeterminate presence. In effect,  $\mathcal{R}$  acts as a *semantic competitor*, weakening the intuitive channel through which the Watching Eye paradigm ordinarily promotes prosocial giving.

This pattern is theoretically coherent within the *Social Intuitionist Model* of moral judgment [16, ?, 17], which holds that moral behaviour is driven primarily by rapid, affect-laden intuitions rather than by reflective cost-benefit deliberation [31, 180, 113]. Under this model, the Watching Eye stimulus shapes

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

These results therefore support an intuitionist, rather than rationalist, interpretation of moral action in this environment. The attenuation effect does not emerge as a failure of explicit reasoning, but as a deformation of the intuitive evaluative processes that precede it. In topological terms,  $\mathcal{R}$  alters the curvature of the moral field: it bends the trajectories along which intuitive appraisals propagate, thereby shifting the probability that moral cues achieve behavioural expression. At the operative *Level of Abstraction* [25, 229, 230], the robot functions as a semiotic intrusion—an entity whose perceived ontology modifies what the agent treats as salient, credible, or normatively relevant.

From a methodological perspective, this interpretation has direct implications for the study of moral cognition. If moral behaviour is mediated by affectively grounded intuitions that are sensitive to environmental structure, then behavioural traces—such as donation decisions—become legitimate datasets for inferring moral evaluations. This aligns with the premises of *Social Signal Processing* [63, 73] and *Affective Computing* [68, ?], which treat observable behaviour as an informational interface through which latent cognitive-affective states may be estimated, modelled, and formalised. The present findings demonstrate that synthetic co-presence can systematically reshape this interface: by altering the distribution of intuitive salience, the robot modifies the behavioural signatures from which moral inference is drawn.

This also intersects directly with the ambitions of *Machine Ethics* [52, 151, ?, 145], which seek to formalise the conditions under which artificial systems may (or may be perceived to) participate in moral contexts. Our results show that even non-interactive robots can perturb moral cognition simply by being *present*—suggesting that artificial agents need not act, speak, or decide in order to exert normative influence. Their moral relevance may emerge from their mere ontological profile, as processed through the observer’s cognitive ecology.

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

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

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

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

From the standpoint of the *Social Intuitionist Model* of moral judgment [16, 31, 17], this pattern is theoretically coherent. Moral action, in this model, is driven primarily by rapid, affectively grounded intuitions rather than reflective deliberation. Our charity poster—operationalising the Watching Eye stimulus—serves precisely as such an intuitive moral prime, designed to trigger empathic concern and reputational awareness. Yet the robot's ambiguous presence dilutes this intuitive channel: the locus of social attention partially shifts from the moral cue to the synthetic body occupying the room, thereby weakening the intuitive pull that ordinarily supports prosocial donation. In topological terms,  $\mathcal{R}$  alters the local curvature of the moral field, redirecting the intuitive flows along which salience is converted into action.

This interpretation is strengthened by Floridi's theory of *Levels of Abstraction* [25, 229]. At the operative LoA of the participant, the robot is encoded not as a machine, nor as a full moral agent, but as an entity whose perceptual affordances (eyes, posture, subtle motion) activate anthropomorphic priors without fulfilling the semantic criteria for agency. In this sense,  $\mathcal{R}$  occupies a liminal ontological position: too animate to be ignored, not animate enough to be treated as an intentional other. The deformation we observe is thus a *semantic deformation*, produced by a presence that inserts ambiguity into the participant's perceptual-moral ecology.

This result has substantial implications for the study of moral cognition. First, it provides empirical support for the thesis that **moral meaning is environmentally scaffolded**: small shifts in perceptual context can reorganise the evaluative machinery that underpins prosocial action. Second, it demonstrates that **moral behaviour is accessible through behavioural signatures**, a fact that aligns with the methodological aims of Social Signal Processing [63] and Affective Computing [68]. If moral action can be systematically perturbed by manipulating

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

**A Critical Note on Machine Ethics.** The present findings also cast a critical light on the current state of Machine Ethics. Much of the Machine Ethics literature has historically been driven by the ambition to design “ethical agents” endowed with explicit moral rules, reasoning procedures, or decision architectures [52, 20, ?]. In the era of LLMs, this ambition has often been rearticulated as the attempt to “align” models with moral norms via fine-tuning datasets, reinforcement feedback, or rule-based guardrails.

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

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

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

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

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

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

The empirical and formal work developed in this chapter allows us to return to Question 5.1 with a more determinate answer. The evidence now supports the following claim: *the silent co-presence of a humanoid robot can, under specific psychological configurations, attenuate the conversion of morally salient cues into prosocial action*. This attenuation is modest in magnitude, probabilistic rather than deterministic, and concentrated within particular evaluative regimes—most notably the Prosocial–Empathic profile—yet it is real, structured, and epistemically tractable.

Topologically, the NAO robot functions as a local deformation of the moral field. The charity poster depicting a child in need, originally introduced as a canonical Watching Eye stimulus, constitutes an affectively loaded attractor in the evaluative landscape: under ordinary circumstances, it pulls intuitive appraisals towards prosocial donation through mechanisms of reputational concern, empathic resonance, and implicit monitoring [65, 2, 212]. Our results indicate that the introduction of  $\mathcal{R}$  partially redistributes this moral salience. For participants in the Prosocial–Empathic regime, the robot operates as a competing focus of attention and an ontologically ambiguous social cue; the intuitive channel that would normally connect the poster to donation behaviour is weakened, re-routed, or locally disrupted.

This pattern aligns with Social Intuitionist accounts of moral judgement, according to which moral action is driven primarily by fast, affect-laden intuitions, with explicit reasoning playing a largely post-hoc justificatory role [16, 31, 17]. In this frame, the Watching Eye effect is not a matter of explicit calculation but of intuitive salience. The robot’s presence does not “argue against” giving; rather, it changes what is experientially foregrounded as normatively relevant. For those whose moral cognition is heavily scaffolded by empathic and reputational cues, NAO’s ambiguous status as quasi-agent and quasi-object suffices to dilute the intuitive force of the poster. For the Analytical–Structured profile, by contrast, the same presence appears to be normatively inert, processed more as a stable environmental feature than as a moral signal. The experiment thus vindicates an ecological, intuitionist interpretation of moral modulation: synthetic presence bends the trajectories of intuitive appraisal rather than intervening at the level of explicit principle application.

Floridi’s theory of Levels of Abstraction (LoA) provides the metaphysical and methodological vocabulary to articulate this deformation [25, 229, 230]. At the operative LoA of the participant, NAO does not appear as a set of internal states or source code, but as a semiotic bundle: body, gaze, posture, micro-movements.

These features instantiate *semantic affordances* that are picked up by different evaluative ecologies in different ways. For the Prosocial–Empathic regime, the robot is encoded as a kind of morally pregnant presence that competes with the child’s image for attentional and normative priority; for the Analytical–Structured regime, the same presence is filtered as structurally irrelevant to the donation decision. The experiment thus realises, in a controlled setting, Floridi’s claim that artefacts can acquire moral salience via their informational role, without being moral agents in any robust sense [230]. NAO is not a locus of *moral agency* here; it is a perturbation in the *informational environment* that reconfigures the mapping from salience to action.

Framed in this way, the present study also exposes a set of limitations in the prevailing discourse of *Machine Ethics*. Much of that literature has centred on the design of explicitly “moral” or “ethical” machines—systems that implement deontological rules, compute consequences, or learn norms, in order to make or justify decisions in ethically acceptable ways [244, 23, 52, 245, 246]. In its canonical formulations, machine ethics presupposes a relatively sharp boundary between human users and artificial moral agents, and locates the core normative challenge in the internal architecture of the latter. Our findings suggest that this focus is, at best, incomplete.

First, the experimental results show that synthetic systems can exert morally relevant influence *without* possessing any explicit ethical architecture at all. NAO neither represents moral principles nor optimises outcomes; it simply occupies space, moves minimally, and is seen. Yet this is sufficient to alter the aggregate pattern of prosocial giving, and to do so selectively across latent cognitive–affective regimes. A research agenda that concentrates on endowing machines with codified moral theories, while neglecting their role as perturbative presences within human evaluative topologies, risks a kind of *conceptual hollowing*: the label “machine ethics” is retained, but the most pervasive moral effects of machines—those mediated through human intuition and social cognition—are left untheorised [244, 23, 230].

Second, the canonical architectures of machine ethics were developed with relatively transparent, modular systems in mind: rule-based agents, deliberative planners, or learning systems whose internal representations could, in principle, be inspected and constrained [23, 29, ?]. Contemporary large language models, recommender systems, and socio-technical platforms do not fit this template. As Coeckelbergh has argued, current AI increasingly generates *simulacra of ethical deliberation*: outputs that *look* like moral reasoning, yet lack robust ties to accountability, context, or genuine normative commitment [54]. In such an environment, the question “how do we encode ethics into a machine?” becomes technically underdetermined and politically misleading. What our data illustrate instead is a different, and arguably more urgent, question: *how do artificial systems and environments shape the informational fields within which human moral cognition operates?*

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

from Social Signal Processing and Affective Computing become central: they treat behaviour, interaction patterns, and expressive cues as data structures from which latent evaluative states can be inferred [63, 68]. Our findings show that the same apparatus can be used not only to analyse human moral action, but to detect and quantify how that action is modulated by synthetic co-presence. The relevant question for machine ethics then becomes not “what principles shall we encode?”, but “how do specific technological affordances reshape the signal-to-inference mapping through which moral salience becomes behaviour?”

Taken together, the chapter’s results therefore support a shift from *agent-centric machine ethics* to an *ecological ethics of synthetic presence*. The NAO robot, as deployed here, is not a moral agent to be judged, but a designed perturbation that reveals structural vulnerabilities in human evaluative systems. Its impact is LoA-dependent, personality-contingent, and epistemically graded. For an ethics of AI and robotics that aspires to be both philosophically serious and empirically grounded, the appropriate research goal is not the engineering of artificially virtuous minds, but the mapping and regulation of the moral topologies in which human and artificial systems are jointly embedded [230, 247, 34]. In this sense, the experiment does not solve the problem of machine ethics; it reframes it. Rather than asking whether robots can be moral, it asks how their mere presence redistributes moral salience, and how such redistributions can be measured, understood, and normatively governed in a world increasingly saturated with synthetic others.

## 6. ETHICAL COGNITION AND NORMATIVE FOUNDATIONS

### 6.1 From Moral Cognition to Ethical Theory

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

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

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

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

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

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

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

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

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

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

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

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

Ethical theory, in its classical formulation, treats moral judgment as the outcome of structured deliberation: a process mediated by reasons, principles, and the articulation of normatively defensible positions. Yet this picture has long been recognised as descriptively incomplete. Human moral behaviour rarely emerges from extended reflection; rather, it unfolds through rapid, affectively mediated evaluations shaped by perception, context, and embodied interaction (see discussion in Chapter 3). The distance between what people *ought* to do, what they *think* they do, and what they *actually* do is substantial. To understand moral action in practice—particularly in technologically saturated environments—ethical inquiry must therefore be coupled with the empirical machinery of moral psychology.

For computing science, this coupling is not optional. Artificial agents are increasingly situated in social contexts where their presence, form, and behaviour modulate human inference, expectation, and decision-making. Fields such as *Social Signal Processing* [63] and *Affective Computing* [68] have already demonstrated that human social cognition is deeply sensitive to subtle cues: gaze, posture, micro-expressions, spatial orientation, and embodied co-presence. These cues structure the “interaction order” [?] within which humans interpret intention, assign agency, and evaluate normatively significant behaviour. When synthetic systems enter this order, they perturb it—not through explicit commands, but by altering the informational and affective landscape in which human cognition operates.

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

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

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

Framed through the lens of ethical theory, the foregoing claim has deeper implications. Ethics, as understood in contemporary philosophy, is a *second-order discipline*: it does not produce moral judgments, but seeks to analyse, justify, or critique them [37, 164, 248]. It examines the *structure* of reasons, obligations, and values, not the psychological mechanisms that generate first-order moral appraisals. The field of Machine Ethics has historically blurred this distinction. By attempting to **engineer** “ethical agents” directly at the level of second-order principles—rule sets, deontic logics, utility functions—it tacitly presumes that moral behaviour can be derived from explicit normative propositions [52, 20]. This presumption is philosophically naïve and empirically untenable. It treats ethics as if it were a generative model of behaviour, rather than a reflective framework that presupposes the very psychological capacities it seeks to evaluate. In doing so, classical Machine Ethics mistakes the normative *grammar* of moral theory for the mechanistic *causality* of moral cognition.

