Synthetic Presence and the Topology of Moral Evaluation:

How Humanoid Robots Modulate Human Prosocial Action in Experimental Settings.

Francesco Perrone

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

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

November 2025

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

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

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

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

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

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

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

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

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

Francesco Perrone

Al mio compagno d'avventure, Francesco.

Abstract

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

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

Interpreted within the mapping $f(\alpha_E, \beta_C, \gamma_R)$ and read at the operative Level of Abstraction, the behavioural pattern is consistent with a structural account in which synthetic presence acts as a perturbation on the evaluative field. NAO's embodied ambiguity appears to modulate intuitive appraisal by shifting attentional weighting and weakening the affective resonance ordinarily elicited by morally salient cues, thereby redistributing salience prior to explicit deliberation. The robot does not contribute moral content; rather, it modifies the perceptual and affective scaffolds through which such content becomes action-guiding.

The thesis proposes that artificial systems, even when devoid of agency or intent, can operate as modifiers of the moral environments in which human decisions unfold. The behavioural patterns observed here are consistent with the view that moral action is field-dependent, topologically structured, and responsive to synthetic co-presence. This perspective highlights a limitation of strictly top-down approaches in Machine Ethics and motivates an ecological orientation—one that emphasises the governance of the evaluative environments shaped through human–machine interaction. The work offers a methodological basis for a computational and topological investigation of moral cognition under conditions increasingly defined by artificial presence.

Synopsis

This thesis investigates whether the mere presence of a humanoid robot can alter the cognitive–affective processes through which morally salient cues are transformed into moral action. Drawing on intuitionist models of moral cognition, it reframes moral judgement as a perceptual and affective process structured by salience, attention, and dispositional architecture rather than explicit reasoning. The theoretical framework is formalised through the evaluative mapping

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

which models how environmental cues, latent psychological structure, and synthetic presence jointly influence behavioural outcomes.

Three hypotheses guide the empirical work: *evaluative deformation*, *synthetic normativity*, and *synthetic perturbation of moral inference*. A controlled experiment examines prosocial donation under a Watching–Eye stimulus, contrasting a human-alone condition with silent robotic co-presence. Across inferential contrasts, cluster-specific analyses, and Bayesian estimation, the results reveal a modest but structured attenuation of prosocial behaviour, concentrated within a Prosocial–Empathic dispositional ecology. The robot does not introduce new norms; instead, its perceptual and ontological ambiguity modulates the salience and affective weight of existing moral cues.

The Discussion situates these findings within Human–Robot Interaction, affective computing, and Floridi's Levels of Abstraction. NAO is interpreted not as a moral agent but as a *morally relevant informational object* whose presence reorganises the evaluative field within which moral cognition unfolds. The thesis therefore advances an ecological reconceptualisation of Machine Ethics: artificial systems exert morally significant influence not through reasoning or agency, but by reshaping the perceptual and affective scaffolds through which humans register and enact moral meaning. This field-theoretic perspective unifies the empirical, computational, and philosophical contributions, showing that synthetic presence already plays a substantive role in the moral ecology of technologically saturated environments.

Contents

Abstract **iii**

Synopsis **iv**

Acknowledgments **xv**

Declaration **xvi**

1 Introduction **1**
- 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

2 Literature Review: Existing Approaches and the Level-of-Abstraction Problem **13**
- 2.1 Introduction: Scope, Objectives, and Theoretical Commitments . 13
- 2.2 Two Levels of Abstraction in Machine Ethics . . . 19
- 2.3 Clarifying the Explanatory Level of the Present Work . . . 21
- 2.4 The Cognitive–Affective Foundations of Moral Judgment . . . 22
- 2.5 Levels of Abstraction and the Structure of Machine Ethics . . . 24
- 2.6 Evaluative Topology, Affective Architecture, and Synthetic Moral Perturbation . . . 25
  - 2.6.1 The Evaluative Field . . . 26
  - 2.6.2 Moral Behaviour as Movement Within an Evaluative Field 27
  - 2.6.3 Synthetic Presence as Evaluative Modulator . . . 28
- 2.7 A Topological Lens on Moral Modulation . . . 28
  - 2.7.1 Toward a Unified Framework . . . 29
- 2.8 Integrative Synthesis: Toward a Cognitive–Affective Framework for Machine-Mediated Morality . . . 30
- 2.9 Global Synthesis: From Inferential Displacement to Synthetic Moral Topology . . . 30
  - 2.9.1 From Question to Framework . . . 30
  - 2.9.2 Rationale for a Multi-Hypothesis Approach . . . 31
  - 2.9.3 What the Literature Establishes . . . 31

3 Cognitive Affective Architecture of Moral Judgement **33**
- 3.1 Descriptive and Normative Domains . . . 35
  - 3.1.1 Why Definitions Vary . . . 37
  - 3.1.2 Minimal Operational Definition for This Thesis . . . 38
- 3.2 Judgments: Factual and Normative . . . 40
- 3.3 Internal Architecture of Moral Judgment . . . 42

3.3.1 Psychological and Neuroscientific Foundations of Moral Decision-Making . . . 44
3.3.2 Functional Integration and Practical Orientation. 46
3.4 From Moral Architecture to Perturbation by Synthetic Agents . . 46
3.4.1 Philosophical Synthesis . . . 49
3.5 Concluding Perspective: Why This Matters for the Thesis 49

4 From Theory to Measurement: Operationalising Dispositions and Moral Perturbation 51
4.1 Perturbation as Measurement: The Experimental Context 53
4.1.1 Purpose and Structure of this Chapter . . . 53
4.2 The Role of Psychometric Tools in the Evaluative–Topological Architecture . . . 55
4.3 Why These Tools: Methodological Criteria and Alignment with the Thesis . . . 57
4.4 The Empathizing Quotient (EQ): Affective Resonance as Evaluative Curvature . . . 58
4.4.1 EQ and Synthetic Presence . . . 59
4.4.2 Methodological Role in the Thesis . . . 60
4.5 The Systemizing Quotient (SQ): Structural Precision in the Evaluative Field . . . 60
4.5.1 Theoretical Background and Psychometric Foundations . . 60
4.5.2 SQ Across Moral Psychology and HRI . . . 61
4.5.3 Systemizing Quotient (SQ): Structural Bias and Evaluative Rigidity . . . 61
4.5.4 SQ, Synthetic Presence, and Field-Level Perturbation . . . 62
4.5.5 Methodological Significance . . . 62
4.6 The Big Five Inventory (BFI): Personality Geometry Within the Evaluative Topology . . . 62
4.6.1 Why Personality Matters for a Topological Account of Moral Cognition . . . 63
4.6.2 Psychometric Stability and Cross-Domain Predictive Value 63
4.6.3 Personality and Moral Behaviour Under Social Presence . 63
4.6.4 Personality Geometry in the Evaluative–Topological Model 64
4.6.5 Cluster Analysis: Mapping the Dispositional Manifold . . . 64
4.6.6 The Key Empirical Result: A Uniform Field-Level Displacement . . . 64
4.6.7 Methodological Significance . . . 65
4.7 The Watching–Eye Paradigm: Amplifying Moral Salience and Revealing Field-Level Deformation . . . 65
4.7.1 Watching–Eye Cues as Topological Amplifiers . . . 66
4.7.2 Why Child-Pair Eyes Provide a Clean Baseline . . . 66
4.7.3 Why Synthetic Presence Dilutes or Distorts the Effect . . . 67
4.7.4 Empirical Finding: Uniform Attenuation of the Watching–Eye Effect . . . 67
4.7.5 Why the Watching–Eye Paradigm Is Indispensable . . . 68
4.7.6 Integration With Costly Prosocial Action . . . 68
4.7.7 Synthesis: A Window Into Moral Topology . . . 68

4.8 General Conclusion: Measurement as the Logic of Synthetic Moral Perturbation . . . 69
4.8.1 Dispositional Mapping: A Structured Manifold, Not a Confound . . . 69
4.8.2 Watching–Eye Cues as Diagnostic Amplifiers . . . 69
4.8.3 Philosophical and Ethical Meaning . . . 70
4.8.4 Methodological Synthesis: The Tools as Epistemic Infrastructure . . . 70
4.9 Transition to Experimental Methods . . . 70

5 Operationalising Evaluative Topology: An Experimental Framework for Moral Perturbation 72
5.1 The Experimental Question as a Test of Field-Level Perturbation 73
5.1.1 Operationalising Moral Action: Prosocial Donation as Behavioural Endpoint . . . 74
5.1.2 Implementing γ_R: The Rationale for Humanoid Synthetic Presence . . . 75
5.1.3 Structuring the Test of Evaluative Perturbation . . . 76
5.2 Experimental Design and Behavioural Paradigm . . . 76
5.2.1 From Architecture to Procedure . . . 77
5.2.2 Experimental Manipulation: Presence as the Only Ontological Difference . . . 77
5.2.3 Participants . . . 79
5.2.4 Ontological Ambiguity as a Perturbation of Evaluative Processing . . . 79
5.2.5 Levels of Abstraction: Why the Robot Can Matter Without Doing Anything . . . 80
5.2.6 Behavioural Paradigm: Donation as Moral Action . . . 81
5.2.7 Preliminary Findings . . . 81
5.2.8 From Behavioural Setup to Evaluative Structure . . . 82
5.3 Synthetic Perturbation of Moral Inference . . . 83
5.4 Inferential Analysis of Experimental Data . . . 85
5.4.1 Demographic Equivalence as a Symmetry Condition . . . 85
5.4.2 Data Preparation and Preprocessing Workflow . . . 86
5.4.3 Preliminary Descriptive Patterns: Orientation Prior to Inferential Analysis . . . 88
5.4.4 Inferential Comparison of Donation Patterns Across Conditions . . . 90
5.4.5 What the Aggregate Divergence Establishes . . . 92
5.4.6 Quantification of Behavioural Modulation: Parametric and Nonparametric Effect Sizes . . . 93
5.5 Dispositional Baseline: Big Five Personality Traits Across Conditions 97
5.5.1 Between-Condition Comparisons of Big Five Personality Traits . . . 97
5.5.2 Predictive and Moderating Roles of Big Five Personality Traits . . . 98
5.5.3 Transition to Structural Modelling of Dispositional Architecture . . . 100

5.5.4 Latent Dispositional Structures and the Modulation of Moral Perturbation . . . 101
5.5.5 Psychometric Interpretation and Semantic Labelling of Latent Personality Clusters . . . 104
5.5.6 Cluster-Specific Regression Analysis of Condition Effects . 107
5.5.7 Bayesian Estimation and the Representation of Epistemic Gradients . . . 110
5.5.8 Closing Reflection: How Synthetic Presence Reconfigures the Moral Field . . . 114

6 Disscussion 116
6.1 Reframing the Central Question . . . 116
6.2 What Can Be Inferred from This Experiment . . . 116
6.3 The Broader Significance: Moral Cognition as a Topological Process 117
6.3.1 Revisiting the Findings Through the Cognitive Architecture of Moral Intuition . . . 118
6.4 Synthetic Presence and the Topology of Moral Salience . . . 120
6.4.1 Rethinking Machine Ethics Through Moral Topology . . . 122
6.5 General Synthesis: Moral Topology, Synthetic Presence, and the Architecture of Human–Machine Moral Ecosystems . . . 124
6.6 Final Synthesis: Moral Topology, Synthetic Presence, and the Boundaries of Interpretation . . . 127

7 Ethical Theory in a Cognitive–Topological Framework 130
7.1 From Moral Cognition to Ethical Theory . . . 130
Bridging Note: From Moral Cognition to Ethical Theory . . . 130
7.2 Introduction: Why Ethics Needs Psychology (and Why Computing Science Needs Both) . . . 131
7.3 Ethical Theory as Second-Order Analysis . . . 133
7.3.1 Ethical Reflection and the Second-Order Stance . . . 133
7.3.2 Levels of Abstraction and the Proper Location of Ethical Explanation . . . 133
7.3.3 Evaluative Topology as a Bridge Between Orders . . . 135
7.4 The Normative Landscape: Structuring Ethical Theories Through LoA and Topology . . . 138
7.4.1 The Three Dimensions of Normative Analysis . . . 138
7.4.2 Why This Framework Matters for the Experimental Chapter 139
7.5 Deontological Structures: The Architecture of Practical Reason . 140
7.5.1 The Source of Normativity: Rational Agency and the Form of Law . . . 140
7.5.2 Deontic Invariants as Topological Constraints . . . 141
7.5.3 Relevance to Synthetic Perturbation . . . 141
7.5.4 Mode of Evaluation: Maxims, Duties, and the Structure of Permissibility . . . 142
7.5.5 Action-Guidance: How Normative Constraints Influence Behaviour . . . 142
7.5.6 Deontological Normativity as Topological Invariance . . . 143
7.5.7 Why Deontology Matters for the Experimental Logic . . . 143

7.5.8 Conceptual Note: Gradient Fields in Consequentialist Topology . . . 144
7.6 Consequentialist Structures: Value Gradients and the Topology of Outcomes . . . 146
7.6.1 The Source of Normativity: Welfare, Impartiality, and the Structure of Reasons . . . 146
7.6.2 Mode of Evaluation: Consequences, Expected Value, and Scalar Normativity . . . 147
7.6.3 Action-Guidance Mechanism: From Value Gradients to Behavioural Pressure . . . 147
7.6.4 Consequentialist Topology: Moral Action as Gradient Following . . . 148
7.6.5 Why Consequentialism Matters for the Experimental Logic 148
7.7 Virtue-Theoretic Structures: Dispositions, Character Topology, and Moral Sensitivity . . . 149
7.7.1 The Source of Normativity: Character, Practical Wisdom, and Moral Perception . . . 149
7.7.2 Mode of Evaluation: Dispositions as Topological Structure 150
7.7.3 Action-Guidance Mechanism: Habituation, Stability, and Situated Sensitivity . . . 151
7.7.4 Virtue-Theoretic Topology: Stability, Curvature, and Susceptibility to Perturbation . . . 151
7.7.5 Why Virtue Ethics Matters for the Experimental Logic . . 152
7.8 Integrated Ethical Interpretation of the Experimental Results . . 153
7.9 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology . . . 155
7.10 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology . . . 156
7.10.1 The Source of Normativity: Sentiment as the Basis of Moral Appraisal . . . 156
7.10.2 Mode of Evaluation: Affective Resonance as Moral Metric 156
7.10.3 Action Guidance: Affective Vector Fields and Behavioural Dynamics . . . 157
7.10.4 Machine Ethics and the Blind Spot of Affective Architecture 157
7.10.5 Experimental Realisation: Synthetic Dampening of Empathic Resonance . . . 158
7.11 Contractualism, Particularism, and Hybrid Normative Models . . 158
7.11.1 Contractualism: Moral Claims as Justification-Equilibria . 159
7.11.2 Moral Particularism: Contextual Salience and the Fragmented Topology of Reasons . . . 160
7.11.3 Hybrid and Pluralist Models: Multidimensional Evaluative Topologies . . . 161
7.11.4 Integrative Ethical Interpretation of the Experimental Findings . . . 162

8 General Synthesis 164
8.1 Introduction: Why the Experiment Requires a Structural Interpretation . . . 164

8.1.1 From Behaviour to Structure: Why a Higher-Level Interpretation is Required . . . 166
8.1.2 Why This Chapter Cannot Be Pure "Discussion" in the Conventional Sense . . . 168
8.1.3 A Structural Reading of the Core Experimental Result . . 169
8.1.4 Why the Synthetic Presence Effect Matters Beyond the Experiment . . . 170
8.2 Cluster-by-Cluster Integrative Interpretation . . . 171
8.3 Global Normative–Topological Synthesis . . . 173
8.4 From the Failure of Machine Ethics to a Reconstruction of Computational Morality . . . 176
8.4.1 Reconstructing Computational Morality: An Empirically Grounded Paradigm . . . 177
8.4.2 Computational Morality as a Scientific Research Programme 178
8.5 Thesis-Wide Synthesis and Closing Reflections . . . 178

9 Conclusion 182
9.1 Returning to the Question: What This Thesis Has Shown . . . 182
9.2 Contributions to Human–Robot Interaction, Affective Computing, and Moral Cognition . . . 183
9.2.1 Contribution to Human–Robot Interaction . . . 183
9.2.2 Synthetic Presence and Floridi's Account of Moral Agency 184
9.2.3 Contribution to Affective Computing and Social Signal Processing . . . 185
9.2.4 Contribution to Moral Cognition Research . . . 186
9.2.5 Integrative Contribution: A Unified Field-Theoretic Approach . . . 186
9.2.6 A Unified Explanatory Structure Rather Than Three Independent Literatures . . . 187
9.3 Final Synthesis and Closing Reflections . . . 188

List of Tables

5.1 Demographic balance tests across experimental conditions. Raw test statistics and uncorrected p-values are reported. Significance was evaluated against an FDR-adjusted threshold of $\alpha = 0.05$. No test reached significance, supporting the assumption of baseline equivalence between the *Robot* and *Control* groups. 86

5.2 Descriptive summaries of behavioural and psychometric variables across the *Control* and *Robot* conditions. The table provides an orienting overview of the sample; its values are descriptive only and do not imply group differences, effects, or perturbation. . . . 89

5.3 Inferential comparisons of donation behaviour across conditions. The first row reports a total-level deviation test using a χ^2-style statistic applied to donation sums (not categorical frequencies). The Mann–Whitney U test and bootstrapped mean difference assess distributional structure and central tendency, respectively. . . 91

5.4 Inferential comparisons of donation behaviour across conditions. The chi-squared test (applied to total coin frequencies), the Mann–Whitney U test, and the bootstrapped mean difference collectively characterise the behavioural contrast. 96

List of Figures

5.1 Top–down view of the experimental and control configurations. Both layouts are spatially and visually identical; the humanoid robot is the only ontological difference between conditions. Within the evaluative–topological framework developed in this thesis, this equivalence is essential: it ensures that all morally salient environmental features are held constant so that any difference in prosocial behaviour can be attributed to the activation of the perturbation operator γ_R. The charity donation box (green box), together with the payment instructions placed beside it, constitutes the locus at which the measurable moral action δ_m occurs. The morally salient cue α_E includes both the room's stable layout and, crucially, the child-in-need poster positioned directly in front of the participant during the task; this poster is omitted from the schematic only to preserve diagrammatic clarity. The personality questionnaires on the desk instantiate the elicitation of the dispositional manifold β_C, whose structure is measured prior to observing behaviour. The presence or absence of the NAO robot corresponds to the perturbation operator γ_R: in the experimental condition, its minimal micro-movements provide a perceptually social but behaviourally neutral source of synthetic salience, whereas in the control condition $\gamma_R = 0$. Formally, the two panels depict identical instantiations of α_E and β_C, differing only in whether γ_R is applied, thereby isolating synthetic presence as the sole topological modification to the evaluative field. 78

5.2 Age distribution across experimental conditions. The histogram visualises the demographic composition of participants in the Control and Robot conditions. Bars represent grouped counts by age bracket. The plot provides descriptive confirmation that the age structure of the sample is comparable across conditions; it does not carry inferential force. 87