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

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

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

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

In the chapters that follow, we develop this integration along three axes. First, we introduce foundational ethical concepts—deontic, consequentialist, and virtue-theoretic—that define the normative landscape in which moral behaviour is interpreted. Second, we examine the empirical architecture of moral cognition, with emphasis on intuitionist and dual-process models [16, 31, 17] that capture the rapid, affectively-driven nature of everyday moral judgment. Third, we link these philosophical and psychological constructs to the computational disciplines that analyse social behaviour—most notably Social Signal Processing and Affective Computing—thereby establishing a unified framework for studying ethical decision-making in environments populated by artificial agents.

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

### 6.3 Ethical Theory as Second-Order Analysis

If the introductory sections of this chapter establish the transition from first-order moral cognition to second-order normative reflection, the next task is to make explicit the methodological consequences of this shift. The distinction is not merely terminological. It determines which claims are explanatory, which are justificatory, and which are subject to empirical constraint. Failure to maintain this distinction has led to recurring conceptual errors in both philosophical ethics and computational modelling [93, 95, 249, 25, 52, 151, 250, 54]. This section therefore articulates a principled account of what second-order ethical theory *is*, what it *explains*, and what it *cannot* plausibly do.

#### 6.3.1 Ethical Reflection and the Second-Order Stance

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

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

ation that are not themselves the proximate causal mechanisms of moral behaviour [16, 31, 60, 251, 88, 252, 253]. Sidgwick already insisted on this point in *The Methods of Ethics*, where he distinguished between the psychology of moral sentiments and the “*method* of determining right conduct” [36, Book I]. Lemos’s treatment of epistemic justification exhibits a similar structural separation between doxastic psychology and the normative assessment of belief [?]. The parallel here is instructive: ethics stands to moral judgment as epistemology stands to belief-formation.

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

### 6.3.2 Levels of Abstraction and the Proper Location of Ethical Explanation

The distinction between first-order moral cognition and second-order ethical theory can be sharpened through Floridi’s framework of *Levels of Abstraction* (LoA) [25, 229]. On this account, every explanatory enterprise selects a perspective defined by its observables, its conceptual resolution, and the class of questions it is equipped to answer. Moral cognition and ethical theory do not merely operate at different LoAs—they answer *different kinds of questions* and employ *different explanatory primitives*.

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

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

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

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

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

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

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

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

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

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

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

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

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

Classical Machine Ethics provides a clear illustration of the dangers of LoA confusion. A recurring methodological assumption in early systems was that normative concepts themselves—obligations, duties, utilities, or virtues—could be implemented at the computational LoA and thereby function as direct generators of behaviour. Early top-down approaches treated ethical theory as if its abstractions could be operationalised without remainder. For example, Arkin’s “ethical governor” encoded deontological constraints derived from Just War Theory as behavioural regulators [22]; Anderson and Anderson’s principlist architectures computationalised Rossian *prima facie* duties as decision rules [254, 20]; and logic-based approaches by Bringsjord and colleagues modelled deontic operators as executable action-selection mechanisms [21, 255]. Parallel lines of work assumed that utility functions could serve as moral evaluators in consequentialist agents [256, 22], while virtue-theoretic systems attempted to reify character traits as algorithmic dispositions governing moral performance [257, 258]. In all these cases, normative structures were treated as if they occupied the same LoA as the cognitive mechanisms responsible for actual moral behaviour.

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

### 6.3.3 Evaluative Topology as a Bridge Between Orders

The challenge, then, is not to collapse first-order cognition into second-order theory, but to articulate a structure that permits principled interaction between them without confusing their explanatory roles. *Evaluative topology*, introduced in the Morality Primer (Chapter 3) and returned to throughout this thesis (see Chapter ??), provides precisely such a structure.

Evaluative topology can be naturally situated within a long-standing tradition in computational cognitive science that conceptualises perception, valuation, and action as parts of continuous, dynamical systems rather than discrete symbolic modules. Research in moral psychology already demonstrates that moral cognition emerges from distributed interactions between perceptual salience, affective appraisal, attentional dynamics, and context-sensitive social meaning. Empirical models—from Haidt’s social intuitionism to Greene’s dual-process account—show that moral perception is shaped by multi-dimensional affective and social fields rather than rule-based computations [16, 31, 60]. Neurocognitive analyses extend this point: Nussbaum’s and Churchland’s treatments of emotion as evaluative perception imply a graded, vector-like structure underlying moral appraisals [83, 259]. Likewise, work in social signal processing models interpersonal evaluation as a shifting landscape of cues that modulate 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 shapes the flow from perceptual input to action readiness. Instead of assuming that be-

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

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

At its core, evaluative topology treats the moral landscape not as a set of discrete judgments or isolated principles, but as a *dynamic field* whose configuration determines the pathways through which perception becomes moral action [16, 31, 259, 60, 83, 88]. Its explanatory primitives include:

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

Unlike classical ethical theory, which specifies norms at an abstract and often idealised level [36, 260, 261, 80, 37], evaluative topology is sensitive to the *real-time architecture* of moral cognition. And unlike purely mechanistic models in psychology, which describe causal processes but lack normative structure, topology captures the relational, structural, and counterfactual properties of moral appraisal [16, 31, 60, 88, 36, 37, 80]: how evaluative trajectories *could* unfold under alternative configurations of salience, affect, or context.

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

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

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

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

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

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

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

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

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

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

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

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

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

#### 6.4 The Normative Landscape: Structuring Ethical Theories Through LoA and Topology

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

This reconstruction is guided by two methodological constraints:

1. **Philosophical fidelity:** the theories must be represented in a manner faithful to their canonical formulations in moral philosophy.
2. **Integrative compatibility:** the theories must be articulated in a form that allows principled interaction with the psychological and topological models of moral cognition established in Chapter 3.

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

significance of the empirical perturbations induced by synthetic presence.

#### 6.4.1 The Three Dimensions of Normative Analysis

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

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

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

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

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

#### 6.4.2 Why This Framework Matters for the Experimental Chapter

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

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

Three claims follow directly from the preceding reconstruction:

1. **Moral action depends on the configuration of the evaluative field.** Normative theories specify different sources of authority and diverse mechanisms of action-guidance, but all agree that moral behaviour arises from structured evaluative relations, not arbitrary choice.
2. **Synthetic presence modulates this field by perturbing salience, attention, and affective resonance.** A humanoid robot does not supply new reasons; it reshapes the environment in which reasons become behaviourally operative.
3. **Normative theories must therefore be reinterpreted through the joint lens of LoA and evaluative topology if they are to explain or critique the behavioural perturbations observed experimentally.**

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

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

## 6.5 Deontological Structures: The Architecture of Practical Reason

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

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

The reconstruction must satisfy three constraints:

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

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

### 6.5.1 The Source of Normativity: Rational Agency and the Form of Law

On the Kantian account, moral authority arises from the structure of rational agency. The categorical imperative does not prescribe concrete actions but establishes a formal test for the permissibility of maxims: whether one's maxim could be willed as a universal law [?, 80, ?]. This places the source of normativity at a *higher LoA* than psychological description. It concerns the *conditions of reflective justification*, not the causal mechanisms that generate everyday judgments.

This distinction is essential. Classical Machine Ethics implemented the categorical imperative as a procedural decision rule—an algorithmic operator [254, 20, 21, 262, 255, 22]. But Kant never intended universality tests to function as cognitive processes.<sup>1</sup> Their purpose is normative: to articulate the standards under which a maxim can be defended as consistent with rational agency. Treating these tests as computational procedures constitutes precisely the LoA confusion diagnosed earlier.

A survey of Classical Machine Ethics reveals this recurring methodological error: the assumption that Kantian constraints, universality tests, or duty-based norms could be directly implemented as procedural decision rules. Early top-down approaches explicitly treated the categorical imperative, or close deontological analogues, as algorithmic operators determining action permissibility. The most widely cited examples are the principlist architectures developed by Anderson and Anderson, where *prima facie* duties are computationalised as weighted decision procedures whose outputs determine ethically “permissible” behaviour [254, 20]. Similarly, logic-based systems developed by Bringsjord and collaborators represent obligations and prohibitions using deontic logic embedded in the cognitive event calculus, thereby converting normative constraints into executable operators that mechanically evaluate action options [21, 262]. Ganascia’s formalisation of ethical rules of warfare follows the same strategy, modelling universally applicable duties as logical conditions that an autonomous agent must satisfy prior to acting [255]. Arkin’s “ethical governor” for lethal autonomous robots

---

<sup>1</sup>See the discussion in [?] and [15] on the reflective rather than psychological status of the categorical imperative.

likewise encodes deontological constraints—derived from Just War Theory and Kantian doctrine—as computational filters that block impermissible actions at run time [22]. In each case, a normative principle originally intended for reflective justification is treated as a psychological mechanism or behaviour-generating operator. As Moor and Coeckelbergh observe, this amounts precisely to the Level-of-Abstraction confusion: normative tests designed for rational self-assessment are misinterpreted as causal algorithms capable of producing moral behaviour [246, 54]. These systems thus instantiate the very conflation at issue—collapsing reflective ethical reasoning into first-order cognitive processing.

### 6.5.2 Mode of Evaluation: Maxims, Duties, and the Structure of Permissibility

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

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

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

### 6.5.3 Action-Guidance: How Normative Constraints Influence Behaviour

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

The answer, consistent with LoA discipline, is twofold:

1. **At the cognitive LoA:** deontological principles do not produce behaviour. Moral action emerges from intuitive appraisal, affective valuation, attentional salience, and controlled modulation—precisely the components analysed in the Morality Primer (Chapter 3).
2. **At the normative LoA:** deontological principles determine which behavioural trajectories can be reflectively justified. They also shape long-term dispositions, thereby influencing the evaluative topology indirectly through moral training, socialisation, and self-constitution.

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

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

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

#### 6.5.4 Deontological Normativity as Topological Invariance

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

These invariants:

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

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

#### 6.5.5 Why Deontology Matters for the Experimental Logic

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

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

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

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

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

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

### **Conceptual Note: Gradient Fields in Consequentialist Topology**

In the topological framework developed across this thesis, a *gradient field* designates a structured evaluative landscape in which each possible action or state of the world is associated with a scalar value—typically representing expected welfare, utility, or outcome-based moral worth. Formally, a gradient field assigns to each point in an abstract space of action–outcome configurations a direction of steepest ascent: the direction in which an incremental shift would produce the greatest increase in expected value. In classical moral philosophy, this structure is implicit in utilitarian reasoning, which assesses actions by their tendency to promote the greatest balance of good over bad consequences [?, 261, 36]. Within this thesis, the notion is used in a non-formal but conceptually rigorous sense: as a way of modelling how consequentialist evaluation imposes directional structure on the moral field, where moral improvement corresponds to movement along the gradient toward higher expected welfare.

A gradient field thus has three key features:

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

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

2. **Directional guidance:** the moral significance of a possible action is given by its vector orientation relative to the gradient; actions are increasingly morally preferable as they align with the direction of steepest ascent.
3. **Sensitivity to empirical structure:** because the gradient depends on expected outcomes, it varies with changes in belief, evidence, context, and the agent's model of the world.

In this topological reconstruction, consequentialist gradient fields do not function as cognitive mechanisms. Human agents do not compute explicit gradients when acting morally, nor do they evaluate global states of the world through analytic integration. Rather, consequentialist structures operate at the *normative Level of Abstraction*: they specify how actions are *justified* in reflective evaluation, not how they are generated in real-time cognition. This LoA separation parallels Sidgwick's distinction between the "point of view of the universe" and ordinary motivational psychology [36, Book IV].

**Interaction with the Evaluative Machinery of Moral Cognition.** Although gradient fields do not describe the causal architecture of moral cognition, they interact with it in conceptually important ways. The evaluative machinery developed in Chapter 3—perceptual salience, affective appraisal, intuitive heuristics, and controlled modulation—does not implement consequentialist reasoning, but it is nevertheless shaped by outcome-related information in several distinct modes:

1. **Salience modulation.** Perceived consequences influence which features of a situation become salient. Potential harm, benefit, or risk amplifies attentional capture, thereby altering the local configuration of the evaluative field even before explicit reasoning occurs.
2. **Affective valuation.** The human affective system registers outcomes (especially those involving harm or welfare) with strong valence. These affective signals act as local gradient approximations: they bias intuitive appraisal toward or away from particular actions in a manner that roughly tracks expected value.
3. **Heuristic extraction.** Over developmental time, agents internalise outcome-sensitive heuristics ("help when it is easy", "avoid causing harm") that serve as psychologically tractable proxies for gradient following. These heuristics allow the cognitive system to approximate consequentialist structure without computing it.
4. **Deliberative correction.** In cases of conflict or ambiguity, controlled processes may approximate aspects of consequentialist evaluation—comparing potential harms or weighing benefits—thereby engaging the gradient field at a coarse-grained level. However, this is slow, effortful, and limited by computational constraints.
5. **Perturbation sensitivity.** Because consequentialist evaluation depends on expected consequences, perturbations to perception, attention, or social meaning—such as the presence of a humanoid robot—can reshape the

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

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

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

## **6.6 Consequentialist Structures: Value Gradients and the Topology of Outcomes**

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

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

### **6.6.1 The Source of Normativity: Welfare, Impartiality, and the Structure of Reasons**

Classical utilitarianism grounds moral authority in the promotion of welfare. In its canonical formulations—Bentham's felicific calculus [?], Mill's qualitative hedonism [261], and Sidgwick's systematic treatment of practical reason [36]—consequentialism maintains that what ultimately matters is the value of outcomes,

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

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

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

As with deontology, these commitments are not descriptive claims about the mechanisms of moral cognition. Sidgwick is explicit that the “point of view of the universe” is *not* the standpoint from which ordinary agents habitually deliberate; it is a standard of justification, not a psychological model of motivation [36, Book IV]. Consequentialism specifies a standard of rightness, not an algorithm that human agents actually implement.

This distinction is crucial for our purposes. Classical Machine Ethics has often treated utilitarian or outcome-based formalisms as if they were *psychologically generative*: reward functions, expected-utility maximisation, or cost–benefit optimisers are proposed not merely as normative ideals but as surrogates for moral cognition itself. Within the LoA framework, this is a category error. Consequentialism operates at the normative LoA; the evaluative machinery described in the Morality Primer (Chapter 3) operates at the cognitive LoA. Any mapping between the two must be justified rather than assumed.

### 6.6.2 Mode of Evaluation: Consequences, Expected Value, and Scalar Normativity

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

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

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

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

But again, computational tractability must not be confused with cognitive realism. Human moral cognition, as reviewed in Chapter 3, does not perform explicit global optimisation over expected outcomes; it operates through heuristic, affective, and context-sensitive processes that are only loosely correlated with the ideals of consequentialist reasoning [252, 31, 16, ?]. Treating human agents as if they literally implemented expected-utility maximisation is therefore another instance of LoA confusion.

### 6.6.3 Action-Guidance Mechanism: From Value Gradients to Behavioural Pressure

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

At the reflective LoA, consequentialism states:

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

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

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

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

#### 6.6.4 Consequentialist Topology: Moral Action as Gradient Following

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

This has several structural implications:

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

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

#### 6.6.5 Why Consequentialism Matters for the Experimental Logic

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

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

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

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

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

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

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

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

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

Deontological invariants and consequentialist gradients capture two important dimensions of the evaluative field, but they remain incomplete without a theory of the *agent* who navigates that field. Virtue ethics—from Aristotle through modern neo-Aristotelian and psychological reconstructions [13, 14, 69, 81]—locates normativity not primarily in constraints or outcomes, but in the *perceptual and dispositional architecture* of the moral agent. This renders virtue ethics particularly well-suited for integration with the experimental findings of this thesis, which show systematic modulation of prosocial action by latent personality dimensions and cluster-level structure in trait space (see Chapter ??).

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

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

perturbations induced by synthetic presence.

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

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

In the virtue-theoretic tradition, normativity originates in the *well-formed character* of the agent, rather than in rules or external valuations. Virtues are not propositional commitments but *stable dispositional patterns* that structure moral perception: they determine what the agent notices, how she evaluates it, and which actions appear salient, fitting, or required [263, 14]. Aristotle's concept of *phronesis*—practical wisdom—captures the idea that virtuous action arises from the *fine-tuned sensitivity* to morally relevant features of a situation [13].

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

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

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

### 6.7.2 Mode of Evaluation: Dispositions as Topological Structure

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

#### (i) Mathematical and Topological Interpretation

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

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

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

We can therefore interpret virtue-theoretic structure as a topological mapping

$$\mathcal{T} : \mathbb{R}^k \rightarrow \mathcal{F},$$

where  $\mathcal{F}$  is the space of evaluative fields. Under this model:

- high-agreeableness clusters exhibit deeper prosocial attractors;
- low-empathy clusters exhibit shallower or displaced prosocial basins;
- high-conscientiousness clusters show increased boundary rigidity for deontic constraints;
- neuroticism modulates sensitivity to evaluation cues (including the Watching-Eye effect).

In virtue-theoretic terms,  $\beta_C$  approximates a parametric description of the agent's *character topology*. This mapping was borne out empirically: different clusters showed markedly different susceptibility to moral deformation under synthetic presence, precisely as a virtue-ethical model predicts.

### *(ii) Connection to Moral Psychology*

Modern moral psychology (e.g. the *moral foundations* approach [264], the *character-based* models of Snow [?], and the *sensitivity-based* accounts of Dancy [66]) emphasises that moral responsiveness is a function of dispositional configuration. Trait-dependent modulation of salience, empathy, and social attentiveness mirrors the classical virtue-theoretic notion that moral judgment depends on habituated perception.

Our empirical data confirm this: the presence of the robot altered prosocial behaviour differentially across personality clusters, demonstrating that the moral field is not homogenous but *character-structured*.

### **6.7.3 Action-Guidance Mechanism: Habituation, Stability, and Situated Sensitivity**

Virtue ethics explains action not by invoking explicit principles or value calculations but through the *habituated, stabilised patterns of salience and response* characteristic of a well-formed agent. This aligns neatly with the dual-process architecture established in Chapter 3:

- intuitive processes are shaped by long-term habituation into affective-perceptual sensitivities,
- controlled processes integrate commitments and identities developed over time,
- behavioural output reflects the stability or fragility of dispositional attractors.

In topological terms, virtues correspond to *deep attractor basins* resistant to perturbation; vices or deficiencies correspond to *shallow or unstable attractors*. This interpretation is supported by both computational models of habit formation [?] and empirical studies of moral perception [?].

### **6.7.4 Virtue-Theoretic Topology: Stability, Curvature, and Susceptibility to Perturbation**

Within the evaluative-topological framework, virtue ethics can be modelled using dynamical systems language:

$$\dot{x} = f(x; \beta_C),$$

where  $x$  is the agent's state in evaluative space and  $\beta_C$  parametrises dispositional curvature. The presence of a synthetic agent introduces a perturbation  $\delta f$  such that

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

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

Crucially:

- for some clusters,  $\delta f$  shifts the trajectory away from the prosocial attractor basin (attenuation of donation);
- for others, the attractor curvature remains sufficiently deep that the perturbation is absorbed;
- for exceptionally prosocial configurations, synthetic presence may even sharpen evaluative focus (rare, but consistent with the upper-tail donors observed).

This constitutes a clear virtue-theoretic phenomenon: moral sensitivity is *trait-dependent*, and synthetic perturbation reveals structural differences in the stability of character.

### 6.7.5 Why Virtue Ethics Matters for the Experimental Logic

Virtue ethics is indispensable for interpreting the experimental results for three interconnected reasons.

**1. Latent Trait Modulation** The experiment confirms that moral perturbation is not uniform: clusters in personality space exhibit distinct patterns of deformation. Virtue theory provides the conceptual vocabulary for understanding these effects as differences in character topology. Prosocial action is more fragile in agents with shallow attractors; synthetic presence perturbs these evaluative structures disproportionately.

**2. Moral Topology Over Trait Space** The mapping

$$\beta_C \mapsto \mathcal{T}(\beta_C)$$

establishes that moral responsiveness is a *function of trait geometry*. This is a virtue-theoretic insight: character is the medium through which the environment's moral affordances are processed.

**3. Machine Ethics Ignores Character Entirely** Classical Machine Ethics frameworks assume that ethical behaviour can be engineered through top-down rules or utility functions. They contain no representation of dispositional structure, no equivalent of  $\beta_C$ , no account of habituation, and no model of trait-dependent sensitivity to perturbation. This makes them incapable of predicting—or even recognising—the character-mediated moral displacement observed in our experiment.

Virtue ethics therefore reveals the deepest limitation of rule-based or utility-based Machine Ethics: moral agency is fundamentally *dispositional*, and no architecture that ignores habituated sensitivity, perceptual tuning, and trait-level topology can claim to model it.

In sum, virtue ethics interprets the experimental findings as a demonstration that synthetic agents perturb moral action by interacting with the agent-specific topology shaped by habituation, character, and perceptual attunement. Where deontology contributes boundary conditions and consequentialism contributes gradient structure, virtue ethics contributes the *curvature of the evaluative manifold itself*: the dispositional geometry that determines how agents absorb, refract, or amplify perturbation.

### **Interim Synthesis: How the Three Normative Frameworks Illuminate the Experimental Findings**

With deontology, consequentialism, and virtue ethics reconstructed through the discipline of Levels of Abstraction and embedded within the evaluative-topological architecture developed across this thesis, we can now articulate their practical significance for the experimental results.

The purpose of this synthesis is not merely classificatory. It is to demonstrate why an empirical study of synthetic social influence *requires* a multi-framework normative lens, and why no single classical theory is sufficient to interpret the perturbations observed in prosocial donation.

#### **1. Deontology: Structural Invariants and the Integrity of Moral Expectation**

In the deontological reconstruction, duties appear as *invariant boundaries* of the evaluative field. The Watching-Eye cue—as developed in Chapter ??—implicitly invokes precisely these invariants: reciprocity, fairness, honesty, and the demand to act as if one’s behaviour were publicly accountable.

- When donation decreases in the Robot condition, the key normative question is whether the perturbation reflects a weakening of sensitivity to these invariants.
- If synthetic presence “flattens” the deontic landscape—attenuating the felt pull of duty—then the perturbation carries ethical weight beyond behavioural variation.
- The deontological analysis therefore provides the vocabulary to distinguish between a mere preference shift and a disruption in the *structural preconditions* of rightful agency.

Empirically, this interpretation is strengthened by the fact that deontic cues (the child’s face, the moral framing of donation) remain constant across conditions. The only structural change is the presence of the synthetic observer. This isolates the robot as a potential *interference with deontic uptake*. A purely psychological description would register the attenuation as behavioural variance; a deontological analysis reveals it as a possible distortion of moral accountability.

## 2. Consequentialism: Gradient Deformation and the Perceived Structure of Outcomes

From a consequentialist standpoint, moral orientation depends on the perceived gradient of expected value. Watching-Eye cues work partly because they shift the perceived payoff structure: being observed increases reputational benefit, reduces social cost, or reinforces anticipated approval.

The robot's presence perturbs this gradient in three ways:

1. It introduces an *ambiguous observer* whose evaluative stance is unclear, flattening or redirecting the perceived payoff of donation.
2. It may compete with, divert attention from, or overshadow the reputational signal emitted by the Watching-Eye stimulus.
3. It may shift the perceived "social meaning" of the interaction, transforming a dyadic human–charity cue environment into a triadic human–robot–cue environment.

In topological terms, the synthetic presence deforms the gradient field: it alters the local slope of the utility landscape surrounding prosocial action. This consequentialist diagnosis captures aspects of the perturbation that the deontological analysis cannot. Whereas deontology cares about the structural integrity of duties, consequentialism cares about the *direction and magnitude of evaluative flow*. The empirical attenuation fits naturally into this model: synthetic presence recalibrates anticipated outcomes, producing a shallower gradient toward the prosocial basin.

## 3. Virtue Ethics: Dispositional Curvature and Cluster-Dependent Sensitivity

The virtue-theoretic reconstruction offers yet another lens—one that matches the empirical findings with remarkable precision. Virtue ethics treats moral responsiveness as a function of dispositional structure: the shape, depth, and stability of an agent's evaluative attractors.