5.3 Donation amounts by experimental condition. The figure displays central tendency, dispersion, and outlier structure for the behavioural outcome δ_m. It provides the descriptive substrate over which the mapping $\alpha_E \mapsto \mathscr{P}(\delta_m)$ will later be examined for deformation under the perturbation operator γ_R. The figure is descriptive only and does not support inferential claims. 88

5.4 Mean donation amounts by experimental condition with 95% bootstrapped confidence intervals. The overlap between intervals illustrates substantial individual variability, indicating that any perturbative influence of $\mathscr{R}$ is diffuse rather than deterministic. . . . 91

5.5 Kernel density estimates of donation distributions across experimental conditions. The Control group exhibits greater mass at higher donation values, whereas the Robot group shows a mild left-shift in density. These plots provide distributional context for the effect-size metrics discussed in the text. 95
5.6 Mean donation amounts with standard error bars by condition. While the Control group donates more on average, the overlapping error bars reflect substantial individual-level variability. The figure complements the density plot by highlighting differences in central tendency rather than distributional shape. 95
5.7 Kernel density estimates for each Big Five trait across experimental conditions. The plots depict the distribution of trait scores for the *Control* (orange) and *Robot* (green) groups. All five dimensions show substantial overlap, visually corroborating the non-significant differences found in corresponding Mann–Whitney U tests 98
5.8 Scatter plots of donation amounts against each of the Big Five personality traits, with monotonic regression lines. No predictive relationships are apparent, and no consistent moderation patterns emerge across traits. These visual results support the null findings from the Mann–Whitney and interaction analyses. 99
5.9 Participants clustered in PCA-reduced psychometric space. Three clusters emerge as coherent and visually distinguishable groupings, providing the structural substrate for subsequent analyses of condition-by-cluster effects. 102
5.10 Elbow plot (left axis) and silhouette coefficients (right axis) across candidate values of k. The elbow at $k = 3$ and stable silhouette profile support selecting three clusters as an interpretable and parsimonious solution. 103
5.11 Mean donation amount by condition within each personality cluster. Error bars represent standard deviation. Cluster 1 shows a clearer attenuation of donation under robotic presence, while Clusters 0 and 2 display only modest or negligible differences. 103
5.12 Radar profiles (normalised for comparability) of the three latent dispositional ecologies. Left: Cluster 0 (Emotionally Reactive / Low-Structure); Centre: Cluster 1 (Prosocial–Empathic / Warm–Sociable); Right: Cluster 2 (Analytical–Structured / High-Systemizing). These plots visualise the relative psychometric configuration of each ecology. 105
5.13 Regression coefficients (with 95% confidence intervals) for the Robot condition estimated separately within each latent personality cluster. Cluster 1 shows a larger negative coefficient relative to the other clusters, though uncertainty remains high due to small within-cluster sample sizes. Clusters 0 and 2 exhibit coefficients near zero. These estimates provide local directional contrasts prior to interaction and Bayesian modelling. 109

5.14 Posterior distribution of the modelled donation difference between conditions. The density is skewed toward positive values (greater expected donations in the Control condition), providing directional probabilistic evidence for attenuation under robotic co-presence. The dashed line marks the point of no effect. 112

Acknowledgments

There is a peculiar stillness that settles around work completed under the accelerating discipline of contemporary academia—a sense that one has been guided less by patient inquiry than by the unhurried rhythm in which ideas ordinarily choose to unfold, and more by the unyielding cadence of an institution convinced that thought must keep pace with its deadlines. If these pages appear composed at a certain distance from themselves, it is because they bear the traces of that tension: the quiet struggle between what might have matured freely and what the present era insists must be shaped, concluded, and surrendered.

In that unsettled interval I have leaned on those whose presence never depended on the coherence of my arguments. This thesis is dedicated to my son, Francesco, who—like an improbable guardian with a light too bright for the corridors I found myself in—used the simple force of his existence to hold back a darkness I could not have held alone. To my mother, Mirella, whose illness and narrowing sight should have been met with my unhurried company rather than the prolonged absences this work imposed; to my father, Alberto, whose quiet vigilance, wisdom, and unwavering care have sustained us in ways that rarely leave a trace on any page; and to my wife, Anna, who has carried more than anyone her age should be asked to bear—not only for reasons that cannot be written here, but because she has returned, again and again, to the limits of my own intellect as though they were a place of rest rather than constraint. To Dr. Herrera Martín (Tony), whose companionship belongs not to the category of colleagues but to that small, accidental inner circle one recognises rather than chooses: in the most unsettled stretches, he reminded me to read, to play games, and to look at myself with a measure of worth I do not naturally grant my own being. And to Ettore Majorana—not the historical figure, but the strange, inward nearness of his temperament and vicissitudes—whose quiet defiance of the world's demands has long served as a reminder that thought can survive even where institutions falter.

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

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

Declaration

I confirm that I am the sole author of this thesis. With the exception of background sections that review existing literature and established theoretical frameworks (chapters 2, 4, 7 and the non-original contextual material in chapters 3 and 8), all research reported here is my own independent work.

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

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

1. Introduction

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

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

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

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

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

Assertions that an algorithm "acts ethically" typically reflect confidence in the coherence of a formal model rather than an examination of how moral behaviour actually arises in human agents. Such models can be elegant and internally consistent, yet their structure often mirrors the assumptions required for formal tractability rather than the dynamics that govern intuitive moral cognition. The difficulty is not that these approaches are misguided, but that they describe a different kind of object: a rule-based construction designed to yield determinate outputs, rather than the perceptual, affective, and context-sensitive processes through which moral considerations gain relevance in human action. Recognising

this distinction clarifies why achievements at the level of formal specification do not straightforwardly illuminate the mechanisms by which people navigate morally charged situations.

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 *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, and—more quietly—in our phones, woven into the background of daily life.

And it is here that we must pause, abruptly, to ask what exactly is being altered.

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.*

This move– from a lived encounter to an analytic definition– is deliberate. We cannot understand how artificial agents affect our moral thinking unless we consider both the concrete situations in which people act, and the underlying processes that give their moral behaviour its structure. With this framework, we designed a study in which participants enter a small room and face a simple moral choice. They may give part of their participation payment to a real charity, or keep the full amount for themselves. This task is not intended to represent moral cognition in all its complexity. Its aim is different. By providing a minimal and controlled setting, it allows some of the elements in our definition of moral decision-making to become empirically observable.

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, or to the coins that could be kept?

Moral decision-making is, at its core, a fundamentally *teleological* process. It unfolds toward action: its purpose is to organise the evaluative conditions under which an agent adopts one course rather than another, consistent with classical action-centred accounts of ethics [13, 14, 15]. When we describe moral decision-making in terms of the perceptual, affective, mnemonic, and attentional operations outlined above, we are recognising the teleological structure through which these elements converge on a practical commitment. This view also aligns with empirical models linking appraisal to action selection [16, 10, 17].

The transition from moral judgment to behaviour is not an optional addendum to the process—it is its natural terminus. A moral evaluation that does not shape the field of possible actions has not yet completed its function; a moral action, conversely, is the crystallised endpoint of evaluative dynamics that have been unfolding long before reflection makes them explicit [18, 19, 16]. The participant's eventual choice to donate or not is the behavioural crystallisation of this entire evaluative process.

This thesis examines how the silent co-presence of a humanoid robot modulates that transformation. The robot does not request, instruct, or communicate, yet its ambiguous social ontology—perceptually agentic, normatively indeterminate—reshapes the conditions under which moral judgments are formed and resolved. In this way, the experiment offers a precise instantiation of the definition of moral decision-making introduced above: a setting in which perceptual cues, affective resonance, attentional dynamics, and implicit social meaning combine

to produce a practical moral commitment, and in which that process can be systematically perturbed.

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

The introduction of synthetic agents into this environment raises a conceptual and empirical challenge. Humanoid robots occupy a liminal ontological space: perceptually social yet not persons, agent-shaped yet not agents. Their presence recruits perceptual and affective systems that evolved for human–human interaction, while simultaneously withholding the ordinary resources through which social meaning stabilises.

This thesis examines the possibility that *such entities reshape the evaluative conditions of moral cognition not by acting, but simply by being present.*

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

Now *place a NAO robot on the same table.* It does nothing. It does not speak, gesture, or request. Yet its humanoid shape, its forward posture, its apparent capacity for attention, reframes the scene. The participant hesitates: the social field has changed. Something in the evaluative machinery has shifted—an attenuation of empathic pull, a dilution of accountability, a re-weighting of salience. We started by looking at something very simple: what happens when a humanoid robot is present in the room while someone is making a moral decision. The robot doesn't talk, it doesn't give instructions, it doesn't ask for anything. It just shares the space—quietly, almost like another person waiting their turn. But that quiet presence turns out to matter. A robot like that sits in an odd position: it looks and moves in ways that make us treat it as an agent, yet we don't quite know what kind of 'being' it is or what norms apply to it. That ambiguity changes the atmosphere. It shifts how people interpret the situation, what they take to be appropriate, and how comfortable they feel committing to one judgment over another. So even without speaking, the robot reshapes the background against which moral choices are made. It nudges the whole process—not by argument or instruction, but simply by being there, hovering between the familiar category of a person and the familiar category of a machine. That's where we see the transformation beginning. This modest behavioural moment is the phenomenon under investigation.What has changed? And why?

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

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

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

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

1.1 From Research Question to Hypotheses: Framing the Investigative Architecture

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

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

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

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

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

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

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

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

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

Hypothesis 1: Evaluative Deformation

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

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

If the robot is visually or socially salient—even without speaking—it might 'bend' the landscape you're navigating. Think of it like a small gravitational field: it

doesn't tell you what to do, but it changes the shape of the space you're moving through. This hypothesis asks:

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

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

Hypothesis 2: Synthetic Normativity of Moral Displacement

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

People may unconsciously treat a robot 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 5 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 converge on a structural point: the research question in this thesis concerns the same broad domain that motivates Machine Ethics—how humans and artificial systems interact in morally relevant contexts—but approaches it from a different explanatory level. Traditional Machine Ethics begins with articulated normative frameworks—rules, utilities, virtues—and develops methods for ensuring that artificial systems behave in ways consistent with those frameworks. This is a coherent and valuable aim when the goal is to design systems whose actions can be aligned with explicit ethical models.

The present work, however, investigates a phenomenon that arises upstream of such normative articulation. Our question is whether the mere presence of a synthetic agent alters the process through which humans move from perceptual encounter to moral action. That transition is shaped by intuitive appraisal, affective resonance, and attentional dynamics long before principles or reasons are invoked. In this sense, Machine Ethics and the approach developed here examine related human–machine moral interactions, but at different explanatory levels: one focuses on the normative specification of artificial behaviour, while the other examines how artificial presence perturbs the evaluative processes through which human moral behaviour is formed.

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 emer-

gent 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 marks a move away from the agenda that has shaped much of Machine Ethics over the past fifteen years—an agenda that addresses the same broad question we do, but at a Level of Abstraction concerned with how machines apply moral principles. Our aim is different. 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 (LoA) 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 Human-Robot Interation (HRI) serve not as neutral measurement devices but as theoretically motivated probes into the latent dispositional manifold (let us call it β_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 attempts to show that each tool targets a distinct dimension of the architecture through which moral salience is encoded, transformed, and expressed in behaviour. Their role is therefore conceptual rather than merely procedural: these instruments define the coordinate system in which the perturbation introduced by synthetic presence becomes detectable as a deformation of the evaluative field rather than as a trait-driven behavioural fluctuation.

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

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

Chapter 5 constitutes the empirical core of the thesis. It operationalises the evaluative–topological model developed in the earlier chapters into a full experimental framework, integrating design, measurement, and statistical inference into

a single methodological architecture. The chapter introduces the controlled observational conditions, reconstructs the Watching–Eye paradigm, and justifies the use of the NAO platform as a parametrically stable source of synthetic presence. It specifies all behavioural measures, psychometric instruments, and salience manipulations, and it details the complete analytical pipeline—from preprocessing and cluster formation to non-parametric tests, regression modelling, and Bayesian estimation.

Its function is foundational: this is the chapter in which the three central hypotheses of the thesis—Evaluative Deformation, Synthetic Normativity, and Synthetic Perturbation of Moral Inference—are formally operationalised and subjected to empirical test. By consolidating the full experimental architecture with the statistical logic required to evaluate deformation in the evaluative field, the chapter provides the decisive evidence for the thesis' central claim: that synthetic 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 interprets the experimental findings within the cognitive–topological framework developed in the earlier chapters. It examines how the robot's silent co-presence attenuates prosocial donation and argues that this modulation is best understood as a subtle reshaping of the pre-reflective pathways through which moral salience is registered and transformed into action. The chapter situates this effect within intuitionist Moral Psychology, integrates it with Floridi's Levels of Abstraction, and contrasts the resulting ecological account of moral influence with agent-centred approaches in Machine Ethics. It concludes by outlining the conceptual, empirical, and governance implications of treating synthetic presence as a perturbation of human moral environments rather than as a candidate locus of moral agency.

Chapter 7 reconstructs the major normative traditions— deontological, consequentialist, virtue-theoretic, sentimentalist, contractualist, particularist, and pluralist—at the Level of Abstraction (LoA) appropriate to the aims of the thesis. The chapter distinguishes their reflective and justificatory role from the cognitive–affective mechanisms that generate everyday moral behaviour, and uses Floridi's LoA framework together with the evaluative–topological model developed earlier to interpret these theories not as implementable rule systems but as structural patterns—constraints, value gradients, dispositional tendencies, and affective vectors—that organise how moral evaluation is formed and directed.

This reconstruction serves a methodological purpose: it provides the normative coordinates required to assess the experimental perturbation ethically rather than merely descriptively, and it establishes the conceptual infrastructure that renders the thesis's hypotheses and their interpretation coherent within both normative theory and cognitive science.

Chapter 8 provides the structural integration of the thesis. It unifies the cognitive–affective architecture, the normative analyses, and the experimental findings into a single theoretical account of how synthetic presence perturbs moral cognition. Building on the experimental result—uniform attenuation of prosocial donation under humanoid co-presence—the chapter attempts to show 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 LoA 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 interpretive: it provides a theoretical lens through which the behavioural pattern observed in the data can be read as one possible indication of underlying evaluative dynamics, while also clarifying the kinds of questions that exceed the methodological reach of Machine Ethics. In this sense, the framework suggests how synthetic presence might be understood as a perturbation of moral appraisal without implying that the experiment discloses the full structure of moral cognition.

Finally, Chapter 9 returns to the thesis's central question—whether synthetic presence perturbs the evaluative transformation from moral cue to moral action—and synthesises how the empirical results, Bayesian modelling, and dispositional analyses jointly support the formal mapping $f(\alpha_E, \beta_C, \gamma_R)$. It suggests that NAO's co-presence subtly reshapes intuitive appraisal rather than explicit reasoning, highlighting moral cognition as a field-sensitive, environmentally modulated process. Structurally, the chapter integrates the thesis's psychological, topological, ethical, and HRI components into a unified explanatory framework, clarifying why synthetic systems possess moral relevance as environmental operators rather than as agents. It closes the conceptual loop opened in the this chapter by consolidating the empirical and philosophical contributions into a coherent account and outlining the broader research programme that follows from this work.

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: Existing Approaches and the Level-of-Abstraction Problem

2.1 Introduction: Scope, Objectives, and Theoretical Commitments

Machine Ethics is often described as the branch of Artificial Intelligence concerned with building machines that can "act ethically." But that description, while technically correct, misses the more interesting question—the one that quietly motivates the field [36, 23, 20].

At its core, Machine Ethics tries to understand what happens when artificial systems begin to participate in the moral situations humans inhabit—not in the abstract, not in the controlled space of philosophical thought experiments, but in the ordinary, textured environments where people make decisions, respond to one another, and register the subtle pressures of social life [37, 38, 39].

What has changed in the past decade is not merely the level of attention the field receives, but the kind of attention. Researchers in Machine Ethics now draw on Psychology, Cognitive Science, Philosophy, HRI, Affective Computing, and Behavioural Economics—not as optional glosses, but as integral sources of evidence about how moral judgment actually works [34, 40, 41, 42, 43, 44, 45]. A parallel shift has been driven by the rise of large language models, which make it possible to examine moral cognition not only through human behaviour but also through models trained on large-scale moral and social corpora [46, 47, 48, 49]. While these systems do not replicate human judgment, they provide a complementary lens for testing how moral cues are represented, transformed, or misrepresented in artificial agents. For the first time, the field can move beyond purely theoretical constructions and test its assumptions against empirical data that reveal how humans negotiate moral space in the presence of artificial systems [35, 50, 51, 52].

In this sense, Machine Ethics has become something like a laboratory for Moral Philosophy. Ideas that once lived comfortably as small-scale thought experiments—trolley problems, dilemmas of autonomy, abstract accounts of agency—*must now* be examined under conditions where artificial systems interact with real people, in real environments, with real consequences [53, 54, 55, 34]. The theories that survive this transition do so not because they are elegant, but because they remain intact when exposed to the messiness of human moral cognition [16, 17, 31]. And this is precisely where the field finds both its opportunity and its challenge.

The question is no longer only how to encode moral principles into machines [36, 20]. It is how to understand the ways in which artificial systems reshape the moral landscape itself—the cues people attend to, the expectations they carry, and the background sense of what is appropriate or permissible [56, 38, 57]. Machine Ethics, in this broader and more mature form, is not simply about

building ethical machines. It is about understanding how machines and humans co-construct the environments in which moral meaning is formed [58, 41, 59].

Together, these developments mark a turning point. Machine Ethics has expanded from a largely theoretical inquiry into a domain where philosophical models, behavioural evidence, and computational systems now interact. Yet precisely because the field has grown in scope, it has also grown in complexity: the assumptions embedded in its frameworks, the *Levels of Abstraction (LoA)* at which its explanations operate, and the mechanisms they presuppose are no longer uniform. Before we can interpret the experiment that follows in Chapter 5, we must therefore clarify the conceptual terrain on which such interpretations depend.

This chapter establishes that terrain. It examines the frameworks that claim to explain how artificial systems acquire moral relevance and identifies the points at which they illuminate—or obscure—the phenomenon under investigation. This review is not background filler; it is the first test of the assumptions on which the later analysis relies: the LoAs they occupy, the cognitive processes they take for granted, and the gaps they leave unexplained. It lets us ask:

> *Whether synthetic presence really does modulate the path from perception to action, which existing frameworks can even register that phenomenon—and which ones are structurally blind to it?*

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 tem-

plates, 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 [60, 61, 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 [62, 63, 64, 65, 66]. They also exhibit high performance on benchmarked moral analogy tasks and moral classification challenges [67, 68]. 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, 69], to dual-process theory [31, 54, 17], to affective neuroscience [70, 71, 72], to embodied and socially embedded models [73, 34, 74]—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 [36, 23, 75, 76].
- **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 [49, 77, 78, 79, 80].

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, 81].