### 6.7.6 Virtue-Ethical Interpretation of Latent Ecologies

The experiment revealed this pattern with striking clarity when analysed through the semantic ecology of the latent trait clusters:

- **Prosocial–Empathic / Warm–Sociable Ecology:** stable, deep prosocial attractors; moral trajectories remained largely invariant under synthetic perturbation. Donation behaviour persisted despite the introduction of the robot, indicating a robust evaluative surface anchored in empathic resonance and interpersonal sensitivity.
- **Emotionally Reactive / Low-Structure Ecology:** shallow, volatile, or displaced attractors; this ecology exhibited the *strongest attenuation* under robotic presence. Their evaluative field displays low structural coherence and heightened responsiveness to contextual cues; the robot's ontological ambiguity therefore diffuses or refracts moral salience at precisely the stage where affective anchoring would normally stabilise action.

- **Analytical–Structured / High-Systemizing Ecology:** intermediate curvature; these participants showed partial but not catastrophic displacement. Their evaluative architecture privileges clarity and rule-structure over affective immediacy, making them comparatively resistant to affective perturbation, but sensitive to disruptions of interpretive coherence.

This ecological differentiation is *not a behavioural epiphenomenon*: it is the virtue-ethical signature of the data. The robot does not act as a uniformly applied suppressor ( $\gamma_R$ ); rather, it functions as a *contextually instantiated perturbator* whose behavioural impact depends on the dispositional topology encoded in  $\beta_C$ .

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

where  $f(x; \beta_C)$  is the endogenous evaluative drift governing each ecological type's baseline moral trajectory, and  $\delta f(x; \mathcal{R})$  is the deformation induced by synthetic presence.

Crucially,  $\delta f$  is *not* constant. Its sign, magnitude, and functional shape vary across ecologies:

- In the **Prosocial–Empathic** ecology,  $\delta f$  attenuates the empathic-affiliative attractor, flattening the gradient that normally drives prosocial donation.
- In the **Emotionally Reactive** ecology,  $\delta f$  interacts with an already unstable evaluative surface, amplifying volatility and producing the sharpest behavioural displacement.
- In the **Analytical–Structured** ecology,  $\delta f$  perturbs coherence rather than affect, leading to partial reconfiguration but not collapse.

This is precisely the prediction of virtue ethics: the moral perturbation induced by  $\mathcal{R}$  is mediated not by rule-following or outcome-calculation, but by the dispositional configuration of the agent—their stable tendencies of attention, valuation, and motivational salience.

Viewed through Floridi's Levels of Abstraction, the latent ecologies constitute **distinct semantic filters**:

- The **Prosocial–Empathic LoA** foregrounds affective and interpersonal cues.
- The **Emotionally Reactive LoA** foregrounds volatility, ambiguity, and contextual instability.
- The **Analytical–Structured LoA** foregrounds coherence, formal clarity, and normative intelligibility.

The robot's ambiguous ontology—neither fully social nor fully inert—is refracted through these LoAs differently, producing *topologically distinct perturbations*. This explains why synthetic presence yields neither a uniform nor a homogeneous effect, but one that is *contingent upon the semantic architecture* that each ecological profile brings to the interaction.

Thus, the virtue-ethical reconstruction, the latent-trait analysis, and the topological model converge: the robot reveals, with unusual diagnostic precision, the dispositional geometry of each ecological type. Moral displacement is not merely a behavioural effect; it is a principled probe into the internal architecture of moral cognition.

## 5. What Machine Ethics Misses

This synthesis also exposes the limitations of classical Machine Ethics:

1. It assumes that behaviour can be derived from explicit rules (deontic codification), ignoring the psychological mechanisms through which duties gain behavioural force.
2. It assumes that welfare optimisation can be implemented directly through utility functions, ignoring the fact that humans do not compute value gradients explicitly and are sensitive to subtle perturbations.
3. It ignores dispositional topology entirely; it has no representation for character, habituation, sensitivity, or cluster-dependent variation.

Thus, Machine Ethics repeatedly commits the LoA mistake: treating normative abstractions as if they were cognitive operators. The experiment demonstrates why this is untenable: moral behaviour emerges from topological dynamics that Machine Ethics has no resources to model.

## 6. Concluding Perspective: Why This Matters Now

The virtue-theoretic, deontological, and consequentialist reconstructions converge on a single insight: the moral significance of synthetic presence cannot be captured by any one normative theory alone, nor can it be reduced to behaviourist regularities. It must be understood as a deformation of the evaluative topology through which agents convert moral salience into action.

The experiment does not merely show that robots change behaviour. It shows *how* they do so: by reshaping deontic sensitivity, altering perceived consequence gradients, and interacting with deep dispositional structures. This threefold interpretation is the normative analogue of the empirical results—and it furnishes the philosophical scaffolding for the sentimental, affect-based account developed next.

The next section introduces sentimental and affect-based accounts, which complement the dispositional framework by modelling the affective vectors that shape the immediate moral landscape and interact with the latent trait structure identified above.

### 6.8 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology

Having reconstructed deontology as topological invariance and consequentialism as value-gradient optimisation, we now turn to the normative framework most

directly implicated in the experimental results: *sentimentalism*. In the sentimental tradition—Hume, Smith, and their contemporary heirs—*moral evaluation arises from patterns of affective resonance*. Moral judgment is not primarily the deliverance of reason nor the outcome of consequence-calculation, but the structured responsiveness of an agent’s affective system to features of the social world [177, 265, 70, 266].

Within the evaluative-topological framework developed in this thesis, sentimentalism can be reconstructed as an **affective vector field**:

$$\mathbf{A}(x) : \mathcal{X} \rightarrow \mathbb{R}^n$$

where  $\mathcal{X}$  is the space of perceived states and  $\mathbf{A}(x)$  encodes the direction and magnitude of affective forces—empathic pull, aversive push, compassion, indignation, warmth, or distress. Moral trajectories emerge from the integration of these affective vectors with attentional, inferential, and regulatory processes.

This reconstruction is not metaphorical. It is empirically realised in the experiment: the robot’s presence attenuates donation behaviour *by dampening the affective vector field*, especially within ecological profiles where empathy, warmth, or affective sensitivity ordinarily serve as the primary drivers of moral salience.

### 6.8.1 The Source of Normativity: Sentiment as the Basis of Moral Appraisal

For sentimentalists, normativity originates in the *patterns of affective response* that constitute the human capacity for moral perception. Hume’s claim that moral distinctions are “more properly felt than judged” [177] is often caricatured; yet properly interpreted, it captures a structural truth about moral cognition: affective resonance is the primary medium through which agents register the moral significance of others.

This maps directly onto the cognitive LoA articulated in the Morality Primer: affective tagging (amygdala, insula), empathic resonance (mPFC–TPJ circuit), and rapid harm appraisal provide the initial topological curvature from which moral trajectories originate.

Where deontology imposes constraints and consequentialism imposes gradients, sentimentalism specifies the *affective geometry* of the evaluative field: how warmth draws the agent toward prosocial action, how distress aversion or fear repel them, and how empathic concern shapes the felt moral landscape.

### 6.8.2 Mode of Evaluation: Affective Resonance as Moral Metric

In the sentimental reconstruction, the mode of evaluation is grounded in:

- **empathic responsiveness** to others’ welfare,
- **reactive attitudes** such as guilt, indignation, gratitude, and resentment,
- **interpersonal attunement** through shared affective states,
- **warmth, sociability, and affiliative motivation**.

These components map precisely onto the **Prosocial–Empathic / Warm–Sociable ecology**. For individuals in this cluster, moral relevance is primarily affective: moral cues are not merely recognised but *felt*, and prosocial donation emerges from empathic attunement to the beneficiary.

Thus, sentimentalism aligns almost point-for-point with the evaluative topology of Cluster *Prosocial–Empathic*. If moral action is the integral of affective forces across the evaluative field, then anything that diminishes the amplitude of  $\mathbf{A}(x)$  will proportionally diminish prosocial behaviour.

### 6.8.3 Action Guidance: Affective Vector Fields and Behavioural Dynamics

The sentimental picture becomes formally precise when expressed as a dynamical system:

$$\dot{x} = f(x) + \mathbf{A}(x),$$

where  $f(x)$  captures neutral evaluative drift and  $\mathbf{A}(x)$  represents affective forces.

Synthetic presence enters the system as a deformation operator:

$$\dot{x}' = f(x) + \mathbf{A}(x) + \delta\mathbf{A}(x; \mathcal{R}),$$

where  $\delta\mathbf{A}(x; \mathcal{R})$  is a vector field that attenuates, displaces, or reorients affective flow.

This model captures the experimental results with exceptional fidelity:

- In the **Prosocial–Empathic ecology**,  $\delta\mathbf{A}$  significantly dampens empathic activation, flattening the trajectory toward donation.
- In the **Emotionally Reactive ecology**,  $\delta\mathbf{A}$  destabilises an already volatile field, producing the strongest behavioural perturbation.
- In the **Analytical–Structured ecology**,  $\delta\mathbf{A}$  is comparatively weak; affective forces are not the dominant drivers of action, so the robot's dampening effect is limited.

In short, the robot modulates the moral field by *reducing the affective curvature* that ordinarily drives prosocial behaviour—a textbook sentimental effect.

### 6.8.4 Contrast with Machine Ethics: The Blind Spot of Affective Architecture

Classical Machine Ethics commits its deepest conceptual error here. It systematically ignores affective architecture, attempting to:

- replace empathic resonance with rule sets,
- replace moral perception with logical inference,
- replace affective appraisal with propositional justification.

No sentimental could make this mistake, because for sentimentalism, affect is neither optional nor ornamental: it is the *substrate* of moral cognition.

Our experiment demonstrates precisely what Machine Ethics ignores: a silent, non-interactive robot can alter human moral behaviour not by violating rules or changing utilities, but by shifting the structure of *affective vectors* that underwrite prosocial responsiveness.

Machine Ethics has no conceptual resources to model this effect. A sentimentalist topology does.

### 6.8.5 Experimental Realisation: Synthetic Dampening of Empathic Resonance

The key empirical finding of the experiment is that robotic co-presence attenuates donation behaviour even in the presence of a strong empathic cue (the Watching-Eye stimulus). From a sentimentalist perspective, this is best understood as:

$$\delta\mathbf{A}(x; \mathcal{R}) < 0$$

for affectively weighted regions of the evaluative field, *where*:

- $x$  denotes the agent's current evaluative state within the moral field;
- $\mathbf{A}(x)$  is the baseline affective vector field that encodes empathic pull, aversive push, warmth, distress, and related affective forces;
- $\mathcal{R}$  represents the presence of the humanoid robot as an environmental perturbator;
- $\delta\mathbf{A}(x; \mathcal{R})$  is the deformation operator modelling how  $\mathcal{R}$  alters the magnitude and direction of affective vectors at  $x$ .

In plain terms, the inequality  $\delta\mathbf{A}(x; \mathcal{R}) < 0$  states that the robot's co-presence *reduces the strength of the affective forces* (especially empathic resonance) that normally propel the agent toward prosocial action. The perturbation does not reverse moral direction; it *dampens* the affective momentum that would otherwise guide the agent toward donation.

for affectively weighted regions of the evaluative field.

The robot introduces ontological ambiguity—neither fully agentic nor wholly inert—which disrupts affective attunement in two ways:

1. **Affective dilution:** the presence of an ambiguous social other diverts empathic focus away from the child-beneficiary.
2. **Affective deflection:** the robot introduces uncertainty about social meaning, reducing the clarity of empathic pathways.

Both phenomena manifest as measurable differences in the experimental clusters:

- **Prosocial–Empathic:** attenuation is diagnostic of diluted empathic resonance.
- **Emotionally Reactive:** attenuation reflects increased volatility of affective vectors.

- **Analytical–Structured:** attenuation is weaker, because affect is not the primary moral driver.

Thus, sentimentalism provides the most *mechanistically precise* explanation of the perturbation: synthetic presence alters the affective landscape through which moral salience becomes moral action.

#### Interpretive Synthesis: Sentimentalism and Synthetic Moral Perturbation

The experimental attenuation of prosocial behaviour under robotic co-presence is a paradigmatic sentimentalist phenomenon. Moral action in the Prosocial–Empathic ecology is driven by affective vector fields whose magnitude is reduced by the robot’s ambiguous ontology; in the Emotionally Reactive ecology, the same perturbation destabilises an already volatile field; and in the Analytical–Structured ecology, affective dampening has limited influence because the evaluative surface is dominated by structural rather than affective curvature.

This tripartite pattern cannot be captured by rule-based models, logical deduction, or utility maximisation. It requires a framework in which *affective forces are constitutive of moral cognition*. Sentimentalism, reconstructed as a vector-field theory of affective appraisal, therefore provides the most illuminating normative interpretation of the experiment.