As several recent critical analyses emphasise, LLMs produce moral reasoning without moral cognition [80, 79, 82]. They resolve dilemmas fluently [1]; they do not reproduce the cognitive–affective processes by which humans come to feel that something is a dilemma in the first place. Moral language is not moral experience. Reflective justification is not perceptual-affective appraisal.

If there is a chronological lesson, it is this: the technologies have changed, but the Level-of-Abstraction mismatch has not. The surface of Machine Ethics has shifted, yet the underlying category error remains fixed. And this is the hinge of the present work: the decisive starting question is not how well artificial systems can perform reflective moral reasoning, but how their presence intervenes in the pre-reflective, perceptual–affective processes that carry moral salience into action. It is this generative layer—not the reflective one—that the experiment must speak to. And the decisive shift is this: artificial systems no longer occupy only the reflective space in which their moral language lives. They enter the very environments that scaffold human moral perception—phones, rooms, shared spaces—altering what is noticed, how situations feel, and how moral salience flows [83, 84, 52, 85, 86, 41]. Presence, not reasoning, becomes the point of intervention [58, 59].

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 often proceeds by treating the principles of an ethical theory as if they could stand in for the cognitive machinery of a moral agent—*as though human behaviour were governed by internal procedures resembling Kantian tests or utilitarian calculations.* 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, 87, 6].

[1] The distinction between reflective and generative Levels of Abstraction (LoAs) is crucial here. Moral justification, principle-balancing, and linguistic explanation occur at a reflective LoA [25, 26]. Human moral behaviour, by contrast, arises from perceptual, affective, and socially embedded processes documented across moral psychology and social neuroscience [16, 31, 72, 71]. 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 [80, 79, 82]. 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, 87, 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, 61, 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].

The literature review therefore brings a particular limitation into focus. Classical Machine Ethics is methodologically refined and theoretically coherent, but it is oriented toward a Level of Abstraction that does not readily register the phenomenon examined here. As work in Moral Psychology, Affective Neuroscience, Social Signal Processing (SSP), and HRI indicates, the dynamics most relevant to our question arise within the evaluative substrate of salience, affect, and social interpretation—not primarily within the reflective principles articulated after the fact [42, 33, 34, 35].

Classical Machine Ethics remains an elegant and logically disciplined project, yet it approaches moral agency through principles, rules, and reflective argumentation [21, 20, 23, 22]. By contrast, the processes that shape immediate moral behaviour appear to operate at an earlier level of appraisal, where salience, emotion, attention, and social interpretation structure what becomes behaviourally relevant [16, 31, 18, 87, 6]. In this sense, the approaches are not in conflict but address different layers of moral cognition, and only one of them engages the generative processes at issue in this thesis.

If we attend to the wrong layer of analysis, the phenomenon simply remains invisible. This is why the same difficulty appears in what is now called Computational Morality. Whether implemented through logic engines, preference aggregators, or LLM-based moral modelling, these approaches treat moral judgment as a problem of symbolic inference—as if the structure of behaviour could be captured by procedures for generating reasons [24, 27, 28, 77, 49].

Empirical research in Moral Psychology, Affective Neuroscience, SSP, and HRI offers a different picture. Moral judgments typically arise from fast, intuitive, and affectively charged appraisal [16, 10, 17]. They are shaped by presence, gaze, tone, perceived stake, and the social and affective cues embedded in the environment [18, 6, 42, 35]. These processes are context-sensitive and dynamically

assembled [19, 32]. When morality is modelled as a chain of propositions—if A then B, if C then D—the generative machinery that produces behaviour is abstracted away. And it is precisely this machinery that our experiment is designed to probe, given that even silent humanoid co-presence can shift the evaluative conditions in which action is selected [33, 34, 74].

In other words, the classical computational models of Machine Ethics are not limited by deficiencies in their logical structure; rather, they address a different component of the moral process than the one examined here. They aim to model explicit reasoning, while much of the behaviourally consequential activity appears to unfold within the evaluative landscape that precedes reflective judgment. It is within this substrate of salience, affect, and social interpretation that synthetic presence is most plausibly understood to exert its influence.

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.

As we will see that the robot employed in this study has no beliefs, goals, intentions, or communicative acts; it is not reasoning or attempting to influence participants. Prior work in HRI and social psychology indicates that even minimally expressive robots can modulate certain aspects of human social behaviour [33, 74, 88, 34]. What remains less well understood—and what the present thesis investigates—is how such synthetic presence interacts with the evaluative processes specific to moral judgement. These dynamics involve shifts in salience, attention, and perceived social presence rather than changes in explicit reasoning [89, 2, 70, 72]. Because such processes operate upstream of deliberation, they fall outside the scope of rule-based, utility-theoretic, or propositional models of morality, which focus on reflective justification rather than the intuitive, affective systems documented across moral psychology [16, 54, 17, 90, 75].

At this stage, the literature points to a difficulty that classical Machine Ethics has not fully resolved. If the aim is to understand how humans behave morally in the presence of artificial agents—and to model those behaviours in a form that artificial systems could eventually operationalise—then the foundational assumptions of principle-first approaches require re-examination. Deontic, Utilitarian, and Virtue-theoretic models presuppose that moral norms can be rendered as explicit rules or evaluative operators [21, 20, 22, 23]. Yet empirical research in Moral Psychology and Affective Neuroscience indicates that morally relevant behaviour typically emerges from cognitively embedded processes of appraisal, salience detection, affective resonance, and social interpretation [16, 31, 18, 87, 6]. These findings do not contradict normative theories, but they suggest that reflective principles capture only the justificatory layer of moral cognition, not the genera-

tive processes on which the present thesis focuses.

From this perspective, moral norms are not best understood as rules to be encoded, but as structural patterns that organise evaluation. They shape constraints, emphases, and trajectories within the space of moral appraisal, operating at a reflective Level of Abstraction that specifies justificatory standards rather than the cognitive mechanisms that produce behaviour [15, 61, 25, 26]. Their influence depends on how they interact with, and are realised through, lower-level processes of perception, affect, and social interpretation [14, 13, 32].

For similar reasons, moral judgement cannot be modelled as pure reasoning or symbolic inference. Dual-process and intuitionist models suggest that intuitive, affectively charged appraisals precede reflective deliberation and constrain the space of admissible reasons [10, 16, 17]. Attention, empathic resonance, perceptual salience, and social–contextual modulation shape the evaluative landscape long before propositional reasoning becomes active [19, 18, 42]. These dynamics provide the psychological foundation for the account developed in Chapter 3.

Nor can artificial agents be understood as carriers or executors of moral values. Work in HRI and Social Signal Processing suggests that artificial systems often function as *modulators* within the environment, shaping patterns of salience, perceived social presence, accountability cues, and evaluative expectations [33, 34, 35, 74]. These influences appear to operate upstream of explicit judgment, within the same cognitive–affective processes that organise the evaluative conditions under which moral decisions are formed.

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, 61], 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 Two Levels of Abstraction in Machine Ethics

What the literature refers to as "Machine Ethics" encompasses two distinct research programmes that share a name but operate at different Levels of Abstraction [25, 26]. One approaches moral behaviour as a computational problem and asks how ethical principles might be implemented in artificial agents. The other investigates how artificial systems participate in, and sometimes perturb, the situations in which human moral behaviour unfolds. Their conflation is understandable, but it obscures the distinct explanatory aims of each approach and makes it less clear which framework can address the phenomenon investigated in this thesis.

To make this distinction explicit, the section introduces each programme in turn,

before clarifying why the present work aligns with one rather than the other.

The first research programme is what may be called *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 foundation comes from Human–Robot Interaction, Media Psychology, and Social Signal Processing. Work in these areas suggests that artificial systems—whether humanoid robots, embodied agents, or even minimally interactive media—can modulate attention, empathy, prosociality, and interpersonal expectations through their mere presence [91, 42, 33, 34, 74, 35]. This programme therefore aligns closely with the phenomenon examined in this thesis: the possibility that a robot's silent co-presence perturbs the evaluative processes through which moral behaviour is generated.

The second programme is *Computational Machine Ethics*. This project seeks to design systems that make ethically adequate decisions by embedding moral theories into computational architectures. Deontic logics, Utilitarian optimisation engines, rule-based governors, and virtue-theoretic templates fall under this category [21, 20, 22, 23, 24, 27]. Here, moral judgement is commonly treated as a reasoning problem: behaviour is modelled as the outcome of applying principles, performing symbolic inference, or satisfying constraints. In this respect, Computational Machine Ethics addresses a different explanatory target than the one motivating the present study. Although the literature often treats these two strands as if progress in one should inform the other, they operate at different Levels of Abstraction and address different questions.

Human–Machine Ethics investigates how artificial systems modulate human evaluative processes, whereas Computational Machine Ethics aims to formalise normative content in ways that could guide artificial agents. Both domains contribute to the broader landscape, but they do so from distinct conceptual vantage points.

The empirical findings of this thesis highlight why maintaining this distinction is important. Research in Psychology and HRI indicates that artificial systems can influence patterns of attention, salience, and social interpretation [6, 18, 16]. Such influences operate within the pre-reflective evaluative processes that shape moral behaviour. By contrast, Computational Machine Ethics focuses on reflective, principle-driven models of ethical decision-making, and therefore does not readily accommodate phenomena that arise upstream of deliberation [31, 10, 87, 32]. The approaches are thus not incompatible, but they illuminate different components of moral cognition.

Seen in this light, the apparent lack of unity in "Machine Ethics" reflects the field's underlying conceptual structure rather than a terminological accident. One strand is empirically grounded, concerned with how humans behave in sociotechnical environments; the other is formally oriented, concerned with how ethical principles might be encoded into artificial agents. The present work belongs to the former: it examines how artificial systems, even when passive, may *modulate the evaluative field* within which human moral decisions take shape.

2.3 Clarifying the Explanatory Level of the Present Work

It is natural to ask how this research should be situated. It touches Affective Computing, with its emphasis on modelling emotion computationally [43]; Human–Robot Interaction, where the behavioural consequences of artificial social agents are examined [33, 34, 74]; and Moral Psychology, which analyses the cognitive and affective substrates of moral behaviour [16, 31, 18, 87]. Each contributes something essential, yet none, on its own, provides the conceptual resources needed to address the phenomenon at stake. For the purposes of this thesis, the disciplinary label is secondary; the priority is to clarify the explanatory level at which the question must be framed.

The central confusion this thesis confronts is therefore not empirical but conceptual. Work collected under the name "Machine Ethics" has often blurred two distinct enterprises: understanding how humans behave morally in sociotechnical settings, and designing machines that behave in accordance with encoded ethical theories. These projects operate at different Levels of Abstraction [25, 26], draw on different forms of evidence, and address different explanatory aims. Treating them as a single field has created a methodological entanglement in which normative elegance can obscure the generative processes that empirical research suggests are crucial for understanding moral behaviour in human–robot contexts.

Applying the LoA distinction clarifies the source of the confusion. Human moral behaviour unfolds at the cognitive LoA: it is shaped by perceptual salience, affective resonance, attentional dynamics, and social-cue interpretation [16, 10, 18, 6]. Normative theories—deontological, utilitarian, contractualist—operate at a reflective LoA concerned with justification rather than generation [15, 61]. Treating high-LoA principles as if they functioned as low-LoA psychological mechanisms tends to obscure the evaluative processes from which behaviour emerges.

The experimental findings of this thesis are consistent with this distinction. A humanoid robot with no beliefs, goals, or communicative acts nonetheless appears to influence the evaluative conditions under which participants move from moral perception to prosocial action. Such modulation is more plausibly associated with changes in salience, affective alignment, and attentional orientation than with reflective reasoning [31, 18, 6]. Frameworks that model moral action primarily as rule retrieval, utility computation, or principle execution therefore address a different LoA and do not readily accommodate these upstream dynamics.

This is why the disciplinary categorisation of the work is secondary. The issue is not where the research should be filed, but what becomes visible once the LoA distinction is applied. Through this lens, the field divides into two legitimate but distinct activities: *Human–Machine Ethics*, an empirically grounded inquiry into how artificial agents modulate human evaluative processes [42, 74, 35]; and *Computational Machine Ethics*, a reflective programme concerned with principled design, drawing on formalisms such as Deontic Logic [21, 20], Utility Optimisation [22], and Virtue-theoretic modelling [23]. Maintaining this distinction clarifies the explanatory space in which the present work is situated: it investigates how artificial systems, even when passive, may modulate the evaluative field within which human moral decisions take shape.

Clarifying the Levels of Abstraction is only a first step. Once the distinction is restored, a different research agenda comes into view. Moral behaviour is not well understood as the output of rule application or principle execution; it is better characterised as emerging from a dynamic evaluative field shaped by affective gradients, perceptual cues, attentional flows, and socially mediated expectations [16, 10, 18, 42]. Artificial agents—robots, avatars, conversational AIs—can interact with this field simply by being present. From this perspective, a productive programme for moral AI begins not with ethics construed as a set of principles, but with the cognitive and affective architecture through which moral perception becomes action.

Several methodological implications follow from this shift in perspective. **First**, empirical grounding becomes essential. Any account of moral behaviour must engage with findings from Moral Psychology, Affective Neuroscience, Developmental Research, HRI, and Social Signal Processing. A model that cannot accommodate the influence of gaze, posture, co-presence, or anthropomorphic cues risks omitting mechanisms that empirical work suggests are behaviourally significant [1, 2, 6]. **Second**, artificial agents are more usefully modelled as operators within the evaluative environment rather than as reasoners applying rules: their influence appears to arise from how they modulate the conditions under which humans act [33, 35, 34]. **Third**, normative theory is best interpreted in structural rather than procedural terms: norms describe constraints, gradients, and dispositions within the evaluative space through which behaviour unfolds [14, 13, 32].

This reframing also clarifies a question that often arises in engineering contexts: what, if anything, is the practical implication? The implication is not a new ethical theory to encode, nor a list of principles to implement. Rather, it is the recognition that artificial agents influence human moral behaviour primarily through their presence—by shaping salience, attention, and social interpretation—rather than through argument or explicit reasoning. Ignoring these evaluative processes risks overlooking the very mechanisms through which artificial systems can come to matter morally.

From this perspective, the future development of moral AI is less about constructing machines that reason like moral philosophers and more about understanding how artificial systems participate in the evaluative conditions that guide human action. A coherent research programme must therefore attend to the perceptual, affective, and social dynamics through which moral behaviour is generated. In bringing the LoA distinction to bear on the literature, this section has identified the conceptual terrain on which such a programme can be built and has clarified why the generative processes of moral cognition—not their reflective-justifications—form the appropriate locus of explanation for the phenomenon investigated in this thesis.

2.4 The Cognitive–Affective Foundations of Moral Judgment

Once the conceptual landscape is clarified, the next step is to start considering the machinery from which moral behaviour is generated. Here the empirical picture is not merely extensive but strikingly consistent. Across moral psychology, affective neuroscience, and behavioural science, evidence suggests that moral judgment

is shaped primarily by fast, intuitive, emotionally charged processes [16, 31, 10]. These processes respond to perceptual salience, attentional capture, empathic resonance, and situational cues long before reflective reasoning is engaged. Slower, deliberative processes do play a role, but typically by refining, justifying, or overriding an appraisal that has already been formed [17, 18, 87]. From a cognitive LoA perspective, it is this intuitive–affective layer that performs the primary generative work in moral cognition.

The major theoretical frameworks in the field converge on this interpretation. Haidt's Social Intuitionist Model [16], Greene's neurocognitive dual-process account [31, 10, 54], and Cushman's action-based inference models [17] each propose that moral evaluation begins with rapid, affectively valenced appraisals, with reflective reasoning entering later and often retrospectively. Neuroscientific work supports this view: affective tagging, motivational relevance, empathy circuitry, and social-interpretive processes are recruited early in the perceptual stream [19, 18, 6]. These findings locate the generative locus of moral cognition firmly within the evaluative processes that precede explicit judgment.

This cognitive–affective picture stands in contrast to the philosophical traditions on which classical Machine Ethics has relied. Kantian ethics, utilitarian frameworks, and contractualism articulate *justificatory* structures—conditions of universalisability, procedures for aggregating value, or standards of interpersonal reason-giving [15, 61]. These theories operate at a reflective Level of Abstraction: they describe how actions can be defended, not the psychological mechanisms through which moral judgments arise.

Classical Machine Ethics, however, adopted only this reflective dimension and treated it as if it described the generative architecture of moral agency. It presumed that humans behave by applying principles and that artificial systems could do likewise by encoding those principles into computational structures [21, 20, 22, 23]. Empirical research suggests a different picture: moral behaviour appears to emerge from a cognitive–affective substrate shaped by salience, emotion, attention, embodiment, and social interpretation rather than from the direct application of explicit rules.

This perspective helps explain why studies of human moral behaviour in context—across HRI, Media Psychology, and Social Signal Processing—consistently identify patterns governed by attentional capture, affective resonance, perceived monitoring, and contextual meaning [91, 42, 33, 34]. A clear illustration is the Watching–Eye effect: minimal cues of observation, even stylised eyes, can modulate prosocial behaviour [1, 2, 5]. Such shifts do not reflect the endorsement of principles but subtle environmental modulation of the evaluative posture from which moral action is selected.

The cognitive level—the terrain of salience, empathy, vigilance, and contextual modulation—is where much of moral behaviour appears to be shaped. It is also the level at which the attenuation effect observed in our experiment is most plausibly situated. The humanoid robot neither reasons nor requests anything; yet its silent co-presence seems sufficient to shift the evaluative field from which prosocial action is selected. This is the cognitive–affective layer at work—the layer classical Machine Ethics did not attempt to model.

From this perspective, a structural implication follows. If moral behaviour emerges from processes of perceptual salience, affective pull, attentional alignment, and social interpretation, then computational models that cast moral agency as rule retrieval or propositional inference address a different phenomenon. They remain coherent at the reflective LoA, but they leave the generative cognitive LoA unrepresented, and thus offer an incomplete account of the mechanisms through which behaviour arises.

This brings the discussion back to Levels of Abstraction. Once the mismatch between reflective theories and generative processes is recognised, it becomes clear why many debates in Machine Ethics have struggled to engage the kinds of phenomena examined here. The remainder of the thesis develops a framework in which moral cognition, evaluative topology, and synthetic presence can be understood in principled alignment. In this sense, the present section has established the conceptual ground on which the subsequent analysis rests.

With this distinction established, the path forward becomes clearer. The present section has identified why reflective, principle-based models do not address the generative processes through which moral behaviour emerges. The next step is to develop a positive framework suited to that task: one that locates moral cognition at the appropriate Level of Abstraction and explains how synthetic presence can perturb the evaluative structures linking perception to action.

2.5 Levels of Abstraction and the Structure of Machine Ethics

A central conceptual tool for clarifying the landscape of Machine Ethics has been Floridi's notion of a *Level of Abstraction* (LoA) [25, 26]. The idea is structurally simple but analytically powerful: any explanation depends on specifying the level at which a system is being described. The LoA determines which variables are observable, the appropriate grain of description, and the kinds of explanations that can be meaningfully offered. Ethical theories operate at a high, reflective LoA: they articulate justificatory structures—principles, universalisability tests, value aggregation procedures, and reason-giving relations [15, 61]. Moral psychology, by contrast, works at a lower, cognitive LoA: it investigates the mechanisms that *generate* moral judgment, including perceptual salience, affective appraisal, attentional dynamics, and social interpretation [16, 31, 10, 17, 18].

Difficulties arise 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 blurs and with it the capacity to explain the processes that underlie behaviour. Much of the classical literature in Machine Ethics illustrates this tension. 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, these approaches implicitly assumed that moral agents—human or artificial—act by applying principles [21, 20, 22, 23]. Yet such principles occupy the reflective LoA: they explain *why* an action may be defensible, not *how* a moral judgment is generated.

When reflective principles are used as behavioural generators—as algorithms intended to produce moral action—the resulting models may be internally elegant but remain out of step with what empirical research suggests about human moral

cognition. Moral behaviour appears to arise not from propositional logic or rule execution, but from patterns of perceptual salience, affective appraisal, attentional orientation, and social interpretation [16, 54, 18, 42]. These cognitive–affective processes form the conditions under which high-level principles acquire practical significance.

This distinction is useful for interpreting the experimental findings presented later in the thesis (in Chapter 5). In the Watching–Eye paradigm, minimal cues of observation typically increase prosocial behaviour [1, 2, 5]. Yet when a silent,non-agentic humanoid robot is introduced into the same environment, this pattern appears attenuated. No reasoning, communication, or intentional signalling is involved. The modulation is more plausibly associated with shifts in salience, affective alignment, and perceived social presence [33, 34, 35]. The accountability cue therefore loses some of its usual influence not because a principle is misapplied, but because the cognitive substrate on which such cues depend has been altered.

Taken together, these considerations suggest that moral action is shaped less by the execution of explicit principles than by the dynamic interaction of perceptual, affective, and social processes. Classical, principle-first models operate at a reflective LoA and therefore illuminate a different aspect of moral life than the one examined in this thesis. What contemporary sociotechnical contexts bring into view—and what the empirical findings are consistent with—is that artificial agents can influence human behaviour through their impact on the evaluative conditions under which moral decisions are formed.

The thesis proceeds from a methodological commitment that follows from this analysis: *before we can design artificial systems with moral capacities or constraints, we must understand how such systems reshape the conditions under which human moral experience unfolds.* This requires reversing the typical order of explanation. Rather than beginning with ethical theory and working downward, the inquiry must begin with the empirical architecture of moral cognition, determine how artificial agents interact with it, and only then ask which forms of ethical oversight or design constraint are appropriate.

Seen through the lens of Levels of Abstraction, the wider landscape of Machine Ethics becomes more intelligible. Distinct questions can be situated at the levels where they can be meaningfully addressed, and debates that once appeared unresolved take on a clearer structure. For the purposes of this thesis, the restoration of LoA discipline provides the conceptual foundation needed for the next stage of the argument: developing an account of moral cognition and synthetic presence that aligns the empirical findings with the explanatory level at which moral behaviour is actually generated.

2.6 Evaluative Topology, Affective Architecture, and Synthetic Moral Perturbation

The previous sections showed that classical Machine Ethics, by operating at a reflective Level of Abstraction (LoA), overlooks the mechanisms that generate moral behaviour. Restoring LoA discipline reveals that the explanatory work must begin not with principles but with the cognitive–affective substrate from

which moral appraisal and action arise. The task of this section is to make that substrate explicit and to articulate the positive framework that replaces the principle-first paradigm.

The central claim is that moral behaviour emerges from the organisation of an *evaluative field*: a structured, dynamic configuration shaped by gradients of salience, affective resonance, attentional orientation, contextual norms, and implicit social meaning. Ethical theories operate within this field as high-level structural constraints rather than algorithmic generators [15, 61]. Their behavioural 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 idea of an evaluative field synthesises three well-established strands of empirical research and places them in the conceptual architecture needed for the remainder of the thesis.

(1) Moral psychology: affect, intuition, appraisal. Dual-process theory [31, 10, 54] and the Social Intuitionist Model [16] converge on the view that moral evaluation begins with rapid, affectively charged appraisals. Affective tagging, empathic resonance, and motivational relevance are recruited early in the process [18, 87, 19]. Perceptual salience, attentional capture, and intuitive heuristics structure the evaluative landscape before reflective reasoning comes into play.

(2) Social Signal Processing and affective computing: cue modulation. Work in Social Signal Processing shows that gaze direction, morphological cues, co-presence, and implicit monitoring reshape attentional and affective weighting prior to explicit cognition [91, 42, 6]. Human–Robot Interaction research extends this finding to artificial systems: humanoid robots and other embodied agents modulate perceived social presence, agency, and interpersonal expectations through mere presence [33, 34, 35, 74]. These modulations operate at precisely the pre-reflective level identified by moral psychology.

(3) Normative theory: structural constraints. Philosophical ethics contributes the insight that moral theories specify *structural constraints* rather than behavioural generators. Deontological principles articulate exclusionary limits [15]; consequentialism defines gradients of evaluative weight; virtue-theoretic accounts identify attractors within patterns of action and perception [14, 92]; sentimentalist and contractualist theories specify affective vectors and justificatory relations [32, 93, 61]. These structures shape the evaluative field but do not, on their own, produce moral behaviour.

When we consider these findings together, a more complete picture emerges. Moral evaluation begins with rapid, affectively charged appraisals. As work in moral psychology shows, what agents notice, how their attention is guided, and the intuitive heuristics they employ shape the evaluative field before any reflective reasoning occurs. These early processes are highly sensitive to social cues. Research in Social Signal Processing, and later in Human–Robot Interaction, shows that gaze, posture, co-presence, and other behavioural signals can alter patterns of attention and affective weighting in ways that precede and influence

explicit judgement. Artificial agents participate in this structure. By their mere presence, they can modify perceived social salience, agency, and interpersonal expectations, thereby reshaping the evaluative landscape in which people decide how to act. Normative theories play a different role. They do not describe the causal mechanisms that generate behaviour. They specify structural constraints—principles that limit certain actions, gradients of evaluative weight, or patterns of perception and response that agents have reason to cultivate. These constraints guide moral assessment, but they are realised only within the cognitive and social processes that shape an agent's immediate evaluative stance. When viewed through Floridi's Levels of Abstraction, these strands form a coherent model. High-level normative theories define the standards by which actions can be assessed; low-level cognitive and affective mechanisms determine the generative processes through which agents arrive at their decisions; and social cues, including those introduced by artificial systems, modulate the field within which both operate. This is the evaluative architecture from which moral action emerges.

2.6.2 Moral Behaviour as Movement Within an Evaluative Field

Within this framework, moral behaviour can be understood as emerging from how an agent moves through the evaluative field formed by perceptual, affective, and social cues. Attention amplifies or suppresses features of the environment, altering which cues become behaviourally salient [94]. Affective processes saturate parts of the field with motivational relevance [87, 18]. Contextual cues shift the weight of norms, expectations, and interpersonal meaning [16, 4]. Social signals further modulate perceived accountability and relational orientation [1, 2, 5].

Several kinds of processes shape the evaluative field in which moral decisions are made. What agents attend to can amplify or suppress particular cues, locally altering the salience landscape [94]. Their affective responses saturate parts of this field with motivational force, influencing how certain possibilities are experienced [87, 18]. Contextual cues may deform the gradients that represent competing obligations, norms, or expectations, shifting how these considerations are weighted [16, 4]. Social signals—including cues of monitoring, accountability, and interpersonal meaning—can further modulate an agent's interpretation of the situation [1, 2, 5]. These influences do not compete with the structural elements described by normative theory. They belong to different explanatory levels. When seen through Floridi's framework of Levels of Abstraction, intuitive and affective mechanisms determine how agents move through an evaluative space whose higher-level structure is provided by reflective principles [25, 26]. This alignment dissolves the familiar rationalist–intuitionist divide. Rationalist accounts specify structural constraints—limits, weights, and relations that determine which considerations count as reasons—while the cognitive–affective domain explains how agents, in practice, navigate within those constraints [54, 61]. The two domains are therefore not rivals but parts of a single architecture, in which justificatory structures and generative mechanisms operate at different Levels of Abstraction.

2.6.3 Synthetic Presence as Evaluative Modulator

The experiment presented in Chapter 5 offers an empirical probe into this evaluative architecture. In the Watching–Eye paradigm, minimal cues of observation typically enhance prosocial tendencies through implicit monitoring [1, 2]. When a silent, non-agentic humanoid robot is introduced into the same environment, this pattern appears attenuated. The shift does not originate in reasoning or principle-application. It is more plausibly understood as a modulation of the evaluative conditions through which the cue acquires behavioural significance [33, 34, 35]. The robot's ambiguous social ontology—perceptually agentic but not fully classifiable—alters the affective and attentional weighting that normally supports the Watching–Eye effect.

Importantly, this modulation is *disposition-sensitive*. The Prosocial–Empathic ecology shows the strongest attenuation, consistent with its reliance on empathic resonance and interpersonal salience. The Analytical–Structured ecology shows moderate attenuation, reflecting its dependence on interpretive coherence more than affective pull. The Emotionally–Reactive ecology exhibits minimal change, as its evaluative patterns are less stable and provide fewer structured gradients for perturbation to act upon.

These differential patterns are consistent with the broader claim developed in this section: synthetic presence exerts its influence *upstream* of deliberation and principle. It modifies the evaluative field within which behaviour is formed rather than overriding or contradicting reflective judgment. Viewed in this light, the humanoid robot functions analogously to a *field operator*: its co-presence perturbs the local configuration of salience, affective alignment, and perceived social relevance, thereby altering the trajectory through which moral perception is carried into action. This operator-like role is not meant to imply a fixed mathematical mechanism, but it captures the structural insight suggested by the data—that synthetic presence acts on the generative conditions of moral cognition rather than on its reflective outputs.

2.7 A Topological Lens on Moral Modulation

Within the framework developed in the preceding sections, the attenuation observed in the experiment can be interpreted through the geometry of the evaluative field. Moral behaviour does not shift because a principle is misapplied or because reflective deliberation has failed. It shifts because the *field* in which such principles acquire behavioural force has been locally deformed. The perturbation occurs upstream of reasoning: at the level of salience gradients, affective vectors, attentional curvature, and perceived social ontology.

- Deontological constraints lose local traction when accountability salience collapses [2, 94].
- Consequentialist value-gradients flatten when contextual meaning becomes ambiguous [4, 95].
- Virtue-theoretic dispositions fail to express themselves when affective attractors weaken [92, 14].
- Sentimentalist vectors diminish when empathic resonance is displaced [32, 93].

· Contractualist justificatory relations loosen when the perceived social field becomes indeterminate [61].

These patterns are not failures of normativity but signatures of a deeper structure: moral behaviour is *field-sensitive*. Norms constrain the space of admissible trajectories, but the trajectories themselves are determined by the low-level dynamics of appraisal, orientation, and situational meaning. When synthetic presence enters the environment, it acts as a perturbation operator that alters the geometry of this field.

The experiment therefore provides a concrete instance of a more general topological insight. The humanoid robot—minimally expressive, non-agentic, and silent—serves as a parametric deformation of the evaluative manifold. Its ambiguous social ontology induces a measurable change in the local structure of the salience landscape, redirecting the flow from moral perception to action. In this sense, NAO functions as a *topological operator*: not a source of moral content, but a factor that modifies the shape of the field within which moral content is realised.

Seen through this lens, the attenuation effect is the behavioural trace of a geometric transformation. What changes is not the agent's principles, nor the agent's reasoning, but the *configuration of the evaluative space* in which those principles and reasons can do work. Synthetic presence thus reveals a property that classical Machine Ethics could not register: moral action emerges from the interplay of structural constraints and the topological dynamics of appraisal. The robot's role is not to replace these dynamics, but to reshape them.

2.7.1 Toward a Unified Framework

The notion of an evaluative topology provides the integrative architecture that the field has long lacked. It offers the structural bridge connecting normative theory, empirical psychology, and computational modelling. By distinguishing the Levels of Abstraction at which these domains operate, the framework clarifies how high-LoA normative structures shape the space of justification, while low-LoA cognitive–affective mechanisms determine the trajectories through which moral perception is transformed into action. Within this architecture, the influence of artificial agents becomes intelligible: they do not contribute beliefs, reasons, or principles, but modulate the salience, affective gradients, and social cues that configure the evaluative field itself.

This synthesis completes the conceptual turn initiated by the literature review. It repositions the problem of "moral AI" from the design of principle-following agents to the analysis of how artificial systems participate in—and occasionally perturb—the generative processes that give moral judgments their behavioural force. The chapters that follow build on this topological foundation to develop a formal account of machine-mediated moral cognition. In that account, artificial systems are not ethical reasoners but *operators on the evaluative field*: elements whose presence reshapes the conditions under which moral meaning is registered, weighted, and ultimately expressed in action.

2.8 Integrative Synthesis: Toward a Cognitive–Affective Framework for Machine-Mediated Morality

The analyses in this chapter converge on a coherent picture of moral behaviour under artificial co-presence. Classical Machine Ethics treats reflective normative theories as behavioural generators [21, 20, 22], yet moral psychology shows that action emerges from a cognitive–affective architecture structured by salience, attention, empathy, and contextual cues [16, 31, 17]. Work in HRI and Social Signal Processing demonstrates that artificial agents modulate these mechanisms through minimal social cues [42, 34, 33, 35]. Evaluative topology integrates these strands by modelling behaviour as movement within a salience-weighted, affectively structured field. The experiment in Chapter 5 is consistent with this view: synthetic presence perturbs the evaluative field upstream of deliberation and attenuates prosocial action.

Three conclusions follow from the literature:

1. **Moral behaviour is generated at the cognitive LoA.** Reflective theories provide justificatory standards [15, 61], but behaviour is shaped by low-LoA affective and social mechanisms [31, 18, 6].
2. **Artificial agents modulate the evaluative field.** Presence alone can shift attention, empathy, and perceived social meaning [91, 42, 33, 74]. The attenuation effect is consistent with this.
3. **A viable programme for moral AI must begin with evaluative topology.** Principle-executing architectures do not account for the mechanisms that generate behaviour [23, 22].

Taken together, these insights reframe moral AI: artificial systems must be understood as *operators on the evaluative field* rather than executors of moral rules. Synthetic presence reshapes salience, affective resonance, and accountability cues—perturbations that arise prior to explicit reasoning.

2.9 Global Synthesis: From Inferential Displacement to Synthetic Moral Topology

Across the reviewed literature, a consistent structure emerges: moral judgment and action arise from a cognitively embedded, affectively mediated evaluative field [16, 31, 18, 42]. Ethical theories operate reflectively, but behaviour is generated by the cognitive architecture through which salience and affect are organised. Artificial agents participate in this architecture by modulating these processes [34, 74].

2.9.1 From Question to Framework

The guiding question of this thesis—whether synthetic presence can perturb the inferential transformation from moral salience to action—emerges naturally from existing tensions. Classical Machine Ethics presumes principle-execution [21, 20], whereas Moral Psychology emphasises affectively charged appraisal [16, 17]. SSP and HRI show that social cues modulate these appraisals [91, 42, 33, 34]. Yet these strands have seldom been integrated.

2.9.2 Rationale for a Multi-Hypothesis Approach

The literature identifies three plausible loci of modulation:

1. **Evaluative deformation**: altered salience, monitoring, or affective weighting [1, 2, 94].
2. **Synthetic normativity**: effects arising from perceived social ontology [35, 33, 34].
3. **Perturbation of inferential pathways**: displaced empathy, attention, or contextual interpretation [18, 54].

No single mechanism captures the full range of effects; a multi-hypothesis framework is therefore required.

2.9.3 What the Literature Establishes

Three findings are robust across disciplines:

1. **Moral behaviour is field-sensitive**, shaped by salience, affect, and context [31, 16, 18].
2. **Artificial agents modulate the field**, altering social meaning, vigilance, and empathic stance [34, 74].
3. **Classical Machine Ethics cannot model this modulation** because it ignores the cognitive LoA where behaviour is generated [23, 22].

These findings motivate a shift: moral AI must be grounded in the structural dynamics of the evaluative field rather than in normative content alone.

The literature review reveals a domain-level misalignment: reflective ethical principles have been treated as if they described the psychological machinery of moral behaviour [21, 20, 22, 23]. Yet reflective norms articulate *conditions of justification* [15, 61], whereas behaviour arises from rapid appraisal and attentional dynamics [16, 31, 17]. Collapsing these levels obscures the mechanisms through which behaviour is modulated.

When empirical findings across Psychology, Affective Neuroscience, SSP, and HRI are placed together, a convergent evaluative architecture becomes visible. Moral judgment begins with rapid, affect-laden appraisal; social and contextual cues modulate evaluative weighting; and explicit reasoning intervenes downstream [18, 87]. Classical Machine Ethics never incorporated this architecture and could therefore not accommodate modulation phenomena.

Finally, the review isolates the theoretical gap motivating the experiment:

can synthetic presence perturb the evaluative field upstream of explicit moral reasoning?

No classical framework poses this question because none operate at the LoA where such perturbations occur. The study is designed to probe this precise gap. In sum, the literature review identifies the cognitive layer at which moral behaviour is generated, clarifies why prior models were misaligned with this layer, and isolates the phenomenon requiring empirical investigation. Within this thesis, the literature review is not merely preparatory; it constitutes the first scientific result.

If there is a single thread running through this chapter, it is that moral life isfar more delicate—and far more interesting—than the theories we usually buildto explain it. We began with frameworks, distinctions, and formal models, but what the literature ultimately reveals is something more human: the way attention pulls us, the way small cues colour our sense of what matters, the way meaning settlesinto a situation before we ever articulate a reason.

Artificial systems step into this landscape not as moral agents, nor as puzzles tobe solved, but as new participants in the background against which moralexperience unfolds. They shift the texture of a moment, sometimes imperceptibly,sometimes decisively, and in doing so they help us see what Moral Psychology, Information Ethics, and HRI each illuminate from different angles.

This chapter has tried to gather those strands into a single frame. The rest of the thesis asks what happens when we move from theory to experiment—when these ideas meet the subtle trajectories of real human behaviour. But the core insight remains the same: to understand how machines and humans share moral space, we must begin with the architecture of the human mind, and with the quiet ways in which meaning is shaped before any reasoning begins.

3. Cognitive Affective Architecture of Moral Judgement

Moral phenomena begin before explicit deliberation and the experiment in Chapter 5 captures a behavioural expression of this pre-reflective evaluative machinery. Moral behaviour is indeed the observable endpoint of a much earlier process: it does not begin with principles, arguments, or rules but in the quiet architecture through which we register what matters in a situation—what draws attention, what feels salient, and what acquires weight before any deliberate reasoning takes shape.