By revealing how synthetic presence modulates affective resonance across latent evaluative ecologies, the experiment demonstrates that moral displacement is not a failure of principles, nor a miscalculation of outcomes, but a deformation of the affective topology through which moral meaning is experienced. This establishes sentimentalist normativity as an indispensable component of any adequate ethical or computational treatment of artificial agents.

## 6.9 Contractualism, Particularism, and Hybrid Normative Models

The preceding sections developed deontological, consequentialist, and virtue-theoretic structures as topological configurations within the evaluative field. To complete the normative landscape relevant to this thesis, we now introduce three additional frameworks—*contractualism*, *particularism*, and *hybrid or pluralist models*. These theories are reconstructed briefly but precisely, preserving their philosophical integrity while integrating them into the LoA discipline and the evaluative-topological architecture that anchors both the theory and the experiment.

The motivation for introducing these additional models is twofold.

First, they represent influential alternatives to the classical triad of deontology, consequentialism, and virtue ethics. Contractualist theories such as Scanlon’s place justificatory relations at the centre of moral evaluation [37]; particularist and perceptual approaches emphasise context-sensitive moral salience rather than rule-governed invariants [263, 248]; and pluralist or hybrid accounts highlight the inherently multidimensional structure of practical reason [92].

Second, these frameworks illuminate aspects of the experimental data that are not

recoverable from topological invariants, value gradients, or dispositional attractors alone. Deontological rules struggle to accommodate dilemmatic or context-dependent cases [267], outcome-based models fail to capture the intuitive and affective dynamics documented in empirical moral psychology [31], and virtue-theoretic attractors are limited by the instability of global traits [92]. By contrast, theories grounded in justifiability, contextual salience, and multidimensional normative interaction provide precisely the structural resources needed to interpret how synthetic presence modulates prosocial behaviour. Empirical studies show that minimal cues of social evaluation—including robotic presence, perceived agency, or merely the appearance of watching eyes—systematically shift cooperative and moral behaviour [51, 268, 204, 34]. These phenomena demand a normative-cognitive model capable of accommodating justification pressures, situational salience, and relational moral dynamics—dimensions captured by these alternative frameworks.

*A further reason for reconstructing these theories is philosophical and pedagogical rather than merely instrumental.* Any comprehensive treatment of normative foundations—particularly one that aims to integrate ethics with empirical findings and computational modelling—must follow the established structure of the discipline. Contractualism, particularism, and pluralist models represent canonical branches of ethical theory, and omitting them would not only break with the methodological tradition of moral philosophy but would deprive the reader of the conceptual resources required to situate the thesis within the broader normative landscape. Their inclusion therefore serves a dual purpose: it preserves continuity with the philosophical canon, and it ensures that the interpretive tools deployed in analysing the experiment are grounded in a complete and pedagogically robust reconstruction of the field. In short, without these frameworks, the chapter would lack both scholarly completeness and exegetical depth; with them, the reader is equipped to understand how the experimental findings resonate across the full range of contemporary normative theory.

### 6.9.1 Contractualism: Moral Claims as Justification-Equilibria

Contractualism, most prominently articulated by Scanlon [37], grounds moral rightness in the principle that actions must be justifiable to others on grounds that no one could reasonably reject. This account locates the *source of normativity* not in rules, consequences, or character, but in the structure of interpersonal justification.

Within the LoA structure adopted earlier, contractualism operates at the reflective normative level. It specifies the conditions under which agents can regard themselves as mutually accountable. However, it also has cognitive implications: establishing justifiability requires a sensitivity to others' claims, expectations, and burdens, which in turn depends on social perception and empathic attunement.

**Topological Interpretation.** Contractualism can be conceptualised as defining regions of justificatory equilibrium within the evaluative field—zones in which an action can withstand the test of mutual recognisability and reasonable non-rejection. On Scanlon's account, moral principles are valid only insofar as they can be justified to others as part of a shared moral relationship [37]; similarly,

Strawson's analysis of reactive attitudes shows that moral assessment presupposes an interpersonal standpoint in which agents acknowledge one another as answerable participants [169]. These equilibria remain stable only when agents register the presence and perspective of others, since the very structure of contractualist judgment requires a perceived field of accountability.

Synthetic presence therefore interacts with contractualist structure in a distinctive way. A minimal cue of observation—such as a watching-eye stimulus—typically increases the salience of interpersonal accountability and strengthens cooperative norms [268, 204]. However, a humanoid robot, being perceptually social yet ontologically ambiguous, perturbs this social-evaluative field. Empirical work shows that robots can elicit social facilitation effects [51] while simultaneously failing to stably occupy the interpersonal roles through which moral demands are ordinarily mediated [34]. The result is a displacement or dilution of the implicit sense of being under another's evaluative regard, thereby disrupting the justificatory equilibrium that contractualism presupposes.

**Relevance to Experimental Findings.** Contractualism helps explain why the Prosocial–Empathic cluster exhibited the strongest attenuation in the presence of the humanoid robot. Contractualist moral cognition is grounded in the demand that one's actions be justifiable to others under conditions of mutual recognisability [37, 169]. Individuals high in prosociality and empathy are especially sensitive to this interpersonal dimension: their behaviour is strongly modulated by cues of accountability and evaluative regard [?]. Under normal circumstances, the Watching-Eye stimulus enhances this perceived mutual accountability, consistent with findings that minimal social-evaluative cues increase cooperation and generosity [268, 204].

The humanoid robot, however, introduces a distinctive perturbation to the justificatory field. Robots can trigger social cognition and elicit affective responses, yet they occupy an ambiguous interpersonal category—they are perceptually social but not reliably recognised as members of the moral community [34, 269]. Empirical studies show that such synthetic presence can both facilitate and destabilise social behaviour [51]. In this context, the robot displaces the implicit sense of being under the evaluative regard of others, thereby weakening justificatory resonance and reducing donation.

Contractualism therefore interprets the observed displacement effect not as a change in underlying preference or a failure of duty, but as a deformation of the justificatory field: a disruption of the interpersonal conditions under which reasons become mutually recognisable and moral motivations are sustained.

### 6.9.2 Moral Particularism: Contextual Salience and the Fragmented Topology of Reasons

Moral particularism rejects the idea that morality is governed by fixed principles, boundaries, or stable evaluative gradients. On the particularist view, the moral relevance of a consideration is wholly context-dependent: a feature that counts in favour of an action in one case may count against it in another, and reasons possess no invariant valence [66]. This holism of reasons is closely aligned with McDowell's account of moral perception, in which the salience of a consideration

emerges from its role within a concrete situation rather than from any codifiable general rule [263]. Related work in moral epistemology likewise emphasises that the moral field is shaped by context-sensitive patterns of attention and evaluative uptake [248].

Within an evaluative-topological framework, particularism corresponds to a landscape devoid of global structure. Instead of stable invariants or fixed gradients, the field consists solely of local salience contours whose shape shifts with variations in context, attention, or affect. Empirical work in moral psychology supports this characterisation: affective intuitions, perceptual cues, and distributed cognitive processes dynamically modulate which features of a situation are experienced as morally salient [16, 31, 88]. On this view, the evaluative field is fragmented and constantly reconfigured by situational parameters, making moral appraisal an exercise in context-sensitive responsiveness rather than rule-governed inference.

**Synthetic Perturbation Under Particularism.** If moral salience is locally determined, then synthetic presence need not override a stable evaluative map—there may be no stable map to override. Instead, the robot alters the immediate pattern of salience in the environment.

Moral relevance shifts with contextual detail, perceptual attention, and affective orientation; reasons have no invariant valence [66, 263, 248, 16]. On this view, the evaluative field is not globally structured but dynamically reconstructed from moment to moment as agents engage with their environment.

Synthetic presence therefore alters moral appraisal not by displacing a fixed evaluative configuration but by reshaping the local pattern of salience. Watching-eye cues, for example, immediately increase the accessibility of accountability norms [268], while the presence of a humanoid robot modifies attention, affect, and perceived social agency in more ambiguous ways [51, 34, 269]. What changes is not a stable moral map but the salience geometry that determines which features of the situation come to the fore. In a locally structured evaluative landscape, such perturbations directly influence moral appraisal by shifting which considerations are taken to matter.

This explains why the Emotionally Reactive cluster remained largely invariant in the experiment: their evaluative field is already highly sensitive to situational micro-variations; the robot adds noise, but not disruption relative to their already-fluid topology.

For the Prosocial–Empathic cluster, particularism illuminates a distinct mechanism. Because moral salience is locally assembled rather than globally fixed, the introduction of a humanoid robot reorganises the initial salience hierarchy in the scene. Findings from Social Signal Processing demonstrate that socially meaningful agents exert strong bottom-up pressure on attentional allocation, reshaping which cues are processed first and with what priority [67, 73]. In HRI, even minimal humanoid cues have been shown to redirect gaze, amplify social relevance, and restructure the perceptual field through which subsequent evaluation occurs [270, 271, 272]. Psychological studies similarly show that agentive or emotionally charged stimuli modulate attentional capture and suppress competing social cues [273, 47, 49].

For individuals high in prosociality and empathy, the Watching-Eye stimulus typically heightens interpersonal accountability and enhances empathic attunement. However, the robot's ambiguous interpersonal status—neither fully social nor fully non-social—introduces a conflict in salience that overshadows the eye cue. The result is a weakening of empathic resonance with the expected evaluative signal, producing the attenuated prosocial output observed in the experiments. This interpretation aligns with philosophical accounts of moral perception in which what becomes salient first, and how long it remains salient, is constitutive of the evaluative episode itself [263, 70].

### 6.9.3 Hybrid and Pluralist Models: Multidimensional Topologies

Hybrid or pluralist normative theories—from Ross's account of irreducible *prima facie* duties [274] to contemporary value pluralism [115]—maintain that normativity is generated by multiple independent evaluative sources. On this view, moral assessment is not grounded solely in duty, utility, or virtue, but arises from an interplay among distinct kinds of considerations: deontic constraints, outcome-based reasons, character-based appraisals, relational obligations, and contextual factors all exert normative force [275, 276, 80, 37].

In topological terms, pluralism corresponds to a multi-dimensional evaluative manifold. Rather than a single moral axis, the evaluative space contains intersecting gradients, constraints, and dispositional attractors that jointly shape moral judgment. Psychological and neurocognitive models of moral cognition reinforce this interpretation: affective intuitions, rule-based processes, and outcome-tracking mechanisms operate semi-independently and interact dynamically [16, 31, 259]. Moral judgment, on this pluralist understanding, involves navigating a field whose geometry reflects the heterogeneity of moral reasons, none of which dominates the space entirely.

**Why Pluralism Fits the Experiment.** The experimental findings align naturally with a pluralist topology:

- The Watching-Eye cue activates deontic expectations (being observed).
- Prosocial donation expresses consequentialist gradients (benefit to others).
- Cluster differences reflect dispositional factors (virtue-theoretic structure).
- The robot's ontology refracts social meaning (contractualist relevance).
- Synthetic perturbation shifts local salience (particularist sensitivity).

No single normative theory fully explains the displacement effect observed in the experiment; the phenomenon does not map cleanly onto duty-based invariants, outcome gradients, virtue-theoretic dispositions, or empathy-driven models alone [277, 267, 31, 92, 49]. Instead, the attenuation of prosocial behaviour in the presence of a humanoid robot appears to arise from a reweighting across multiple normative dimensions simultaneously. This is precisely what a pluralist topological model predicts. If moral judgment is guided by a manifold of intersecting evaluative gradients—deontic constraints, empathic pull, reputational expectations, contextual norms, and outcome-based considerations—then perturbing the

structure of the environment can alter the geometry of this manifold as a whole [274, 115, 259].

The experimental findings provide empirical support for this view. The robot's mere presence displaced prosocial action across participants irrespective of dispositional differences: personality traits, empathizing and systemizing profiles, and even latent psychometric clusters failed to moderate the effect. This indicates that the perturbation operates not at the level of individual evaluative tendencies but at the level of the evaluative field itself. The robot's perceptual salience and ontological ambiguity modulate several normative gradients at once—attenuating empathic resonance, altering implicit social expectations, and shifting the perceived normative demand of the situation [34, 186, 272, 269, 51, 217]. In particular, the Watching-Eye cue's typical facilitation of prosocial behaviour is dampened by competing or ambiguous social cues, consistent with findings on accountability cues, attentional capture, and salience competition in social signal processing and psychology [268, 204, 47, 49, 67, 73]. Rather than reinforcing or suppressing any single source of moral motivation, the robot reconfigures the topology within which diverse moral reasons are weighed and integrated.