If this thesis asks whether the mere presence of a humanoid robot can alter moral behaviour, it must first offer an account of the evaluative machinery through which such behaviour is formed. Before anything else can be said about synthetic moral modulation, the central problem must be stated plainly. The experiment does not begin with behaviour,robots, or statistics; it begins with a conceptual uncertainty about the structure of moral cognition itself. If moral evaluation is shaped upstream—by attention, affect, salience, and social interpretation—then even a silent artificial body might, in principle, shift how those evaluative processes unfold. The project therefore opens with a single question, and everything that follows is an attempt to answer it:

> *Can the presence of a synthetic, non-agentic entity reshape the cognitive–affective transformation through which morally salient cues become action?*

Providing an answer is, in truth, an immense task. Questions about how perception acquires moral significance, how emotions and intuitions guide judgement, and how evaluative meaning crystallises have been debated for centuries across Moral Psychology, Phenomenology, Ethics, and Cognitive Science. No single thesis can reproduce or resolve that tradition. The aim here is more circumscribed and methodologically precise: to identify, at the appropriate Level of Abstraction, the cognitive–affective mechanisms that are *causally operative* in the phenomenon under investigation. What follows is therefore not a philosophical survey but a functional outline of the evaluative processes that operate upstream of deliberation and through which any perturbation—including synthetic presence—must pass in order to influence moral action.

Within this cognitive–affective frame, moral judgement emerges through a layered evaluative sequence: attention filters the scene, affective appraisals colour what is noticed, intuitive evaluations establish a sense of "how things stand," and only then do reflective reasons enter the picture. These mechanisms belong to a Level of Abstraction beneath the familiar structures of moral theory. They do not specify what agents *ought* to do; they determine how moral meaning becomes behaviourally actionable in the first place.

Robotic presence, if it influences moral judgment at all, must do so here—within the processes that govern salience, resonance, and social interpretation. The experiment described later in this thesis examines precisely such an influence, but its interpretation depends on identifying the evaluative substrate through which moral cues acquire force. Without that substrate clearly articulated, the phenomenon risks being misdescribed as a change in belief or a failure of reasoning, when it is instead a deformation of the field from which both arise.

Reflective moral theories—Kantian maxims, utilitarian calculus, contractualist reasoning— are all ethical frameworks that 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, 61, 60]. Any attempt to explain the experimental effect by appealing to ethical principles therefore begins at a Level of Abstraction the phenomenon does not inhabit [25, 26].

What is required instead is an account of moral cognition as an action-guiding evaluative process: one in which affect, attention, salience, social interpretation, and contextual meaning jointly determine how moral cues acquire behavioural force. Although rationalist traditions in moral psychology treat judgment as the outcome of reflective reasoning, a large body of empirical work demonstrates that mechanisms such as affective appraisal, empathic resonance, intuitive evaluation, and attentional modulation constitute the causal substrate of moraljudgement [16, 31, 10, 17, 18, 87]. Only within such a framework can the influence of synthetic presence be meaningfully specified. Without it, the experimental result would risk being mischaracterised as a shift in belief or a failure of deliberation, rather than a perturbation of the evaluative field that precedes both [6, 4, 94].

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 these mechanisms from the reflective structures of ethical theory; and it establishes the Level of Abstraction at which the central research question resides. Only on this foundation can the experiment be interpreted correctly. The study does not test principles, preferences, or doctrines. It probes the stability of the evaluative machinery through which moral meaning becomes action.

A recurrent theme across both the philosophical and empirical literature is that difficulties in Machine Ethics may arise from two related forms of misalignment: *category errors* and *LoA conflation*. Category errors occur when reflective normative principles are treated as if they described the psychological mechanisms that generate behaviour; LoA conflation arises when justificatory standards are interpreted as causal processes, or when descriptive regularities are taken to impose normative constraints [25, 26]. These misalignments can obscure the possibility that moral behaviour is shaped at a lower, cognitive–affective LoA documented in Moral Psychology and Social Cognition [16, 10, 18]. Without clarity about which LoA is operative, empirical patterns—such as the attenuation effects examined in Chapter 5—may be misread as failures of reasoning rather than as potential deformations in the evaluative conditions through which moral salience becomes action.

The remainder of the chapter therefore proceeds in a principled order. Before we can explain how synthetic presence perturbs the evaluative substrate of moral action, we must clarify the two domains that moral inquiry straddles: the descriptive domain concerned with how moral judgments are *generated*, and the normative domain concerned with how actions can be *justified*. Only by keeping these domains distinct can the phenomenon of synthetic moral perturbation be located with conceptual precision. Before turning to that distinction, it is worth pausing for a moment. Much of what follows is technical, but the motivation is simple: moral life is fragile in ways our theories do not always capture. A small shift in how a scene is perceived can change what feels salient, what feels asked of us, and how we respond. It is this subtle architecture—the quiet machinery beneath judgement—that the next sections aim to bring into focus, one layer at a time.

3.1 Descriptive and Normative Domains

The term "morality" spans two analytically distinct domains, which must be kept separate if the questions in this thesis are to be made precise. The first is *descriptive morality*: the empirical study of how human beings in fact form moral judgments, experience moral emotions, and act in ways that carry normative significance. This domain includes developmental accounts of the emergence of moral judgment [96], social–intuitionist and dual-process models of evaluative response [69, 97], affective and cognitive neuroscience of empathy and valuation [19, 18], and evolutionary explanations of cooperation and prosocial motivation [98, 99]. These approaches seek to describe the mechanisms that generate moral behaviour. They aim to explain *how* people come to judge, feel, or act, without making claims about whether those responses are justified.

The second domain is *normative morality*: the philosophical study of how one *ought* to act, and which moral principles one has most reason to accept. Here we find deontological theories that articulate constraints on action, consequentialist views that evaluate outcomes, contractualist accounts grounded in justifiability to others, and virtue-theoretic traditions emphasising character and cultivated dispositions [92, 100, 101, 102]. These theories do not describe psychological processes. They offer standards by which beliefs, emotions, and actions may be assessed as justified or unjustified.

Keeping these domains distinct is not a matter of terminology but of explanatory discipline. Descriptive theories trace the *causal pathways* through which moral behaviour is generated—how salience is allocated, how affect is recruited, how evaluative weight takes shape before any explicit reasoning occurs. Normative theories, by contrast, articulate the *standards of justification* by which actions may be assessed or criticised.

These domains are distinct but interdependent. Descriptive accounts reveal how agents in fact perceive and respond to a situation, while normative theories rely on that psychological architecture to remain action-guiding rather than fictional. Likewise, empirical models of moral cognition acquire meaning partly through the normative vocabulary within which moral judgments are articulated, even as those normative frameworks must remain constrained by what human agents are

psychologically capable of performing or understanding.

The phenomenon investigated in this thesis—whether synthetic presence can perturb the formation of moral judgments—belongs squarely to the descriptive domain. Yet it can only be understood with conceptual clarity when set against this normative background. Without the separation, one risks mistaking psychological mechanisms for ethical reasons, or treating justificatory principles as predictors of behaviour. The analyses that follow rely on this distinction to interpret the experimental findings at the correct Level of Abstraction.

The descriptive–normative distinction is introduced here because it marks the final conceptual boundary required before the thesis can turn to a precise account of moral cognition. Without it, two confusions would repeatedly undermine the scientific aims of the project.

First, moral terminology in technical disciplines is often used with unexamined ambiguity. Terms such as *obligation*, *responsibility*, *harm*, or *trust* circulate freely in HRI, AI ethics, and behavioural science, but their usage slides—often imperceptibly—between describing how people in fact respond to a situation and prescribing how they ought to. When these domains are not kept distinct, conceptual instability follows: empirical results are mistaken for ethical insights, and normative claims are misread as behavioural 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?*

Answering this question requires working at the LoA where moral behaviour is *generated*, not where it is *justified*. If this boundary is not drawn explicitly, one risks treating behavioural attenuation as a moral failure or misinterpreting reflective theories as if they were mechanistic explanations.

The descriptive–normative distinction therefore performs three indispensable functions.

1. It 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 an artificial agent enters the evaluative field. Such phenomena demand descriptive tools: models of salience, affect, attention, and social meaning [16, 10, 18, 94, 6].
2. It prevents the **misinterpretation of empirical effects as moral judgments**. If a robot's presence reduces prosocial behaviour, this is a psychological modulation—not an ethical lapse. It indicates a perturbation in the evaluative machinery that gives moral cues their behavioural force [1, 2, 5, 33, 34].
3. It 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, 87, 42, 6]. Normative theories explain justificatory standards [61, 15, 60], but they do not generate behaviour. Keeping the domains apart secures the LoA at which the empirical phenomenon actually occurs [25, 26].

Finally, clarifying this boundary ensures that normative theory can re-enter the discussion later—without conceptual conflation. When deontic invariants, consequentialist gradients, virtue-theoretic attractors, or contractualist equilibria appear in later chapters [54, 14, 92, 103], they will do so not as behavioural engines, but as structural constraints within the evaluative topology. This reinterpretation is only coherent when the descriptive and normative domains have been separated with care.

In sum, the distinction introduced here is not a philosophical detour but a boundary condition. It specifies the domain in which the thesis makes its claims, prevents methodological conflation, and secures the conceptual precision required to study moral perturbation under synthetic presence. The orientation is now clear: moral cognition is the object of analysis; normative theory provides the vocabulary of justification; and coherence requires that these domains remain distinct.

The project therefore turns, at this point, to a minimal operational definition of morality. This may feel abrupt, but its placement is deliberate. Having established the conceptual boundaries of the inquiry, we now require a definition that is precise enough to support empirical analysis yet modest enough to avoid the commitments of substantive normative theory. What follows provides that definition, and with it, the foundation on which the remainder of the thesis will build.

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 [104]; affective traditions emphasise the emotional systems that underpin moral concern [105]; rationalist accounts privilege normative reasoning [102]; social-scientific models attend to conventions and cultural norms [106]; evolutionary frameworks focus on the adaptive functions of cooperation and prosociality [98, 99].

Philosophical traditions likewise diverge in grounding morality in rationality, sentiment, virtue, utility, social contracts, or evolutionary pressures. Rationalist approaches trace their lineage to Kant's account of morality as grounded in pure practical reason and universal maxims [107, 108], later developed in pluralistic form by Ross [109] and in contemporary constructivist terms by Korsgaard [15]. Sentimentalist traditions locate the foundations of morality in affective experience, drawing on the classical work of Hume and Smith [110, 111, 112] and extended by modern accounts of emotion as evaluative perception [105, 113]. Virtue-ethical approaches interpret morality through character and practical wisdom, originating in Aristotle's treatment of virtue as cultivated sensitivity to salience [114] and revitalised by Foot, MacIntyre, and Hursthouse [115, 116, 92]. Consequentialist theories instead ground normativity in outcomes and aggregated welfare, from Bentham and Mill's classical utilitarianism [117, 118] to Sidgwick's formal unification [119] and later developments by Singer and Parfit [120, 121]. Contractualist and social-contract traditions ground morality in principles that

no one could reasonably reject or that would be chosen under conditions of fairness, tracing back to Hobbes, Locke, and Rousseau [122, 123, 124], and given contemporary articulation by Rawls and Scanlon [125, 126, 61]. Evolutionary approaches ground morality in the adaptive functions of cooperation, altruism, and shared intentionality, beginning with Darwin's account of sympathy and moral sense [127], and extended through foundational work by Trivers, Wilson, Boehm, and Tomasello [128, 129, 130, 99].

Computational treatments often inherit only one strand of this theoretical diversity. They tend to default to *rule-based perspectives*, particularly those associated with rationalist and deontological traditions in moral philosophy. These include Kantian approaches, in which moral evaluation proceeds through the application of universal maxims and categorical imperatives [107, 108], and Rossian pluralism, which frames moral judgement as the resolution of prima facie duties through principled deliberation [109]. Classical utilitarian, decision-theoretic accounts also lend themselves to rule-like operationalisation insofar as they cast moral evaluation as the application of calculable decision procedures over outcomes [118, 119]. It is these rule-governed structures—deontic imperatives, duty taxonomies, and outcome-based decision rules—that have historically been adopted within Machine Ethics as if they captured the generative machinery of human moral cognition [36, 20, 21]. This adoption appears to reflect not claims about descriptive accuracy, but the structural convenience such models offer for computational implementation [21, 20, 22, 23]. This structural convenience has, at times, encouraged an oversimplified picture of moral behaviour as though it were principally a matter of rule following. Such an interpretation can obscure the cognitive–affective processes through which moral judgements are formed, processes documented across moral psychology, cognitive neuroscience, and behavioural research [16, 31, 17, 18].

A central aim of this chapter is therefore to provide a corrective orientation: to articulate a framework informed by contemporary findings in Moral Psychology, Cognitive Science, and Social Signal Research [91, 42, 6]. Such a framework offers a more appropriate basis for the empirical and conceptual analyses required by the research question developed in this thesis.

3.1.2 Minimal Operational Definition for This Thesis

The survey above does not deliver a single, unified theory of morality—nor could it. Instead, it reveals the depth and heterogeneity of the concept. Across philosophical, psychological, evolutionary, and computational traditions, the term "morality" designates different objects: rules, reasons, emotions, virtues, practices, social contracts, adaptive strategies. What this overview provides, however, is the necessary backdrop for the next conceptual move.

To study how synthetic presence perturbs moral behaviour, the thesis must shift from the broad question of *what morality is* to the more specific and empirically tractable question of *how moral evaluations are generated*.

This is where the notion of *moral cognition* enters. The term does not replace the philosophical debates surveyed above, nor does it adjudicate between them. Rather, it identifies the cognitive–affective machinery through which moral con-

cern becomes psychologically operative. If "morality" names a family of normative and cultural frameworks, "moral cognition" names the set of processes that allow an agent to recognise normatively salient features, form evaluative judgments, and translate these into behavioural dispositions. It is at this level— the level of perception, salience, affect, and intuitive appraisal—that the phenomenon of synthetic moral perturbation must be located.

With this shift of focus, the thesis can now adopt the minimal, action-oriented definition required for empirical analysis. The definition is not intended to capture the philosophical richness of moral theory; it is intended to identify the causal substrate that the experiment must probe.

> *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.*

Two clarifications follow immediately. First, this definition is deliberately *non-substantive*: it does not commit the thesis to any particular normative theory. It specifies a cognitive role—detecting, evaluating, acting—rather than a moral content. Second, the definition locates moral behaviour within an information-processing architecture rather than within a system of rules. This orientation is what allows the empirical work of the thesis to proceed. To understand how synthetic presence perturbs moral action, we must describe the mechanisms that convert perceptual and affective input into evaluative output.

The remainder of this chapter develops precisely that architecture. Having surveyed the diversity of moral theories and clarified the sense in which the thesis is concerned with *moral cognition*, we can now turn to the cognitive–affective processes that make moral evaluation behaviourally operative.It is within this machinery—not within ethical doctrines—that the phenomenon of synthetic moral perturbation will be found.

This definition is intentionally modest. It avoids entanglement with substantive normative theories while isolating the components necessary for empirical investigation: evaluation, judgment, and action. It aligns with 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 scaffolding developed across the thesis: evaluative topology, Levels of Abstraction [25, 26], and the perturbative role of synthetic presence as documented in HRI and SSP [34, 74, 42].

Under this definition, moral cognition functions as a mapping from situational cues to action policies, shaped by dispositional traits [131, 132] and by the attentional and affective structures of the evaluative field [6, 94]. It supplies the minimal conceptual anchor required to analyse how synthetic presence modulates the transformation from moral perception to moral action.

Before turning to the distinction between factual and normative judgments, it is worth making explicit what has been achieved in the preceding sections. Although the discussion has been conceptual, its role has been scientific rather than merely preparatory. Three contributions are central.

The clarificatory work carried out so far serves three related functions. First, it locates the phenomenon under investigation at the correct Level of Abstraction. The literature review shows that the effects of synthetic presence occur within the cognitive–affective processes that precede deliberation; identifying this level is not a matter of presentation but a substantive result. Only by situating the phenomenon upstream of explicit reasoning can the experimental attenuation be interpreted without normative distortion or speculative inference.

Second, clarifying the distinction between descriptive and normative domains eliminates a set of category errors that routinely distort empirical interpretation. The aim is not conceptual tidiness but methodological discipline. Without this distinction, prescriptive content is too easily imported into cognitive models, and descriptive regularities too easily mistaken for ethical conclusions. Avoiding these forms of conflation is a necessary condition for generating reliable explanatory claims.

Third, the chapter establishes a minimal and action-guiding definition of moral cognition. By specifying the evaluative process that links situational cues to action selection, it identifies the mechanisms that may legitimately be invoked—salience, affect, attention, social meaning—and excludes those belonging to the wrong Level of Abstraction. This definition also provides the conceptual interface through which the empirical findings connect to the evaluative–topological framework developed later in the thesis.

Taken together, these contributions supply the explanatory architecture required to interpret the experimental results correctly and to explain how synthetic presence perturbs the evaluative field from which moral behaviour emerges.

Collectively, these achievements secure the conceptual foundations of the project. They define the explanandum, delimit the operative explanatory layer, and prevent Level-of-Abstraction conflation [25, 26]. Only once this groundwork is established can the thesis introduce finer distinctions—such as that between factual and normative judgments—that refine the architecture of moral cognition at the point where synthetic perturbation takes effect [61, 15].

This is why the next section follows naturally. To understand how synthetic presence modulates moral behaviour, we must first understand *which kind of judgment is being modulated*. Synthetic presence does not alter factual beliefs; it alters the evaluative force through which normative appraisals acquire behavioural significance. Distinguishing factual from normative judgment is therefore not an ornamental philosophical exercise: it is the next analytic step in identifying the mechanism through which the evaluative field is reshaped [17, 16, 54].

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 has a long and influential history in moral philosophy, yet it is frequently blurred in computational and psychological treatments of morality [133, 134]. Its importance in the present context lies in the fact that:

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

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

Factual judgments derive their correctness from empirical features of the world; their truth depends on observation or inference. Normative judgments embed reasons for action—they carry prescriptive force even when tacitly represented [135, 102]. This is more than a semantic contrast. It marks a functional division within the cognitive architecture: judgments about what is engage classificatory and predictive systems, whereas judgments about what ought to be done recruit mechanisms that assign motivational weight, integrate affective cues, and generate the directional force that links evaluation to action.

This division maps directly onto the psychological conception of moral cognition, understood as the ensemble of perceptual, affective, and inferential processes that register morally salient features and transform them into evaluative representations [31, 16]. Moral cognition includes explicit moral judgment as well as the upstream mechanisms that detect salience, encode social meaning, and initiate the transition from appraisal to behaviour [17, 81]. 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, 136].

This division maps directly onto our operational definition of moral cognition. Under this definition, the cognitive system performs at least two analytically distinct functions. First, it forms *factual judgments* about how the situation stands on the basis of perceptual and descriptive input. Second, it performs a further *normative transformation*: it evaluates those facts in terms of what ought to be done, producing action-guiding judgments that structure behaviour. These two processes operate at different functional levels within moral cognition and must be kept distinct for the empirical questions of this thesis to be intelligible [31, 16, 19, 17, 136].

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 evalua-

tive 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 [137, 138, 139, 140, 141]. 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 [105, 104].
3. **Practical commitment**: the formation of an action-guiding stance, in which the judgment functions as a reason for or against a particular behaviour [101, 102].

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, 142, 143, 144]. Conversely, deliberative processes involve explicit reasoning, counterfactual consideration, and appeals to principles or character traits [92]. 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 (more on this point will follow in Chapter 4) whereas others disrupt or attenuate it, we must examine the architecture through which moral salience is perceived, represented, and acted upon. The transition from perception to appraisal, and from appraisal to action, is mediated by identifiable affective, perceptual, and executive systems, each contributing distinct computational roles within the broader evaluative ecology.

As the next section shows, contemporary psychological and neuroscientific research converges on a model of moral cognition as a distributed, dynamically

interactive network. This framework clarifies how humans ordinarily navigate morally charged environments and provides the conceptual foundation for understanding how these processes may be perturbed—subtly yet systematically—by the presence of agents whose social and ontological status is ambiguous, such as humanoid robots. In this sense, the empirical foundations surveyed below serve as the substrate upon which the subsequent experimental analysis is built.

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

The experiment in Chapter 5 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 [6, 5, 1, 50, 35, 34, 57].

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

Before moving further into technical terrain, it is worth pausing to recall the shape of the question guiding this chapter. If moral cognition is the process by which agents register what matters in a situation and translate it into action, then any adequate account must eventually touch the mechanisms that realise this process in the mind and brain. The philosophical distinctions just developed prepare the ground; what follows identifies the cognitive architecture that makes those distinctions behaviourally meaningful.

A substantial body of cognitive neuroscience demonstrates that moral decision-making does not arise from a single "moral centre." Instead, it emerges from coordinated activity across affective, social-cognitive, and executive networks. These systems jointly determine how agents detect morally salient cues, generate evaluative appraisals, and select behaviour. The architecture is therefore inherently practical: the neural substrates implicated in moral judgment are also those responsible for valuation, behavioural control, and action selection.[1] Contemporary research thus situates moral judgment within a distributed computational system whose governing question is not "What is right?" but "What should I do here?" [145, 19, 142].

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 [145]. 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 [146, 147, 148]. The amygdala detects threat, intentional aggression, and aversive outcomes [149, 70], while the anterior insula responds to disgust, norm violations, and aversive interoception [150, 151, 152]. Electrophysiological studies indicate that these affective signals often precede conscious deliberation [153, 154], functioning as rapid gating mechanisms for downstream moral appraisal.

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

Social-Cognitive and Interpretive Systems. Moral judgments frequently hinge on beliefs, intentions, and reasons [155, 156]. The temporo-parietal junction (TPJ), medial prefrontal cortex (mPFC), and posterior superior temporal sulcus (pSTS) form a network specialised for mental-state attribution [157, 158, 159, 160]. TPJ activation, for example, is reliably observed when distinguishing intentional from accidental harms or attributing blame or forgiveness [161, 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 [162, 163]. Classic moral dilemmas recruit ACC activity when intuitive emotional responses and reflective considerations collide [10, 164]. This conflict-monitoring function indicates that moral cognition involves arbitration among multiple evaluative forces [165, 166].

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 [167, 168]. Disruption of dlPFC activity via TMS alters willingness to endorse instrumental harm [169, 170], demonstrating that this region contributes to structuring action policies that integrate affective, deontic, and goal-directed considerations [136, 171].

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 [172, 173, 174]. Recent accounts describe this network as computing action-guiding commitments rather than abstract evaluations [175, 176].

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 (Chapter 5) 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.

3.3.2 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, 177, 178]. This integration is fundamentally practical. The vmPFC and OFC compute the affective value of potential actions [179, 180]; the TPJ and mPFC generate intention-sensitive interpretations of agents' behaviour [158, 161]; the ACC detects conflict between competing behavioural tendencies [162, 163]; and the dlPFC regulates whether intuitive impulses should be suppressed, enacted, or balanced against normative constraints [167, 169]. 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 [70, 152].

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 [142, 181]. This interpretation aligns with philosophical accounts emphasising the intrinsically action-directed nature of moral evaluation [101, 102], 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 chapter establishes the conceptual foundations on which the remainder of the thesis depends. The experimental analysis that follows examines how minimal synthetic presence alters the evaluative conditions under which moral behaviour is formed, while the theoretical chapters (chapters 6, 7, 8, 9) articulate the structural consequences of this finding for our understanding of moral agency, normative theory, and the design of artificial systems. What unifies these strands is the recognition that moral action cannot be understood in abstraction from the cognitive–affective and social scaffolds that make it possible.

Seen through this lens, artificial agents are neither moral subjects nor passive tools. They function as *operators on the evaluative field*: entities capable of shifting salience, displacing empathic resonance, and modulating the pathways through which moral meaning becomes behaviourally operative. The full significance of this perspective emerges only when the empirical and philosophical analyses are brought into alignment, but the core insight is already visible.

Understanding how moral behaviour arises under conditions of social and ontological ambiguity is not ancillary to the thesis; it is the *conceptual linchpin* that renders the central research question intelligible. Only with this architecture in place can the influence of synthetic presence be explained with the clarity and precision the phenomenon requires.

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. A LoA, in Floridi's sense, fixes the informational distinctions that matter for explanation. Here, our LoA does not concern the metaphysics of moral agency nor the justification of principles, but the *functional transformation* of perceptual and affective cues into action-guiding evaluation. At this LoA, robots are not modelled as moral agents but as *modulators of the evaluative field.*[2]

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

This topological perspective also helps to clarify why synthetic agents may exert moral relevance even when they remain behaviourally minimal [37, 182]. At the operative LoA adopted here, the salient property of a robot is its potential to *modulate attentional and affective gradients* that guide human appraisal [183, 184]. In this capacity, a robot can act as a normative deflector or semantic attractor, subtly reshaping the pathways through which moral salience gains behavioural traction [185, 186]. Chapter 5 examines these possible redistributions and illustrates how they invite a shift from locating moral significance solely in the artificial agent to considering the *perturbations its presence may induce* [56, 57].

Seen through this joint lens of LoA and evaluative 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?*

To state this more precisely, it is helpful to introduce a minimal formalism:

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

The symbols here name the components already implicit in the discussion:

- Σ is the *evaluative input space*: the perceptual, affective, and contextual cues available to the agent in a situation.
- $f : \Sigma \to \Delta$ is the *baseline evaluative mapping*: the cognitive–affective process by which those cues are transformed into downstream moral responses (for example, a decision to donate).
- $\mathcal{P}_{\mathscr{R}} : \Sigma \to \Sigma'$ represents the *perturbation induced by the robot's presence*: it deforms the evaluative field by shifting which cues are salient and how they are weighted.
- $f_{\mathscr{R}} = f \circ \mathcal{P}_{\mathscr{R}}$ is the *robot-conditioned evaluative mapping*: the overall transformation from cues to action once the field has been perturbed by synthetic presence.

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

Nothing stronger is claimed. This formalism serves as a conceptual anchor for the central thesis: the humanoid robot does not introduce new moral content or explicit reasons; it acts as a *perturbation operator* on the evaluative field. The experiment asks whether $\mathcal{P}_{\mathscr{R}}$ is empirically detectable—that is, whether the silent presence of the robot measurably alters the mapping from moral perception to prosocial action.

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.

3.5 Concluding Perspective: Why This Matters for the Thesis

The preceding analysis points toward a common insight: **robots may reshape the evaluative topology of moral life**, not by reasoning or instruction, but by modulating the perceptual–social gradients through which moral meaning acquires behavioural relevance.

Chapter 5 examines this possibility empirically, and the normative chapters explore how it intersects with, and in some cases complicates, the assumptions that underlie classical Machine Ethics. Taken together, these strands suggest a technomoral thesis: as artificial agents become embedded in human environments, their presence may contribute—often quietly and without intention—to shifts in the *topology of moral experience*. This is the sense in which synthetic presence matters, and it is why the conceptual groundwork developed in this chapter is needed for what follows.

Although the claim that artificial agents may influence the *topology of moral experience* is illustrated here through embodied robots, its scope extends to the contemporary landscape shaped by large language models. As the earlier discussion of LLMs and the "post–Machine Ethics" era indicates, modern AI systems differ markedly from the rule-based architectures that motivated earlier theoretical frameworks. They operate through statistical patterning, implicit social modelling, and affectively charged conversational dynamics; in doing so, they can recalibrate attention, shape expectations, and influence interpretive stance. In this broader context, the perturbational role attributed to synthetic presence becomes a general feature of artificial systems that participate in human moral environments.

In this sense, even without bodies, **LLMs can be understood as potential perturbation operators on the evaluative field**. What differs is the channel through which the perturbation arises. Robots tend to influence *perceptual* and *embodied* salience, whereas LLMs act through *semantic*, *discursive*, and *interpersonal* forms of salience. Both may interact with the intuitive layer of moral

cognition—the layer that precedes explicit deliberation and structures the evaluative background in which reasons and principles acquire behavioural significance.

Viewed from this perspective, the technomoral thesis is not confined to robotics. It is a broader claim about how artificial systems—whether embodied or disembodied—may reconfigure the cognitive–affective conditions under which human moral judgement unfolds. The task of this chapter is to render this conceptual shift explicit. Without a clear account of moral cognition as an *action-guiding*, *field-sensitive*, and *LoA-dependent* architecture, discussions of LLM "moral competence" or "machine virtue" risk losing contact with the processes that actually shape moral behaviour.

Classical Machine Ethics tended to locate the moral significance of artificial systems in the principles encoded within them. The present analysis suggests that a more informative focus lies in the *perturbations such systems may induce in us*.

In this light, the technomoral thesis challenges Machine Ethics not because LLMs resolve the traditional problems of rule encoding, but because they indicate that those problems may be peripheral to the dynamics that matter. 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 is therefore foundational. It assembles the cognitive, psychological, and philosophical resources required to motivate the reframing that guides the thesis. The distinctions developed here—between descriptive and normative domains, between factual and moral judgement, between intuitive and deliberative processing, and the Levels of Abstraction that organise them—offer the conceptual discipline needed to interpret the experiment without drift or overreach. Taken together, these distinctions suggest that synthetic systems matter not as bearers of reasons or values, but as *environmental operators*: entities whose presence may modulate the patterns of salience, affect, and accountability through which moral cognition gains behavioural traction.

With these boundaries in place, the central phenomenon of the thesis becomes clearer. The attenuations in prosocial behaviour which we will observe under robotic co-presence in the experimental setting in Chapter 5 is not readily understood as a shift in belief or principle-application, but as a possible deformation of the evaluative field that precedes both.

What follows therefore turns to this cognitive–affective substrate—the level at which synthetic perturbation is most plausibly understood to take effect.

4. From Theory to Measurement: Operationalising Dispositions and Moral Perturbation

To measure moral cognition is to confront a familiar difficulty: the processes that matter most for guiding action—those early shifts in attention, affect, and evaluative weight—rarely announce themselves. They move quietly, shaping what feels salient long before any explicit judgment is formed. An experiment that aims to detect perturbations in this machinery must therefore proceed with more than methodological care; it must choose instruments that open the right conceptual windows onto a process that is not directly observable.

The task of this chapter is to make that access possible. The measures introduced below are not selected for convenience, nor for their prevalence in psychological-research, but because each provides a principled entry-point into the evaluative architecture described earlier. They allow upstream processes to leave traces that can be analysed without distorting the phenomenon itself. What follows, then, is not a catalogue of questionnaires, but the construction of the measurement interface through which evaluative perturbation becomes empirically visible.

Moral appraisal does not enter empirical analysis directly [31, 187, 188, 189, 190]. What becomes observable are the structured traces it leaves—dispositions, affective responses, attentional shifts, behavioural choices—from which the underlying evaluative process can be inferred [191, 192, 170, 1, 2, 193]. Within the evaluative–topological framework developed earlier, measurement is therefore not passive registration but a *mode of access* to structure. The evaluative topology designates the organisation of the evaluative field: the structured set of task-relevant moral appraisals, affective orientations, social expectations, and contextual cues that jointly determine how an agent interprets a situation as calling for one course of action rather than another. It is a "field" in the technical sense: a dynamically organised configuration of influences—cognitive, affective, and social—that exerts graded pressures on judgement and behaviour, such that changes to any one component (e.g., the presence of a humanoid robot) can shift the overall pattern of moral response. To measure moral perturbation is to measure deformation within this topology.

The task of this chapter is therefore not to catalogue instruments, but to justify how specific psychometric measures and perturbational manipulations—most centrally, the silent co-presence of a humanoid robot—serve as theoretically motivated probes of that topology. Each tool samples a different dimension of the evaluative field; taken together, they provide an empirical framework capable of detecting whether NAO's presence modulates the pathways through which moral salience is translated into moral action.

These instruments function not as neutral devices but as *conceptually disciplined interventions*. Each targets a specific dimension of the evaluative struc-

ture represented by the mapping introduced in Section 4.1.1, which formalises how empathic–affective variation, trait-level dispositions, and synthetic presence contribute to the generation of moral behaviour. The formalism does not claim mathematical precision beyond this interpretive role; it provides a controlled vocabulary for expressing how different components of the evaluative field become experimentally tractable without collapsing their complexity into reductive summary scores. Their epistemic role is analogous to the instruments of physics or the conceptual scaffolds of analytic philosophy: they do not merely record a value but *constitute the mode of access* through which the phenomenon becomes empirically available. The analogy is exact. In physics, a gravitational wave or an electron is not simply "detected"; it is rendered detectable by an apparatus that fixes the relevant Level of Abstraction and the dimension of variation. Philosophical analysis performs an analogous function when specifying the conceptual lens through which inference, normativity, or agency become visible. Measurement is thus an act of disciplined construction, not passive reception.

The same principle governs the tools used here. Briefly, measures of empathising and systematising dispositions (EQ/SQ) do not purport to exhaust personality; they isolate axes of cognitive–affective variation known to shape sensitivity to salience and intuitive processing in moral cognition [131, 194, 195]. The Big Five Inventory (BFI) captures dispositional gradients that interact with evaluative weighting, vigilance, and behavioural inhibition [196, 197, 198]. These instruments are therefore not psychological "thermometers" but theoretically justified interventions into the *evaluative field*, each selecting the dimensions along which individual differences become experimentally meaningful.

This methodological stance, in turn, informs the design of the experimental setting following the Watching–Eye paradigm as an experinental template. The stimulus is not treated as a behavioural curiosity but as a calibrated perturbation:

> *a contextual operator on accountability salience and affective vigilance at the cognitive LoA identified earlier.*

The Watching–Eye cue functions as a controlled modification of the evaluative landscape, providing a channel through which implicit social meaning acquires behavioural force. As in physics, the measurement depends not only on the quantity under investigation but on the entire apparatus that renders the phenomenon measurable.

What follows in this chapter therefore treats the instruments not as psychological artefacts but as components of an *epistemic strategy*. Each measure is examined in light of three questions: (i) which aspect of the evaluative architecture it operationalises; (ii) what assumptions it encodes about cognitive–affective processing; and (iii) how it constrains the interpretation of the behavioural perturbations that constitute the empirical core of the thesis.

Having established the epistemic role of these instruments, we can now turn to their empirical deployment. Every measure introduced here is a way of gaining access to something that never appears on its own—the evaluative machinery by which agents register what matters. If the previous chapter showed how moral ap-

praisal is shaped by attention, affect, and social meaning, the task now is to make those structures experimentally visible. The next subsection therefore describes how dispositional architecture and controlled perturbation are operationalised in practice, and how the measurement suite becomes the means by which subtle deformations of the evaluative field reveal themselves in behaviour.

4.1 Perturbation as Measurement: The Experimental Context

The experiment developed in this thesis does not measure moral cognition by presenting participants with explicit dilemmas or by eliciting articulated judgments. Instead, it probes the evaluative architecture indirectly, through a precisely calibrated perturbation of the perceptual environment. The silent presence of a humanoid robot serves as this perturbation. Prior work in Human–Robot Interaction shows that even non-interactive robots can reshape perceived social affordances, redirect attentional gradients, and alter expectations about norm-relevant behaviour [35, 199, 200]. Their ambiguous social ontology—perceptually agentic yet behaviourally indeterminate—disrupts default social priors and reorganises the salience structure within which moral reasons become behaviourally operative.

Within this framework, robotic presence is not a background feature of the experimental setting; it is the experimental instrument. It functions as a controlled perturbation of the evaluative field, allowing the study to examine how dispositional invariants interact with contextual cues to produce measurable differences in prosocial behaviour. The Watching–Eye cue provides a second, orthogonal perturbation, modulating accountability salience and intuitive social monitoring. Together, these manipulations define the salience structure within which participants navigate moral action.

4.1.1 Purpose and Structure of this Chapter

The work undertaken in this chapter is not organisational housekeeping; it is an attempt to expose the structure that makes the experiment scientifically meaningful. If moral behaviour is the surface expression of deeper cognitive–affective dynamics, then any empirical study must first specify *which dimensions of that structure are being accessed, and by what kind of instrument.* The chapter therefore has a dual purpose, each corresponding to a different layer of the evaluative architecture.

1. First, it establishes the empirical and theoretical foundations of the psychometric and contextual tools employed in the study. These instruments are not selected for convenience: each targets a specific component of the evaluative substrate—affective resonance, systemising precision, personality curvature, or salience modulation—whose interaction determines how moral salience becomes behaviourally operative.
2. Second, to show how each tool contributes to the modelling of the dispositional term β_C in the formal expression

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

Here $\mathscr{P}(\delta_m)$ denotes the probability of a morally relevant behavioural outcome (e.g. donation) in a given trial; α_E collects the task and environmental parameters (Watching–Eye cue, payoff structure, contextual framing); β_C encodes the participant's latent cognitive–affective disposition; and γ_R indexes the presence or absence (and configuration) of the humanoid robot. The mapping f is not introduced as a fully specified quantitative model but as a structured vocabulary for talking about how environmental conditions, dispositional architecture, and synthetic presence jointly shape the evaluative field. 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 instruments employed in this thesis—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 [194, 131, 195, 196, 197, 198, 1, 2, 5]:

1. **Theoretical relevance.** Each tool targets a component of the evaluative architecture identified earlier: affective resonance, evaluative precision, personality curvature, or salience modulation[18, 87, 6, 42].
2. **Empirical robustness.** Each tool has been validated across multiple cultures, large samples, and decades of psychological research, and has been employed in studies of prosociality, moral sensitivity, social attention, and HRI [201, 202, 203, 132, 1, 2, 3, 33, 34, 74].
3. **Computational suitability.** Each tool yields variables suitable for integration into regression models, cluster analysis, and topological interpretation, allowing dispositional and contextual parameters to be related systematically to behavioural outcomes[17, 204, 136].

Given the relatively modest sample size ($N = 71$), it is natural to ask how this measurement framework remains empirically credible:

> *How does the theoretical weight of the measurement architecture reconcile with a study based on seventy-one participants, and have these tools been employed in comparable contexts?*

The answer lies in the nature of the constructs under investigation. The tools used in this thesis were not chosen because they require large samples for exploratory factor recovery, but because they index *structurally stable* psychological dimensions. Their psychometric properties—factor structure, reliability, discriminant validity—have been established in large-scale studies involving thousands of participants across diverse populations [194, 131, 195, 196, 197, 198, 201, 202, 203]. In that sense, the present study does not re-validate the instruments; it leverages constructs whose statistical scaffolding is already in place.