In this respect, the experiment does more than illustrate a behavioural effect: it offers empirical evidence for the central insight of normative pluralism—that moral judgment is sensitive to the configuration of multiple evaluative dimensions, any of which may be displaced by contextual perturbation [274, 115, 37]. The displacement effect observed here thus constitutes a concrete instantiation of pluralist topology: an environmental shift that alters the manifold of moral reasons as a whole rather than modulating a single axis of moral evaluation.

It is important to emphasise that the field-level displacement effect demonstrated in the experiment does not contradict the presence of stable dispositional differences identified through our clustering analysis. The three psychometric clusters reflect distinct starting positions within the evaluative manifold—different dispositional orientations that shape how individuals ordinarily navigate moral contexts. However, the robotic perturbation operated not on these dispositional baselines but on the shared topological structure of the evaluative field itself. This is why all clusters, despite their psychological differences, exhibited the same directional attenuation in prosocial action. In pluralist topological terms, the robot alters the geometry of the manifold as a whole rather than modulating any single dispositional gradient. The cluster analysis and the displacement effect thus describe two complementary layers of moral cognition: stable trait-like orientations, and a context-sensitive evaluative field capable of being globally reshaped by environmental factors such as synthetic social presence.

#### **6.9.4 Integrative Ethical Interpretation of the Experimental Findings**

Bringing the reconstructed frameworks together, we can now articulate the ethical significance of the experimental results in a way that reflects both the normative pluralism developed earlier and the dual-layer structure of moral cognition revealed by the empirical findings. The donation attenuation produced by robotic presence does not reflect the modulation of a single moral principle, evaluative dimension, or dispositional trait. Rather, it arises from a global perturbation of

the evaluative field in which diverse moral reasons are ordinarily weighed and integrated.

1. **From a deontological perspective**, the robot weakens the felt presence of a morally relevant observer and thereby attenuates the sense of duty-oriented accountability that the Watching-Eye cue is designed to amplify. The displacement effect represents a disruption of implicit normative expectations rather than a violation of explicit moral rules.
2. **From a consequentialist perspective**, the robot alters the perceived payoff structure of helping behaviour by flattening the social-evaluative gradient. The expected “return” of prosocial action—whether reputational, emotional, or anticipatory—becomes less sharply defined, shifting the cost–benefit landscape in a way that depresses altruistic output.
3. **From a virtue-ethical perspective**, the perturbation does not target trait-dependent motivational tendencies directly. Rather, it reveals that even robust dispositional architectures (as captured in the three psychometric clusters) can be globally displaced by contextual features that reshape the evaluative field within which character expresses itself. The fact that all clusters show the same behavioural shift indicates that virtue is not a self-contained driver of action, but a gradient subject to field-level modulation.
4. **From a contractualist perspective**, the robot disrupts the local justificatory equilibrium by diminishing the perceived presence of agents to whom reasons are owed. The justificatory landscape becomes noisier and less structured, thereby reducing the motivational force of the requirement to “act in ways that others could not reasonably reject.”
5. **From a particularist perspective**, the robot modifies the fine-grained salience structure of the environment, reconfiguring which contextual features become normatively operative. The eyes cue remains physically present, but its moral traction is displaced by a new source of salience—an ambiguous agent whose social meaning is not yet assimilated into the participant’s normative schema.
6. **From a pluralist-topological perspective**, the findings are precisely what we should expect when multiple normative gradients interact with a global perturbation to social meaning. The donation attenuation is not the suppression of a single moral principle; it is the displacement of the evaluative manifold itself. This explains why no single theory—deontological, consequentialist, virtue-based, contractualist, or particularist—captures the full phenomenon, and why the effect persists across dispositional clusters.

Taken together, these interpretations converge on a unified thesis:

### Integrative Conclusion: The Ethical Signature of Moral Displacement

The presence of a humanoid robot reshapes the multi-dimensional evaluative topology through which moral salience becomes action. This perturbation operates at the level of the evaluative field, modulating deontic expectations, consequentialist gradients, dispositional attractors, justificatory relations, and contextual salience structures simultaneously. No monolithic ethical framework captures the phenomenon. The experimental results therefore vindicate a pluralist, topological, empirically grounded account of moral cognition—one that reveals how synthetic agents can globally displace moral evaluation in ways systematically overlooked by classical Machine Ethics.

By reconstructing the major normative theories through LoA discipline and embedding them within a topologically structured model of moral cognition, this chapter provides the conceptual architecture required to understand the ethical significance of synthetic moral perturbation. The experiment that follows empirically demonstrates how such perturbation manifests as a field-level displacement effect, thereby linking the normative, psychological, and computational analyses into a unified account of how synthetic agents influence moral behaviour.

## 7. General Discussion and Theoretical Integration

### 7.1 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 co-presence 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, 278, 248, 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, 181]. Because such events culminate in observable behavioural outputs, they are empirically tractable and available to systematic measurement and analysis [208, 209, 207]. 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, 248, 279, 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, 280, 281, 282, 283, 284, 285]. 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 [286, 287, 288, 289, 290, 291]. Even approaches motivated by psychological plausibility, such as computational models of ethical reasoning [250, 24, 292], 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, 293, 294, 295, 54, 246]. 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, 296, 67, 181]. 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 [208, 209, ?]. 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, 297, 298].

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.

### 7.1.1 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  $\gamma_R$  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 [299], personality variation captured by the BFI-10 [184], and broader accounts of moral-psychological “ecologies” that organise evaluative processing [181, 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 [259, 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 [268, 204], 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 6. Each theory identifies a different locus of normative structure, and each provides a distinct—yet convergent—reading of the deformation caused by  $\mathcal{R}$ :

- *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:  $\delta\mathbf{A}(x; \mathcal{R}) < 0$ . 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.

### 7.1.2 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.

### 7.1.3 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  $\gamma_R$  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 [299, 88]. 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, 263].

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.

#### 7.1.4 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.

## 7.2 Cluster-by-Cluster Integrative Interpretation

The experimental results demonstrate that robotic co-presence  $\mathcal{R}$  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  $\mathcal{R}$  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 [299]. 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 [300, 301]. 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 [302, 303].

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 [304, 305, 306]. 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  $\mathcal{R}$  to deform.

Within such a landscape, the experimentally observed pattern—minimal or noisy attenuation—is theoretically revealing. The Watching-Eye stimulus  $\sigma_{WE}$  generates only a modest prosocial gradient for this cluster [268, 204], and the robot-induced perturbation  $\gamma_R$  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  $\mathcal{R}$  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 [268], 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  $\delta\mathcal{E}(x; \mathcal{R})$  acts *upstream*, modifying the salience structure itself:

$$\delta\mathcal{E}(x; \mathcal{R}) < 0, \quad \delta\mathbf{A}(x; \mathcal{R}) < 0.$$

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 [299]. 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  $\mathcal{R}$  introduces representational and social-ontological ambiguity [186], subtly undermining the interpretive regularities on which this ecology relies. The perturbation operates primarily on semantic and predictive structure:

$$\delta\mathcal{E}(x; \mathcal{R}) \approx 0^-, \quad \delta\mathbf{A}(x; \mathcal{R}) \approx 0.$$

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  $\gamma_R$  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.

### 7.3 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  $\mathcal{R}$  acts as a field-level perturbation, not a trait-level driver.**

In topological terms, the robot introduces a deformation operator

$$\gamma_R : \mathcal{E} \rightarrow \mathcal{E}',$$

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  $\mathcal{R}$  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.  $\mathcal{R}$  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  $\mathcal{R}$  is present, participants implicitly alter their model of who counts as a moral interlocutor—a phenomenon well-documented in human–robot interaction literature. This re-categorisation 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  $\gamma_R$  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.

### 7.4 From the Failure of Machine Ethics to a Reconstruction of Computational Morality

The preceding analyses show that robotic co-presence  $\mathcal{R}$  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, 286]. These approaches presuppose that normative theories function as *implementable specifications*. However, as Floridi's Levels of Abstraction make clear [25, 307], 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 [259], and schema-based social cognition [88]. 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 [268, 204]—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 [92] and schema ecologies [88] 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  $\gamma_R$  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.

#### 7.4.1 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 [259]. Robotic presence is formalised as a perturbation operator:

$$\gamma_R : \mathcal{E} \rightarrow \mathcal{E}',$$

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.

#### 7.4.2 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.

## 7.5 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  $\gamma_R$  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,  $\gamma_R$  disrupts the recognition of accountability cues implicit in the Watching-Eye stimulus;
- consequentially, it flattens the perceived payoff gradient of beneficence;
- virtuously, it weakens the stabilising force of prosocial dispositions;
- sentimentally, it dampens empathic vector fields that anchor reactive moral emotions;
- contractually, it disrupts justificatory visibility between moral agents;
- particularistically, it shifts the situational salience profile.

These converging interpretations reveal the central structural insight of the thesis: *the robot does not add a new norm; it shifts the evaluative conditions under which norms become behaviourally operative*.

## 6. Final Position of the Thesis

We may now return to the guiding hypotheses:

**H1 — Evaluative Deformation** *Confirmed.* The evaluative process  $f$  linking perception to action is systematically altered by synthetic presence.

**H2 — Synthetic Normativity** *Confirmed.* Synthetic agents acquire derivative normative force by altering the field of salience and accountability.

**H3 — Synthetic Perturbation of Moral Inference** *Confirmed.* The robot refracts the transition from moral appraisal to prosocial behaviour, attenuating the expressive force of the Watching-Eye cue.

Accordingly, the thesis takes the following stand:

### Final Thesis Position (Definitive)

*Human moral agency is not internally autonomous. It is structurally coupled to the perceptual-social field in which it is embedded. Synthetic agents, even when lacking sentience, intentionality, or communicative acts, act as modulators of that field. They reshape attentional gradients, dampen empathic resonance, and deform the topological structures through which moral appraisal acquires behavioural expression. Moral displacement under synthetic presence is therefore not a behavioural curiosity, but a structural fact about the architecture of moral cognition.*

## 7. Implications for the Future of Computational Morality

This final insight reshapes the methodological landscape. Artificial agents cannot be treated as moral subjects but must be understood as **moral modifiers**: entities whose design implicitly reconfigures the evaluative field. Future research in computational morality must therefore move beyond rule encoding and value annotation toward a structural science of moral environments, moral salience, and field-sensitive interaction.

In this sense, the thesis does not simply present an experimental result; it offers a new conceptual foundation for the empirical and ethical study of artificial agents. It reorients the field toward a *topological, empirically grounded, and LoA-disciplined* understanding of moral cognition—one capable of addressing the forms of synthetic presence that will increasingly populate human social life.

*With this synthesis, the thesis closes. Its central claim is now complete: moral behaviour is field-dependent, and synthetic presence reshapes that field.*

## Bibliography

- [1] Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge University Press.
- [2] Anderson, J., Rainie, L., and Luchsinger, A. (2018). *Artificial Intelligence and the Future of Humans*. Pew Research Center.
- [3] Allcott, Hunt, Braghieri, Luca, Eichmeyer, Sarah, and Gentzkow, Matthew (2020). *The welfare effects of social media*. American Economic Review, 110(3), 629–76.
- [4] Auxier, Brooke, and Anderson, Monica (2021). *Social media use in 2021*. Pew Research Center.
- [5] Allen, Colin and Wallach, Wendell and Smit, Iva. (2006). *Why machine ethics?*, In: IEEE Intelligent Systems, 21(4), pp. 12–17. IEEE.
- [6] Allen, C., & Wallach, W. (2012). *Moral machines: contradiction in terms or abdication of human responsibility*. In *Robot ethics: The ethical and social implications of robotics* (pp. 55–68). MIT Press Cambridge. Mass.
- [7] Aristotle. (1984). *The Complete Works of Aristotle: The Revised Oxford Translation*. Princeton University Press.
- [8] Bail, Christopher A. (2021). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.
- [9] Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ, 1986.
- [10] Bryson, J. J. (2010). *Robots should be slaves*. In Close engagements with artificial companions: Key social, psychological, ethical and design issues (pp. 63-74). John Benjamins Publishing.
- [11] Bird, A. (2000). *Thomas Kuhn*. Princeton University Press.
- [12] Bricmont, J. (2016). *Making Sense of Quantum Mechanics*. Springer.
- [13] Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and individual differences*, 13(6), 653-665.
- [14] Chalmers, A. F. (2013). *What is this thing called science?* Hackett Publishing.
- [15] Laudan, L. (1987). Progress or Rationality? The Prospects for Normative Naturalism. *American Philosophical Quarterly*, 24(1), 19-31.
- [16] Woodward, J. (2007). *Making things happen: A theory of causal explanation*. Oxford university press.