Crucially, the goal of the experiment is not to discover new personality factors or infer latent structure from the data. It is to examine how *known dispositional invariants* interact with a controlled perturbation of the evaluative field (see 4, page 51). Small-to-moderate sample sizes are standard in this domain: replications of the Watching–Eye effect typically employ samples between 40 and 120 participants [1, 2, 4, 5], and HRI studies on the behavioural impact of robotic

presence operate in similar ranges [35, 34, 74]. Studies integrating personality measures into moral or prosocial decision paradigms (e.g. BFI or EQ/SQ as predictors of prosociality, empathic concern, or social attention) likewise often rely on samples of comparable scale [94, 132, 200, 205].

The inferential strategy adopted here reflects this precedent. Dispositional measures are treated as low-dimensional, theoretically structured parameters that influence the deformation of evaluative gradients; the analysis does not attempt to reconstruct the topology of moral cognition from the present dataset alone. The statistical burden falls on detecting systematic, directional shifts in behaviour induced by the experimental manipulations, not on extracting high-dimensional latent structure from sparse data. For that purpose, a well-powered design requires a clean manipulation, validated constructs, and an analysis aligned with the theoretical architecture [17, 204, 136] rather than sheer sample size.

In short, the experiment does not aim to estimate the entire evaluative topology *de novo*. It examines how a synthetic agent perturbs an already well-characterised structure. The sample size is calibrated not to psychometric exploration but to experimental contrast:

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

Evidence from HRI, SSP, and moral psychology indicates that perturbations of this magnitude are robustly detectable with samples of the size employed here [91, 42, 6, 74].

With this clarification in place, we can now turn to the instruments themselves. Their role in the thesis is not a matter of psychometric convenience, but of *theoretical access*: they expose the latent structures through which evaluative salience is processed and transformed—structures that, as the experiment will show, can be subtly but measurably deformed by the presence of a synthetic agent.

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

In the framework developed thus far, moral behaviour is modelled as the endpoint of a trajectory across an evaluative field. Contemporary work in moral psychology and cognitive science converges on the view that such trajectories are shaped by the coordinated influence of environmental cues, dispositional structure, and perturbational forces[9, 187, 10, 11, 12, 206]. This relationship was formalised in the previous section through the schematic decomposition

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

whose components were introduced and motivated earlier. What matters here is how psychometric tools allow the dispositional term β_C to become empirically visible.

Within this architecture, the role of psychometrics is not descriptive but structural. The Empathizing Quotient (EQ), Systemizing Quotient (SQ), and Big

Five Inventory (BFI) operationalise latent properties of the evaluative field: affective bandwidth, analytic structuring tendencies, and broad personality curvature. Each instrument isolates a dimension along which individuals differ in how they register, weight, and transform morally salient information. These constructs do not exhaust the space of possible dispositions, but they provide *theoretically grounded coordinates* for the manifold through which evaluative trajectories flow [7, 207, 208, 198].

The methodological necessity of these tools becomes clear once we consider the counterfactual. Without a dispositional coordinate system, heterogeneity among participants would remain unmodelled, and any perturbation effect could be misattributed to uncontrolled trait variance. As it will become clearer in I Chapter 5, the cluster structure recovered from EQ, SQ, and BFI scores demonstrates that this heterogeneity is not noise but a *non-trivial topology*: profiles centred on empathic warmth, analytical structuring, and emotional reactivity emerge as distinct basins in the dispositional landscape. Participants did not enter the experiment as psychologically interchangeable agents.

What follows is the critical empirical observation. Despite this internal structure, the humanoid robot produced a *uniform directional attenuation* of prosocial behaviour across all dispositional profiles. No component score, no trait factor, and no latent cluster moderated the effect. This aligns with findings in Human–Robot Interaction showing that passive robots can globally reshape social affordances and attentional dynamics[35, 199, 200]. The present study **extends that literature** by demonstrating that the robot acts not on trait-linked pathways, but on the evaluative field itself: shifting salience gradients, dampening affective trajectories, and deforming the topology through which all dispositions acquire behavioural expression.

The psychometric instruments are indispensable for isolating this phenomenon. They allow the analysis to separate the *shape of the dispositional manifold* from the *geometry of the perturbation.* In the formalism introduced and explained earlier (page 53), the contrast of interest can be expressed as

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

a comparison that holds the dispositional coordinates β_C fixed while introducing the perturbation operator γ_R.

The function f represents a *structural mapping* introduced earlier to express how observable moral behaviour depends on three classes of determinants. Given a fixed environmental configuration α_E (e.g., the Watching–Eye cue and task context) and a stable dispositional profile β_C, the expression $f(\alpha_E, \beta_C)$ denotes the pattern of moral behaviour expected *in the absence* of synthetic presence. The term $f(\alpha_E, \beta_C, \gamma_R)$ represents the corresponding pattern *when the robotic perturbation is active.* The difference

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

therefore isolates the contribution of the perturbation operator γ_R while holding both environmental cues and dispositional structure fixed. It is a formal way

of expressing a counterfactual comparison: how the same evaluative architecture behaves with and without the synthetic agent.

The empirical result in Chapter 5 is unambiguous: although β_C exhibits a structured internal topology, the robot's presence induces a deformation that is effectively global. This is the behavioural signature of a field-level operator, not a trait-contingent stimulus.

The purpose of this section is to clarify the epistemic function of the tools we use. Psychometric tools make visible the latent evaluative substrate; the experiment reveals how synthetic presence acts upon that substrate. Only by mapping dispositional structure can we show that the robot's influence bypasses trait-specific channels and instead targets the conditions under which moral trajectories unfold.

With this distinction in place—between the manifold of dispositions and the global perturbation imposed upon it—we can now examine the individual tools in their own right. Each measure is introduced in light of three questions:

(i) which aspect of the evaluative architecture it operationalises; (ii) what assumptions it encodes about cognitive–affective processing; and (iii) how it constrains the interpretation of the behavioural perturbations that form the empirical core of the thesis.

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

The measurement strategy of this thesis cannot be grounded in convenience or disciplinary habit. Once moral behaviour is modelled as the endpoint of a trajectory shaped by environmental cues, dispositional architecture, and perturbational forces, the empirical task becomes clear: the instruments must allow these components to be *distinguished in practice*. In particular, the methods must permit a clean separation between variation arising from stable cognitive–affective dispositions and variation induced by a deformation of the evaluative field.

The psychometric and experimental tools selected here were chosen precisely because they satisfy this requirement. Each instrument targets a theoretically motivated dimension of the evaluative topology and has a well-established empirical profile that makes its role interpretable within the framework developed in Chapters 2 and 3. The criteria below articulate the methodological reasons for their inclusion *prior* to any empirical result.

(1) Cross-paradigmatic relevance. The EQ, SQ, BFI, and Watching–Eye paradigm each derive from empirical traditions spanning moral psychology, social cognition, personality research, and Human–Robot Interaction. Across these literatures, they have been used to study prosociality, empathic concern, harm aversion, cognitive style, and the integration of affective and deliberative processes in moral appraisal [9, 187, 10, 11, 12, 206]. The Big Five Inventory operationalises broad personality architecture with robust behavioural predictive validity [208, 198, 209]. The Empathizing and Systemizing Quotients provide validated assessments of affective resonance and analytic curvature [7, 207]. The Watching–Eye paradigm is among the most reliable manipulations of prosocial

salience, repeatedly demonstrating that minimal cues of observation modulate cooperative behaviour [1, 2, 4, 6, 5]. These convergences align the present study with a mature empirical landscape while remaining faithful to the evaluative–topological framework.

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

- **EQ**: affective attractors shaping early intuitive appraisal;
- **SQ**: analytic curvature influencing interpretive structure;
- **BFI**: personality geometry modulating salience, attention, and regulatory control;
- **Watching–Eye**: a validated perturbation of accountability salience operating at the cognitive LoA.

Together, these measures provide the granularity required to instantiate the dispositional term β_C (see page 53 for an introduction to the terms) in the expression

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

and thereby allow dispositional variation to be distinguished from field-level perturbation. This is central to the empirical question: whether robotic presence acts *on traits* or *on the evaluative field* within which traits express themselves.

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

- **Stability**: each has extensive psychometric validation across cultures and populations;
- **Analytical tractability**: each yields variables suitable for clustering, regression, and topological comparison;
- **Interpretability**: each maps onto established accounts in moral psychology, enabling behavioural effects to be related back to theoretical structure.

Most importantly, these tools provide the precision required to dissociate *who participants are* from the *evaluative conditions* under which they act. As the analysis will show, the psychometric suite revealed a structured dispositional landscape, yet robotic presence attenuated prosocial behaviour *irrespective* of that structure. The instruments therefore make visible the distinction between dispositional architecture and field-level deformation—an interpretive separation that would be impossible without them.

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

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

The Empathizing Quotient (EQ) provides a validated measure of affective resonance—an individual's capacity to detect, register, and respond to the emotional and psychological states of others [7]. Originally developed within the

Empathizing–Systemizing framework [210, 211], the EQ captures both emotional reactivity and cognitive perspective-taking, two mechanisms repeatedly shown to influence prosocial behaviour, harm aversion, and sensitivity to moral salience [9, 187, 11, 12].

Why EQ Matters Conceptually

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

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

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

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

Historical and Psychometric Grounding

Empirically, the EQ has a robust record: strong internal reliability, stable factor structure across cultures [201], convergence with related constructs (empathic concern, emotional intelligence), and predictive validity for prosociality in behavioural economic tasks. Neurocognitive studies further show correlations between EQ scores and activation in vmPFC, anterior insula, and TPJ—regions central to affective resonance and mental-state attribution [19, 18] (refer to section 3.3.1, page 44 for a discussion on brain regions).

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

4.4.1 EQ and Synthetic Presence

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

The data did not support this possibility. EQ did *not* moderate the displacement effect: both higher- and lower-empathy participants tended to show reduced prosocial donation in the robot condition. This pattern suggests the following interpretive reading:

"The robot may have influenced the evaluative field itself rather than trait-dependent gradients within it."

This interpretation is consistent with findings in HRI indicating that robotic presence can modulate attentional and social–evaluative processing in ways that do not strongly depend on empathic predisposition [199, 200].

4.4.2 Methodological Role in the Thesis

The EQ served two indispensable methodological purposes:

1. **Controlling for affective heterogeneity**. Without a measure of empathic sensitivity, reductions in donation could have been attributed to unmeasured differences in participants' empathy levels. The EQ rules out this confound.
2. **Modelling the affective dimension of β_C**. EQ provides the affective coordinate of the dispositional manifold, enabling cluster analysis and regression models to distinguish dispositional shape from field-level perturbation.

Thus, even though affective resonance plays an important role in moral cognition, the experimental pattern suggests that the perturbation introduced by the humanoid robot operated *upstream* of empathy—modulating aspects of the evaluative topology rather than amplifying or dampening empathic traits.

Having considered the affective dimension of β_C, we now turn to its analytical counterpart: the Systemizing Quotient.

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

Where the Empathizing Quotient (EQ) indexes affective resonance, the Systemizing Quotient (SQ) [212, 202, 201] quantifies a cognitive style characterised by rule extraction, structural analysis, and the search for causal regularities. Within the evaluative–topological model developed in this thesis, the SQ corresponds to the *analytical curvature* of the evaluative field: the extent to which agents encode situations via stable structural relations rather than affective gradients.

4.5.1 Theoretical Background and Psychometric Foundations

The SQ emerged from the Empathizing–Systemizing framework [210, 211], originally designed to capture the dissociability of affective versus rule-based processing in autism research. Subsequent work broadened this motivation: systemizing is now associated with mechanistic reasoning, pattern extraction, predictive modelling, and a preference for low-noise, high-coherence causal schemas [202]. Psychometric studies demonstrate high internal reliability, cross-cultural robustness, and predictable correlations with analytic problem-solving and rule-consistent behaviour.

Neurocognitively, higher SQ scores have been associated with increased lateral prefrontal and parietal activation during analytic reasoning [213, 214, 215, 216], and with reduced activation in affective salience networks during certain social tasks [8]. Taken together, these findings are consistent with interpreting SQ as indexing a cognitive style that places greater weight on structural stability than on affective modulation.

4.5.2 SQ Across Moral Psychology and HRI

In moral psychology, systemizing predicts greater reliance on deliberative processing, reduced affective interference, and increased endorsement of principle-based judgments in high-conflict dilemmas [54, 11, 131, 217, 218]. In behavioural economics, high-SQ individuals show more consistent strategic patterns and reduced susceptibility to framing effects [219, 220, 221].

In Human–Robot Interaction, systemizing tendencies shape expectations about synthetic agents: high-SQ participants tend to interpret robots through structural and functional cues rather than anthropomorphic ones and attribute competence and reliability more readily than emotional or social qualities [35, 199, 200, 222, 223]. This makes the SQ especially relevant in the present experiment, where the perturbation introduced by the robot is primarily structural rather than affective.

4.5.3 Systemizing Quotient (SQ): Structural Bias and Evaluative Rigidity

Within the evaluative–topological framework developed in Chapter 3, each dispositional measure contributes a distinct dimension to the latent configuration β_C. Whereas the Empathizing Quotient (EQ) captures the affective attractors that steepen or flatten moral salience, the Systemizing Quotient (SQ) indexes a different property of the evaluative field: the degree to which agents privilege structural invariants, analytic coherence, and rule-based regularities when interpreting a situation.

In the functional decomposition

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

SQ enters as a shaping parameter of β_C: it modulates the *rigidity*, *smoothness*, and *resistance to perturbation* of evaluative trajectories. High-SQ agents tend to encode situations through stable relational patterns; their evaluative landscape is less susceptible to low-level shifts in affect or fleeting cues of social meaning. Conversely, lower SQ reflects a more flexible, affect-sensitive terrain in which subtle perturbations may redirect the trajectory from appraisal to action.

This topological intuition can be expressed heuristically by treating the evaluative potential $V(x)$ as a surface whose curvature reflects the systemizing tendency:

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

The Laplacian here is not introduced as a mechanistic claim about neural computation but as a conceptual device: higher curvature corresponds to a stiffer evaluative surface—one whose gradients change slowly and whose trajectories are less easily diverted by perturbational forces. Lower curvature marks a softer topology, where local salience cues exert proportionally greater influence on evaluative flow.

In this way, SQ provides one of the coordinate axes through which the dispositional manifold becomes empirically visible. Its contribution to β_C is not descriptive ornamentation but a structural constraint on how agents weigh context,

interpret relational structure, and integrate perturbational input. As the later cluster analysis shows, variation in SQ participates in shaping the dispositional geometry; but the perturbation induced by robotic presence operates at a field level that exceeds these trait-contingent contours—a dissociation that the formalism above is designed to make intelligible.

4.5.4 SQ, Synthetic Presence, and Field-Level Perturbation

The experiment in Chapter 5 indicated that SQ did *not* moderate the behavioural attenuation observed in the presence of the humanoid robot. Participants with higher SQ—who might be expected to rely more heavily on rule-based evaluative strategies—showed a similar directional reduction in prosocial behaviour to both higher- and lower-empathy participants.

This pattern is conceptually informative. It suggests that the influence of robotic presence operated not on participants' cognitive styles but on aspects of the *evaluative field itself*. Systemizing tendencies did not appear to buffer, amplify, or redirect the behavioural effect.

The perturbation introduced by the robot appears to have been global rather than trait-specific.

This aligns with existing HRI work showing that ambiguous synthetic agents alter social affordances and attentional dynamics independently of analytic or empathic predispositions [35, 199].

4.5.5 Methodological Significance

SQ served two methodological functions within the experiment:

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

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

4.6 The Big Five Inventory (BFI): Personality Geometry Within the Evaluative Topology

The Big Five Inventory (BFI) is one of the most extensively validated instruments in differential psychology [196, 197, 198]. Decades of lexical, psychometric, and cross-cultural research have shown that Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism constitute a robust low-dimensional structure for describing stable individual differences [224, 225]. Within the evaluative–topological framework developed in Chapter 3, these traits supply a principled

coordinate system for locating each participant's contribution to the dispositional term β_C in the functional decomposition

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

where α_E captures environmental salience, β_C captures stable evaluative tendencies, and γ_R denotes the perturbation introduced by the humanoid robot.

4.6.1 Why Personality Matters for a Topological Account of Moral Cognition

Personality traits shape how situations are perceived, which cues are weighted, and how affective signals enter downstream judgement [226, 227, 228, 229]. They influence the steepness of prosocial attractors, the volatility of evaluative trajectories, and the degree to which social presence modulates behaviour. For a thesis concerned with *perturbations of the evaluative field* (see page 51), the BFI provides the structural background against which any deformation must be interpreted. Without a model of trait geometry, one could not determine whether the attenuation induced by robotic presence reflected genuine field-level displacement or merely the aggregation of heterogeneous personality effects.

4.6.2 Psychometric Stability and Cross-Domain Predictive Value

The BFI exhibits strong internal reliability, stable factor structure across cultures, and predictive value across a wide range of behavioural domains: cooperation, social engagement, norm adherence, and affective regulation [209, 230, 132]. Short forms such as the BFI-10 preserve this structure while remaining suitable for time-constrained experimental settings [198]. These qualities make the BFI particularly well suited to modelling β_C within an evaluative–topological framework that requires dispositions to act as stable constraints on the flow of salience and affect.

4.6.3 Personality and Moral Behaviour Under Social Presence

Each Big Five trait has a theoretically grounded role in shaping evaluative processing:

- **Agreeableness** steepens prosocial basins and increases sensitivity to interpersonal harm.
- **Conscientiousness** stabilises evaluative gradients and supports rule-consistent trajectories.
- **Neuroticism** introduces volatility and amplifies contextual reactivity.
- **Extraversion** enhances responsiveness to social presence and implicit monitoring cues.
- **Openness** broadens contextual sampling and moderates tolerance for ambiguity.

HRI studies show that these dimensions influence how artificial agents are perceived—as social partners, observers, or normatively relevant entities [35, 41]. The BFI therefore enables the experiment to test whether personality-dependent interpretations of robotic presence contribute to prosocial modulation or whether the effect arises at a level independent of personality variation.

4.6.4 Personality Geometry in the Evaluative–Topological Model

Within the topological formalism, the BFI quantifies the geometry of β_C: the curvature, stability, and directionality of each participant's evaluative field. Traits can be interpreted as shaping particular geometric properties:

· Agreeableness deepens altruistic attractors.
· Conscientiousness smooths and stabilises the evaluative surface.
· Neuroticism introduces local turbulence and heightened gradient sensitivity.
· Extraversion strengthens channels responsive to social cues.
· Openness expands the contextual manifold explored during appraisal.

These interpretations allow personality structure to be integrated without reducing behaviour to trait levels: traits constrain trajectories but do not determine them.