- [17] Dennett, D. C. (1971). *Intentional systems*. The Journal of Philosophy, 68(4), 87-106.
- [18] Dwyer, Ryan J., El-Bardicy, Mostafa, and Hakami, Tahani (2020). *Seeking and avoiding digital distractions in the workplace*. Information Systems Journal, 30(5), 845-874.
- [19] Floridi, L. (2008). *Levels of Abstraction and the Foundation of Computational Ethics*. APA Newsletter on Philosophy and Computers, 8(1), 3-5.
- [20] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- [21] Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California law review*, 94(4), 945-967.
- [22] Hampton, K. N., Sessions, L. F., Her, E. J., and Rainie, L. (2009). *Social isolation and new technology*. Pew Internet and American Life Project.
- [23] International Federation of Robotics (IFR). (2019). *World Robotics Report*. IFR.
- [24] Mendelson, E. (2009). *Introduction to mathematical logic*. CRC Press.
- [25] Minsky, M. (1985). *The Society of Mind*. Simon and Schuster.
- [26] Moor, J. H. (2006). *The nature, importance, and difficulty of machine ethics*. IEEE intelligent systems, 21(4), 18-21.
- [27] Pantic, I. (2014). *Online social networking and mental health*, Cyberpsychology, Behavior, and Social Networking, volume 17, number 10, Mary Ann Liebert Inc 140 Huguenot Street 3rd Floor New Rochelle NY 10801 USA.
- [28] Pantic, Maja and Vinciarelli, Alessandro (2014), *Social signal processing*, The Oxford handbook of affective computing, page 84
- [29] Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. Y., Rosen, D., Colditz, J. B., Radovic, A., and Miller, E. (2017). *Social media use and perceived social isolation among young adults in the U.S.*, American Journal of Preventive Medicine, 53(1), 1-8. DOI: 10.1016/j.amepre.2017.01.010
- [30] Russell, B. (1919). *Introduction to Mathematical Philosophy*. London: George Allen & Unwin.
- [31] Russell, S., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited.
- [32] Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J.C., Lyon, T., Etchemendy, J. (2018). *The AI Index 2018 Annual Report*. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University.
- [33] Silver, D. et al. (2018). *A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play*. Science, 362(6419), 1140-1144.

- [34] Stone, P. et al. (2016). *Artificial Intelligence and Life in 2030*. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University.
- [35] Taylor, C. (1985). *Human Agency and Language: Philosophical Papers, Volume 1*. Cambridge University Press.
- [36] Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic books.
- [37] Zermelo, E. (1908). *Investigations in the foundations of set theory I*. In From Kant to Hilbert: A Source Book in the Foundations of Mathematics, Ewald, W. (ed.), Oxford University Press.
- [38] Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- [39] James, W. (1884). What is an Emotion?. *Mind*, 9(34), 188-205.
- [40] Misra, S., Cheng, L., Genevie, J., and Yuan, M. (2016). *The iPhone Effect: The Quality of In-Person Social Interactions in the Presence of Mobile Devices*. Environment and Behavior, 48(2), 275-298.
- [41] Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.
- [42] Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2), 207-232.
- [43] Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- [44] Vosoughi, Soroush, Roy, Deb, and Aral, Sinan (2018). *The spread of true and false news online*. Science, 359(6380), 1146-1151.
- [45] Haidt, Jonathan (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon
- [46] Xerxa, Yllza and Rescorla, Leslie A and Shanahan, Lilly and Tiemeier, Henning and Copeland, William E., (2023) *Childhood loneliness as a specific risk factor for adult psychiatric disorders*, Psychological Medicine, Volume 53 number 1, pages 227–235, Cambridge University Press.
- [47] Oda, R., Kato, Y., & Hiraishi, K. (2015). *The watching-eye effect on prosocial lying*. Evolutionary Psychology, 13(3), 1474704915594959. Los Angeles, CA: Sage Publications.
- [48] Atran, S. & Norenzayan, A. (2004). *Religion's Evolutionary Landscape: Counterintuition, Commitment, Compassion, Communion*. Behavioral and Brain Sciences, 27(6), 713-770.
- [49] Bering, J.M., McLeod, K., & Shackelford, T.K. (2005). *Reasoning about dead agents reveals possible adaptive trends*. Human Nature, 16(4), 360-381.
- [50] Shariff, A.F. & Norenzayan, A. (2007). *God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game*. Psychological Science, 18(9), 803-809. Los Angeles, CA: SAGE Publications.

- [51] Sharkey, A., & Sharkey, N. (2010). *The crying shame of robot nannies: an ethical appraisal*. Interaction Studies, 11(2), 161-190.
- [52] Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford: Oxford University Press.
- [53] Lin, P., Abney, K., & Bekey, G.A., eds. (2012). *Robot ethics: The ethical and social implications of robotics*. Cambridge, MA: MIT Press.
- [54] Bryson, J.J. (2010). *Robots should be slaves*. In Close engagements with artificial companions: Key social, psychological, ethical and design issues (pp. 63-74). Amsterdam: John Benjamins Publishing Company.

## Bibliography

- [1] 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] G. P. Goodwin and J. M. Darley, "The psychology of meta-ethics: Exploring objectivism," *Cognition*, vol. 106, no. 3, pp. 1339–1366, 2008.
- [86] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, *et al.*, "'economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies," *Behavioral and Brain Sciences*, vol. 28, no. 6, pp. 795–855, 2005.
- [87] J. Mikhail, "Universal moral grammar: Theory, evidence, and the future," *Trends in Cognitive Sciences*, vol. 11, no. 4, pp. 143–152, 2007.
- [88] 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.
- [89] D. Narvaez, "Triune ethics: The neurobiological roots of our multiple moralities," *New Ideas in Psychology*, vol. 26, no. 1, pp. 95–119, 2008.
- [90] L. Young and A. Waytz, "Moral cognition: A review," in *The Handbook of Social Psychology*, pp. 1–47, Oxford University Press, 5 ed., 2013.
- [91] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. Wojcik, *et al.*, "Moral foundations theory: The pragmatic validity of moral pluralism," *Advances in Experimental Social Psychology*, vol. 47, pp. 55–130, 2013.
- [92] J. M. Doris, *Lack of Character: Personality and Moral Behavior*. Cambridge University Press, 2002.
- [93] M. Black, "The factual and the normative," in *Human Science and the Problem of Values*.
- [94] J. Deigh, *An introduction to ethics*. Cambridge University Press, 2010.
- [95] R. M. Hare, *Moral Thinking: Its Levels, Method, and Point*. Oxford University Press, 1981.
- [96] 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.

- [97] 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.
- [98] F. Cushman and L. Young, “The psychology of dilemmas and the philosophy of morality,” *Ethics*, vol. 119, no. 4, pp. 597–636, 2009.
- [99] F. Cushman and J. D. Greene, “Finding faults: How moral evaluations arise from normative frameworks,” *Cognition*, vol. 136, no. 2, pp. 30–43, 2012.
- [100] 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.
- [101] J. D. Greene, “Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics,” *Ethics*, vol. 124, no. 4, pp. 695–726, 2014.
- [102] P. Railton, “Moral learning: Conceptual foundations and normative relevance,” *Cognition*, vol. 167, pp. 172–190, 2017.
- [103] M. Smith, *The Moral Problem*. Blackwell, 1994.
- [104] P. Railton, “Moral realism,” *The Philosophical Review*, vol. 95, no. 2, pp. 163–207, 1986.
- [105] S. Blackburn, *Ruling Passions*. Oxford University Press, 1998.
- [106] A. Gibbard, *Wise Choices, Apt Feelings*. Harvard University Press, 1990.
- [107] C. M. Korsgaard, *The Sources of Normativity*. Cambridge University Press, 1996.
- [108] 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.
- [109] 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.
- [110] 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.
- [111] 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.
- [112] J. LeDoux, *The Emotional Brain*. Simon and Schuster, 1998.
- [113] 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.

- [114] 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.
- [115] 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.
- [116] 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.
- [117] 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.
- [118] J. Mikhail, “Universal moral grammar: Theory, evidence, and the future,” *Trends in Cognitive Sciences*, vol. 11, no. 4, pp. 143–152, 2007.
- [119] 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.
- [120] F. Cushman and L. Young, “The psychology of dilemmas and the philosophy of morality,” *Ethics*, vol. 119, no. 4, pp. 597–636, 2009.
- [121] 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.
- [122] 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.
- [123] 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.
- [124] F. Van Overwalle, “Social cognition and the brain: A meta-analysis,” *Human Brain Mapping*, vol. 30, no. 3, pp. 829–858, 2009.
- [125] 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.
- [126] 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.
- [127] 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.
- [128] 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.

- [129] 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.
- [130] 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.
- [131] E. K. Miller and J. D. Cohen, “An integrative theory of prefrontal cortex function,” *Annual Review of Neuroscience*, vol. 24, pp. 167–202, 2001.
- [132] 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.
- [133] 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.
- [134] 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.
- [135] 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.
- [136] 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.
- [137] 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.
- [138] 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.
- [139] H. Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, 2 ed., 2014.
- [140] J. D. Greene, “The cognitive neuroscience of moral judgment and decision-making,” *Handbook of Neuroethics*, pp. 161–178, 2014.
- [141] 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.
- [142] 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.
- [143] 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.

- [144] 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.
- [145] M. Coeckelbergh, “Robot rights? towards a social-relational justification of moral consideration,” *Ethics and Information Technology*, vol. 12, no. 3, pp. 209–221, 2010.
- [146] M. Alfano, *Character as Moral Fiction*. Cambridge University Press, 2013.
- [147] 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.
- [148] 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.
- [149] 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.
- [150] P. Railton, “Moral learning: Conceptual foundations and normative relevance,” *Cognition*, vol. 167, pp. 172–190, 2017.
- [151] L. Floridi and J. W. Sanders, “On the morality of artificial agents,” *Minds and Machines*, vol. 14, no. 3, pp. 349–379, 2004.
- [152] M. Coeckelbergh, *AI Ethics*. MIT Press, 2020.
- [153] J. D. Greene, “Why are vmpfc patients more utilitarian? a dual-process theory of moral judgment,” *Annals of the New York Academy of Sciences*, vol. 1124, pp. 114–126, 2007.
- [154] J. S. B. T. Evans, “Dual-processing accounts of reasoning, judgment, and social cognition,” *Annual Review of Psychology*, vol. 59, pp. 255–278, 2008.
- [155] K. R. Scherer, “The dynamic architecture of emotion: Evidence for the component process model,” *Cognition and Emotion*, vol. 23, no. 7, pp. 1307–1351, 2009.
- [156] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.
- [157] P. Railton, “Moral learning: Conceptual foundations and normative relevance,” *Cognition*, vol. 167, pp. 172–190, 2017.
- [158] E. A. Crone and N. Steinbeis, “Neural perspectives on cognitive control development during childhood and adolescence,” *Trends in cognitive sciences*, vol. 21, no. 3, pp. 205–215, 2017.
- [159] C. D. Batson, *Altruism in Humans*. Oxford University Press, 2011.
- [160] E. Fehr and S. Gachter, “Altruistic punishment in humans,” *Nature*, vol. 415, pp. 137–140, 2002.

- [161] J. Henrich *et al.*, “Economic man in cross-cultural perspective,” *Behavioral and Brain Sciences*, vol. 28, no. 6, pp. 795–855, 2005.
- [162] F. Warneken, “Precocious prosociality: Why do young children help?,” *Child Development Perspectives*, vol. 9, no. 1, pp. 1–6, 2015.
- [163] 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.
- [164] S. Darwall, *The Second-Person Standpoint*. Harvard University Press, 2006.
- [165] 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.
- [166] M. Alfano, *Character as Moral Fiction*. Cambridge University Press, 2013.
- [167] M. Bratman, *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.
- [168] D. Velleman, *The Possibility of Practical Reason*. Oxford University Press, 2000.
- [169] P. F. Strawson, “Freedom and resentment,” *Proceedings of the British Academy*, vol. 48, pp. 1–25, 1962.
- [170] N. Arpaly, “Unprincipled virtues,” *Oxford University Press*, 2003.
- [171] P. Railton, “Moral learning: Conceptual foundations and normative relevance,” *Cognition*, vol. 167, pp. 172–190, 2017.
- [172] Aristotle, *Nicomachean Ethics*. Hackett, 1999.
- [173] J. McDowell, “Virtue and reason,” *The Monist*, vol. 62, no. 3, pp. 331–350, 1979.
- [174] M. Burnyeat, “Aristotle on learning to be good,” in *Essays on Aristotle’s Ethics* (A. Rorty, ed.), pp. 69–92, University of California Press, 1980.
- [175] I. Kant, *Groundwork of the Metaphysics of Morals*. Cambridge University Press, 1998.
- [176] H. E. Allison, *Kant’s Groundwork for the Metaphysics of Morals: A Commentary*. Oxford University Press, 2011.
- [177] D. Hume, *A Treatise of Human Nature*. Oxford University Press, 2000.
- [178] D. Hume, *An Enquiry Concerning the Principles of Morals*. Oxford University Press, 1998.
- [179] R. Cohen, *Hume’s Morality: Feeling and Fabrication*. Oxford University Press, 2008.
- [180] J. Greene and J. Haidt, “How (and where) does moral judgment work?,” *Trends in cognitive sciences*, vol. 6, no. 12, pp. 517–523, 2002.
- [181] M. Fedyk, *The Social Turn in Moral Psychology*. Cambridge, MA: MIT Press, 2017.

- [182] 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.
- [183] 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.
- [184] 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.
- [185] 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.
- [186] 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.
- [187] 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.
- [188] S. Baron-Cohen, “The extreme male brain theory of autism,” *Trends in cognitive sciences*, vol. 6, no. 6, pp. 248–254, 2002.
- [189] S. Baron-Cohen, “Autism and the empathizing-systemizing (es) theory,” *Developmental social cognitive neuroscience*, pp. 125–138, 2009.
- [190] 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.
- [191] A. Wakabayashi, S. Baron-Cohen, S. Wheelwright, N. Goldenfeld, J. De-laney, 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.
- [192] J. A. Bartz, J. Zaki, N. Bolger, E. Hollander, N. N. Ludwig, A. Kolevzon, and K. N. Ochsner, “Oxytocin selectively improves empathic accuracy,” *Psychological science*, vol. 21, no. 10, pp. 1426–1428, 2010.
- [193] 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.

- [194] 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.
- [195] 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.
- [196] 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.
- [197] R. R. McCrae and P. T. Costa, “The five-factor theory of personality,” *Handbook of Personality: Theory and Research*, pp. 159–181, 2008.
- [198] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, “The mini-ipip scales: Tiny-yet-effective measures of the big five factors of personality,” *Psychological Assessment*, vol. 18, no. 2, pp. 192–203, 2006.
- [199] 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.
- [200] 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.
- [201] 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.
- [202] 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.
- [203] 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.
- [204] 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.
- [205] A. Senju and G. Csibra, “Gaze following in human infants depends on communicative signals,” *Current Biology*, vol. 18, no. 9, pp. 668–671, 2008.
- [206] 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.
- [207] R. Rosenthal and R. L. Rosnow, *Essentials of Behavioral Research: Methods and Data Analysis*. New York: McGraw-Hill, 3 ed., 2008.

- [208] H. T. Reis and C. M. Judd, *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press, 2000.
- [209] A. E. Kazdin, *Research Design in Clinical Psychology*. Boston: Pearson, 5 ed., 2017.
- [210] 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.
- [211] 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.
- [212] 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.
- [213] 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.
- [214] Aldebaran Robotics, “Nao: Product overview and technical specifications,” tech. rep., Aldebaran Robotics, Paris, France, 2013. Official product documentation.
- [215] 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.
- [216] 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.
- [217] 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.
- [218] 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.
- [219] E. Husserl, *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy: First Book*. The Hague: Nijhoff, 1913. Original 1913; various translations available.
- [220] D. Zahavi, *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: MIT Press, 2005.
- [221] S. Gallagher, *How the Body Shapes the Mind*. Oxford: Oxford University Press, 2005.

- [222] 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.
- [223] F. Brentano, *Psychology from an Empirical Standpoint*. Routledge, 1874. Original work; various editions.
- [224] J. Searle, *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press, 1983.
- [225] T. Crane, *Elements of Mind: An Introduction to the Philosophy of Mind*. Oxford: Oxford University Press, 2001.
- [226] S. E. Guthrie, *Faces in the Clouds: A New Theory of Religion*. New York: Oxford University Press, 1993.
- [227] 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.
- [228] D. C. Dennett, *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.
- [229] L. Floridi, *Information: A Very Short Introduction*. Oxford: Oxford University Press, 2010.
- [230] L. Floridi, *The Ethics of Information*. Oxford: Oxford University Press, 2013.
- [231] 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.
- [232] J. K. Hietanen, “Social attention orienting induced by eye gaze and head orientation,” *Visual Cognition*, vol. 9, no. 1–2, pp. 1–22, 2002.
- [233] D. R. Carney, A. J. 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.
- [234] M. Argyle, *Bodily Communication*. London: Methuen, 1975.
- [235] G. Rhodes, “The evolutionary psychology of facial beauty,” *Annual Review of Psychology*, vol. 57, pp. 199–226, 2006.
- [236] G. Johansson, “Visual perception of biological motion and a model for its analysis,” *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.
- [237] 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.
- [238] 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.

- [239] R. E. Kleck and A. Strenta, “Perceptions of the gaze of another,” *Journal of Personality and Social Psychology*, vol. 39, no. 5, pp. 725–732, 1980.
- [240] J. K. Hietanen, “Does your gaze direction reflect your attention?,” *Visual Cognition*, vol. 6, no. 1, pp. 97–120, 1999.
- [241] 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.
- [242] 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.
- [243] 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.
- [244] C. Allen, W. Wallach, and I. Smit, “Why machine ethics?,” *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 12–17, 2006.
- [245] J. H. Moor, “The nature, importance, and difficulty of machine ethics,” *Machine ethics*, pp. 13–20, 2011.
- [246] J. H. Moor, “The nature and limits of machine ethics,” *AI and Society*, vol. 39, no. 1, pp. 33–51, 2023.
- [247] F. De Brigard, W. Sinnott-Armstrong, A. E. Monroe, N. Carroll, and J. May, “The agent?patient asymmetry in moral cognition: Evidence of a social bias in moral judgment,” *Cognitive Science*, vol. 45, no. 4, p. e12965, 2021.
- [248] R. Audi, *Moral Perception*. Princeton, NJ: Princeton University Press, 2015.
- [249] C. G. Hempel, “Aspects of scientific explanation,” 1965.
- [250] B. M. McLaren, “Computational models of ethical reasoning: Challenges, initial steps, and future directions,” *IEEE*, 2006.
- [251] L. Kohlberg, “Stage and sequence: The cognitive-developmental approach to socialization,” *Handbook of socialization theory and research*, vol. 347, p. 480, 1969.
- [252] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- [253] 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.
- [254] M. Anderson and S. L. Anderson, “Machine ethics: Creating an ethical intelligent agent,” in *AI Magazine*, vol. 28, pp. 15–26, AAAI Press, 2007.
- [255] J.-G. Ganascia, “Modelling ethical rules of warfare,” in *International Conference on Computer Ethics: Philosophical Enquiry (CEPE)*, pp. 181–190, 2007.

- [256] 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.
- [257] 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.
- [258] C. Thornton, “Rethinking machine ethics in the light of virtue ethics,” *Ethics and Information Technology*, vol. 15, no. 4, pp. 291–297, 2013.
- [259] P. S. Churchland, *Braintrust: What Neuroscience Tells Us About Morality*. Princeton, NJ: Princeton University Press, 2011.
- [260] J. Rawls, *A theory of justice*. Harvard university press, 2020.
- [261] J. S. Mill, *Utilitarianism*. Hackett Publishing, 1861.
- [262] 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.
- [263] J. McDowell, “Virtue and reason,” *The Monist*, vol. 62, no. 3, pp. 331–350, 1979.
- [264] J. Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Vintage, 2012.
- [265] A. Smith, *The Theory of Moral Sentiments*. Cambridge: Cambridge University Press, 1759. Edited by D. D. Raphael and A. L. Macfie (1976 edition).
- [266] S. Nichols, *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press, 2004.
- [267] 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.
- [268] 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.
- [269] 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.
- [270] 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.
- [271] 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.

- [272] 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.
- [273] J. A. Bargh and T. L. Chartrand, “The unbearable automaticity of being.,” *American psychologist*, vol. 54, no. 7, p. 462, 1999.
- [274] D. Ross and W. D. Ross, *The right and the good*. Oxford University Press, 2002.
- [275] J. Griffin, *Well-Being*. Oxford: Oxford University Press, 1986.
- [276] M. Stocker, *Plural and Conflicting Values*. Oxford: Oxford University Press, 1990.
- [277] P. Foot, “The problem of abortion and the doctrine of double effect’, in her virtues and vices,” *Berkeley and Los Angeles: University of California Press. FootThe Problem of Abortion and the Doctrine of the Double Effect19Virtues and Vices1978*, pp. 19–32, 1978.
- [278] R. J. Wallace, “Practical Reason,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, spring 2020 ed., 2020.
- [279] H. S. Richardson, “Moral Reasoning,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, fall 2018 ed., 2018.
- [280] P. Lin, G. Bekey, and K. Abney, “Autonomous military robotics: Risk, ethics, and design,” tech. rep., California Polytechnic State Univ San Luis Obispo, 2008.
- [281] K. Atkinson and T. Bench-Capon, “Action-based alternating transition systems for arguments about action,” in *AAAI*, vol. 7, pp. 24–29, 2007.
- [282] 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.
- [283] M. Hjelmbom, *Deontic action-logic multi-agent systems in Prolog*. Högskolan i Gävle, 2008.
- [284] 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.
- [285] 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.
- [286] 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.
- [287] A. R. Honarvar and N. Ghasem-Aghaei, “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.
- [288] 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.
  - [289] 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.
  - [290] R. Montague and R. H. Thomason, “Formal philosophy. selected papers of richard montague,” *Erkenntnis*, vol. 9, no. 2, 1975.
  - [291] R. Carnap, *Introduction to symbolic logic and its applications*. Courier Corporation, 2012.
  - [292] L. M. Pereira and A. Saptawijaya, “Modeling morality with prospective logic,” *Cambridge University Press*, 2007.
  - [293] “The problem of machine ethics in artificial intelligence,” *AI and SOCIETY*, vol. 35, no. 1, pp. 103–111, 2020.
  - [294] J. McDermid, V. C. Muller, T. Pipe, Z. Porter, and A. Winfield, “Ethical issues for robotics and autonomous systems,” 2019.
  - [295] D. Howard and I. Muntean, “Artificial moral cognition: moral functionalism and autonomous moral agency,” in *Philosophy and computing*, pp. 121–159, Springer, 2017.
  - [296] M. Pantic and A. Vinciarelli, “Social signal processing,” *The Oxford handbook of affective computing*, p. 84, 2014.
  - [297] R. W. Picard, *Affective computing*. MIT press, 2000.
  - [298] R. A. Calvo, S. D’Mello, J. M. Gratch, and A. Kappas, *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.
  - [299] S. Baron-Cohen, *The Essential Difference: The Truth about the Male and Female Brain*. London: Penguin, 2003.
  - [300] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Cambridge, MA: Perseus Books, 1994.
  - [301] R. D. Beer, “A dynamical systems perspective on agent–environment interaction,” *Artificial Intelligence*, vol. 72, no. 1–2, pp. 173–215, 1995.
  - [302] L. B. Smith and E. Thelen, “Development as a dynamic system,” *Trends in Cognitive Sciences*, vol. 7, no. 8, pp. 343–348, 2003.
  - [303] K. Friston, “The free-energy principle: a unified brain theory?,” *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.
  - [304] 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.

- [305] 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.
- [306] T. Hollenstein, *State Space Grids: Depicting Dynamics Across Development*. New York: Springer, 2015.
- [307] L. Floridi and J. W. Sanders, “On the morality of artificial agents,” *Minds and machines*, vol. 14, no. 3, pp. 349–379, 2004.