4.6.5 Cluster Analysis: Mapping the Dispositional Manifold

The cluster analysis reported in Chapter 5 revealed three recurrent dispositional configurations:

1. **Prosocial–Empathic**: high Agreeableness combined with high EQ, producing steep affective attractors for interpersonal concern.
2. **Emotionally Reactive**: high Neuroticism and unstable evaluative curvature, yielding heightened sensitivity to contextual fluctuation.
3. **Analytical–Structured**: high Conscientiousness together with high SQ, generating rigid, rule-oriented evaluative pathways.

These profiles show that participants did not enter the experiment as psychologically interchangeable agents. The dispositional manifold exhibited *structured heterogeneity*: distinct attractor regions with characteristic evaluative tendencies. In the evaluative–topological framework, the BFI and related measures provide the coordinate system through which this manifold becomes empirically visible. Without such structure, the subsequent perturbation analysis would lack an anchor for distinguishing individual variation from field-level dynamics.

4.6.6 The Key Empirical Result: A Uniform Field-Level Displacement

Against this backdrop of meaningful dispositional diversity, the experiment produced a striking and theoretically informative result:

> *The humanoid robot induced a uniform attenuation of prosocial behaviour across all clusters.*

No Big Five dimension—and no latent cluster—moderated the effect.

This pattern is not merely a negative finding; it is the empirical signature of a *field-level perturbation*. Had the effect been trait-contingent, clusters would have separated in their behavioural response. Instead, the attenuation aligned across the entire dispositional manifold. The perturbation introduced by the robot acted not on personality-specific pathways, but on the evaluative topology

itself: reshaping salience gradients, dampening affective pull, and shifting the conditions under which moral appraisal becomes behaviourally operative.

In this sense, the psychometric tools were indispensable. They enabled a clean dissociation between:

· the *shape of the dispositional manifold* (β_C), and
· the *geometry of the perturbation* (γ_R).

Only with a well-characterised β_C could the experiment demonstrate that the robot's influence was globally oriented—an operator on the evaluative field, not a stimulus whose meaning depended on personality variation.

With the role of personality geometry clarified, we now turn to the final psychometric component of β_C: the Systemizing Quotient.

4.6.7 Methodological Significance

The BFI provides the evidential basis for distinguishing between:

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

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

The analysis of BFI scores brings the dispositional component of the model into clear view:

> *The personality manifold is structured, yet the attenuation induced by robotic presence does not track personality. It deforms the evaluative field globally.*

The BFI delivers three results essential for the evaluative–topological framework:

1. It confirms that participants exhibit theoretically meaningful dispositional variation rather than psychological homogeneity.
2. It provides the coordinate basis for the cluster structure of β_C, making the dispositional manifold empirically visible.
3. It shows that the robot's influence is not trait-contingent: the observed attenuation arises from a field-level shift, not from personality-driven pathways.

These results complete the characterisation of β_C. With the geometry of dispositional structure established—and with trait-based explanations ruled out—we can turn to the design of the perturbation itself: the Watching–Eye paradigm and the silent humanoid robot that reconfigures the evaluative field.

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

Across behavioural ethics, social psychology, and field studies of cooperation, a remarkably stable finding recurs: minimal cues of being observed—stylised eyes, schematic pupils, or even simple dot configurations—increase charitable giving, norm compliance, and prosocial behaviour[1, 2, 4, 6]. No instruction, coercion,

or explicit social information is required. The effect arises because these cues selectively heighten the salience of norm-relevant action.

Within the evaluative–topological framework developed in earlier chapters, watching–eye stimuli are best understood as *topological amplifiers*: they increase the weight of prosocial directions in the evaluative field by signalling that one's behaviour carries implicit social meaning. This makes them the ideal baseline perturbation against which to test whether synthetic presence deforms evaluative trajectories in a fundamentally different way.

4.7.1 Watching–Eye Cues as Topological Amplifiers

Early explanations cast the watching–eye effect as an implicit reputational computation—a sense that one's behaviour is observable and therefore subject to social judgement[1, 2]. Contemporary models provide a more mechanistic account: the effect reflects a coordinated modulation of

- **attentional uptake**, increasing sensitivity to norm-relevant cues;
- **affective arousal**, particularly self-conscious emotions linked to evaluation;
- **interpretive expectation**, via implicit social-monitoring systems.

In the formalism already introduced, the environmental input α_E encodes contextual features that influence early evaluative appraisal. A watching–eye cue increases the prosocial component of this input. Formally, we express this as:

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

This notation does not imply a specific quantitative metric. It indicates that the cue adds a *directional* contribution to the evaluative field: prosocial gradients steepen, making cooperative trajectories more accessible in the resulting cognitive–affective dynamics.

4.7.2 Why Child-Pair Eyes Provide a Clean Baseline

Child-eye posters are widely used in prosociality research because they combine high perceptual salience with minimal interpretive content. Decades of evidence show that stylised child eyes:

- robustly increase prosocial behaviour across cultures and modalities[1, 2, 3, 95, 4];
- evoke empathic and care-oriented affective responses[231];
- heighten vigilance without implying the presence of a moral agent[6].

They therefore provide the ideal experimental baseline: a *high-salience, low-interpretation* perturbation. The cue is strong enough to elevate prosocial gradients, yet conceptually simple enough to avoid confounds involving mind attribution, intentionality, or interpersonal inference.

In the context of this thesis, the watching–eye paradigm serves a precise theoretical role. It establishes how the evaluative field responds to a well-understood salience amplifier. Against this backdrop, the perturbation introduced by the

humanoid robot can be evaluated not simply in terms of outcome, but as a deformation of the underlying evaluative topology.

4.7.3 Why Synthetic Presence Dilutes or Distorts the Effect

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

Humanoid robots are perceptually social but ontologically indeterminate. They are seen, but not reliably understood, as bearers of evaluative or moral capacities [35, 199, 200]. This ambiguity weakens all three mechanisms that normally support the watching-eye effect:

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

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

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

The experiment yields a clear result:

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

Participants normally responsive to cues of observation—including those high in Agreeableness, Extraversion, or Empathizing Quotient (EQ)—showed the same directional reduction in prosocial action when the robot was present.

To express this within the evaluative–topological formalism, recall that:

- α_E denotes the environmental input that shapes early evaluative appraisal;
- $\delta\alpha_{\text{eye}}$ represents the salience increment produced by the watching–eye cue (a positive shift in prosocial weighting);
- $\Delta_{\mathscr{R}}$ denotes the deformation of the evaluative field induced by robotic presence.

The behavioural pattern can then be summarised as:

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

This notation does not specify a numerical metric. It encodes the structural fact that the robot imposes a *countervailing field-level displacement* ($\Delta_{\mathscr{R}}$) that reduces the effective prosocial gradient introduced by the eye cue. Crucially, the displacement occurs *regardless of the participant's dispositional configuration* β_C. No Big Five trait, no EQ or SQ dimension, and no latent cluster moderated the

attenuation. This is the signature of a perturbation that acts on the *evaluative field itself*, not on specific trait-dependent pathways.

4.7.5 Why the Watching–Eye Paradigm Is Indispensable

Within this thesis, the watching–eye paradigm is not an auxiliary manipulation; it is a methodological anchor. It serves four essential functions:

- **A reliable high-salience baseline:** the eye cue consistently steepens prosocial gradients, allowing attenuation to be detected with precision.
- **Continuity with moral psychology:** the paradigm embeds the experiment within a long empirical tradition, enabling direct interpretive comparison with decades of prosociality research.
- **Isolation of perturbation effects:** attenuation is meaningful only when there is something to attenuate. The eye cue provides this necessary initial elevation of salience.
- **Revelation of topological structure:** the contrast between amplification (eye cue) and deformation (robot) allows the evaluative topology to be interrogated rather than merely described.

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

4.7.6 Integration With Costly Prosocial Action

Donation tasks provide observable moral behaviour rather than abstract moral judgement. Their integration with watching–eye cues allows the experiment to trace the evaluative sequence from:

1. **cue uptake** (the perceptual registration of observation),
2. **salience amplification** (increased weighting of prosocial trajectories), to
3. **action selection** (allocation of real resources).

The robot's attenuation of this sequence demonstrates that synthetic presence modifies the mapping from perceptual cue to moral action. In topological terms, the robot does not erase the watching–eye effect; it *counteracts* it by introducing a deformation of the evaluative field through which all trajectories must pass.

4.7.7 Synthesis: A Window Into Moral Topology

The watching–eye paradigm functions as the critical hinge of the measurement architecture. By steepening prosocial gradients, it renders the structure of the evaluative field empirically visible. By attenuating these gradients, the humanoid robot exposes the central empirical insight of the thesis:

> **Synthetic presence acts on the evaluative field itself, not on personality-dependent pathways.**

The watching–eye effect thus provides the diagnostic contrast through which the robot's field-level deformation becomes observable.

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

This chapter has not merely catalogued instruments. It has constructed the *measurement logic* that makes synthetic moral perturbation empirically legible. The Empathizing Quotient (EQ), Systemizing Quotient (SQ), Big Five Inventory (BFI), and the Watching–Eye paradigm form an integrated system of epistemic probes: each is theoretically grounded, psychometrically validated, and methodologically necessary for making the evaluative topology of moral cognition accessible without flattening it.

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

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

where α_E encodes environmental salience, β_C the dispositional manifold, and γ_R the perturbation induced by synthetic presence. Each measurement tool targets one of these components:

- **EQ** probes affective attractors within β_C.
- **SQ** probes structural curvature and analytic rigidity within β_C.
- **BFI** provides the coordinate geometry through which the topology of β_C becomes measurable.
- **Watching–Eye cues** modulate α_E by steepening prosocial gradients, enabling displacement by γ_R to be detected.

Individually, these instruments measure meaningful constructs. Collectively, they make the evaluative topology *empirically visible.*

4.8.1 Dispositional Mapping: A Structured Manifold, Not a Confound

The EQ, SQ, and BFI jointly revealed a structured dispositional manifold: Affective–Prosocial, Emotionally Reactive, and Analytical–Structured profiles. This establishes that participants entered the experiment with real, theoretically interpretable heterogeneity.

Yet the experiment demonstrated a decisive dissociation:

The humanoid robot attenuated prosocial action uniformly across all clusters.

The dispositional manifold β_C was not the locus of modulation. The perturbation acted on the evaluative field itself. Without psychometric resolution, this inference would have been impossible: the uniform attenuation could have been mistaken for trait-driven variance rather than field-level deformation.

4.8.2 Watching–Eye Cues as Diagnostic Amplifiers

The watching–eye manipulation provided the complementary half of the measurement logic. By steepening prosocial gradients in α_E, it yielded the high-salience baseline necessary to detect synthetic attenuation. The robot did not merely reduce generosity; it *neutralised a well-established amplifier of moral salience.* This

interaction is the clearest behavioural signature of a perturbation operating at the level of evaluative topology.

In theoretical terms: the eyes amplify the gradient; the robot deforms the landscape.

4.8.3 Philosophical and Ethical Meaning

Placed against the philosophical frameworks introduced earlier, the findings reveal:

- **Against rationalist models:** the perturbation bypasses deliberation.
- **Against virtue-ethical accounts:** stable dispositions do not shield agents from synthetic deformation.
- **Against sentimentalist explanations alone:** high-empathy profiles are attenuated equally.
- **Against classical Machine Ethics:** the moral significance of AI lies not in artificial agency but in field-level modulation.

Synthetic systems do not enter moral space as agents but as *operators*: they bend the evaluative field through which moral meaning becomes action.

4.8.4 Methodological Synthesis: The Tools as Epistemic Infrastructure

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

1. moral behaviour must be analysed through distinct layers of environmental input, dispositional structure, and perturbational force;
2. psychometrics provides the resolution needed to map β_C and rule out trait-based explanations;
3. observational cues provide the leverage needed to modulate α_E in a controlled manner;
4. synthetic presence must be interpreted as a deformation of *evaluative topology*, not as a trait-contingent stimulus.

The tools are therefore not auxiliary components of the study—they are the *conditions of intelligibility* for its central empirical claim.

4.9 Transition to Experimental Methods

The work of this chapter has been to give the experiment something it cannot supply for itself: an architecture within which its measurements become intelligible. We now have the conceptual scaffolding needed to understand what a perturbation of moral cognition would look like, and—equally important—what it would not. The evaluative field, the dispositions that curve it, the cues that tilt its gradients: these are no longer abstractions but the coordinates from which empirical observation becomes possible.

The next chapter turns from architecture to instrumentation. It shows how these evaluative structures were translated into stimuli, procedures, and statis-

tical models; how deformation is rendered measurable; and how dispositional parameters can be tracked as they respond to contextual change.

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

The guiding question for the remainder of the thesis can now be stated with precision:

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

What follows is the methodological counterpart to the conceptual work developed here. The instruments introduced in the next chapter are not merely data-gathering devices; they are the means by which the evaluative landscape becomes empirically visible. With these foundations in place, the thesis moves—*from structure to measurement, from possibility to test*—into the experimental core of the argument.

5. Operationalising Evaluative Topology: An Experimental Framework for Moral Perturbation

The preceding chapters developed a theoretical architecture in which moral behaviour is understood as the output of a structured evaluative field, and in which synthetic presence functions as a perturbation of that field. Here, we translate this architecture into an experimental design. What has so far been formulated conceptually must be rendered observable and open to empirical adjudication. This requires a different form of precision: every construct must be operationalised, every assumption made measurable, and every inference disciplined by explicit procedure.

The experiment is organised around a single research question:

Question 5.1: *Inferential Displacement*

Does the silent presence of a humanoid robot—perceptually social yet ontologically indeterminate—alter the evaluative process that transforms moral perception into prosocial behaviour?

This question is methodological rather than rhetorical. It identifies the causal layer at which the experiment intervenes and fixes the structure of the mapping to be tested:

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

Introduced in Chapter 4 (section 4.1.1 page 53), is the conceptual and methodological anchor of the experiment. Although the components of this expression were defined earlier, a second, compact explication is warranted here. The experiment operationalises this mapping directly, and empirical transparency requires that each symbol be reintroduced with full precision.

The mapping $\mathscr{P}(\delta_m)$. The left-hand side denotes the probability (or propensity) that a participant will produce a measurable moral action δ_m—in this study, the donation decision. It is not assumed to reflect a stable moral trait; it is the *behavioural output* of an evaluative process.

The function $f(\cdot)$. The function f represents the evaluative mechanism through which environmental cues, dispositional structure, and perturbational forces jointly shape behaviour. It encodes the cognitive–affective transformations that convert moral salience into action. No parametric assumptions are imposed; f is a structural placeholder for the evaluative architecture developed in the previous chapters.

Environmental input α_E. The term α_E denotes the morally relevant features of the environment. In this experiment, it includes:

- the Watching–Eye cue, which amplifies prosocial salience;
- the task context, instructions, and perceptual setting;
- the baseline social meaning of the environment.

Dispositional manifold β_C. The term β_C denotes the participant's latent dispositional configuration, as operationalised through the Empathizing Quotient (EQ), the Systemizing Quotient (SQ), and the Big Five Inventory (BFI). It is called a *manifold* because it is structured: a multi-dimensional geometry rather than a single trait score. The mapping f is sensitive to β_C in the sense that individuals differ in how they encode, weight, and integrate moral cues.

Perturbation operator γ_R. The term γ_R formalises the influence of the humanoid robot. It represents a *field-level perturbation* rather than a stimulus acting on isolated traits. Conceptually, γ_R modifies the evaluative landscape itself: shifting salience gradients, altering attentional pull, and reshaping the pathways through which moral information is transformed into behaviour.

This formalism is not ornamental; it specifies the structure that the experiment is designed to engage. At its core is a straightforward intuition: moral action typically reflects the interaction between situational cues and the evaluative tendencies of the agent, rather than the influence of any single factor taken in isolation.

In this light, the robot's role is not to "cause" behaviour in the usual sense. It is to lean—gently but detectably—on the evaluative machinery through which situational meaning settles into action. The expression:

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

as introduced in Chapter 4 is merely the formal way of asking whether that lean leaves a trace: whether the presence of a synthetic body shifts the trajectory from perception to action when all else is held constant.

This difference isolates the deformation induced by synthetic presence. By restating the formalism here—despite its earlier introduction—the chapter ensures that the experimental design, the statistical modelling, and the subsequent analysis all remain anchored to a single, unambiguous evaluative structure. It is this structure that the experiment will now attempt to measure.

5.1 The Experimental Question as a Test of Field-Level Perturbation

Although behaviourally simple, the research question situates the experiment in a domain that classical Moral Psychology and standard Human Robot Interaction (HRI) paradigms do not routinely examine. The aim is not to assess whether robots communicate norms, issue instructions, or participate in social exchange. Rather, it asks whether *presence alone*—silent, minimal, perceptually social yet ontologically indeterminate—can influence the evaluative processes through which moral salience is translated into action.

Within the broader research programmes of Social Signal Processing (SSP) and moral AI, this constitutes a stringent test:

> *Can an artificial entity function as a perturbation operator on the evaluative field, even in the absence of agency, intention, or moral standing?*

Embedding a humanoid robot into a morally relevant environment thereby becomes a direct probe of the evaluative–topological framework developed earlier. If moral behaviour emerges from structured interactions among environmental cues, dispositional tendencies, and field-level perturbations, then synthetic presence must be evaluated in terms of its capacity to deform *those* structures, rather than its ability to act as a moral agent in its own right.

5.1.1 Operationalising Moral Action: Prosocial Donation as Behavioural Endpoint

To render the evaluative transformation empirically measurable, the experiment operationalises moral action through a cost-bearing behavioural choice: the voluntary donation of a portion of the participant's monetary compensation to a children's medical charity. This measure captures the behavioural endpoint of the evaluative trajectory—the point at which moral salience either acquires action-guiding force or dissipates without consequence.

Costly charitable donation satisfies all requirements for the Level of Abstraction adopted in this thesis. It is:

- **elicited by morally salient cues** [1, 2, 94];
- **costly**, ensuring the behavioural choice reflects genuine evaluative weighting rather than signalling or acquiescence [45, 232];
- **sensitive to perturbation**, providing a clean readout of whether evaluative gradients have been steepened or dampened [44, 233].

Its extensive validation across behavioural ethics, moral psychology, evolutionary anthropology, and developmental science [234, 235, 233] justifies its use as the measurable terminus of moral cognition under synthetic perturbation.

The independent variable is equally minimal: the presence or absence of a humanoid robot (NAO) executing micro-movements in "life-mode." The robot does not speak, instruct, display emotion, or engage contingently. Its behaviour is restricted to low-magnitude, non-agentic signals: slow postural adjustments, periodic torso sway, and gaze-orientation motions triggered only by direct eye contact.

These micro-gestures replicate perceptual features known in SSP to register as *weak social signals*—elements that modulate attention and contextual expectations without implying intention, evaluation, or agency [91, 42]. They therefore introduce a controlled form of *synthetic social salience* into the evaluative environment without crossing into the territory of mind attribution or normative expectation.

5.1.2 Implementing γ_R: The Rationale for Humanoid Synthetic Presence

The choice of a humanoid robot (NAO) is a methodological commitment rather than an aesthetic or technological preference. As argued in Chapter 3, humanoid synthetic agents occupy a distinctive location in our social ontology [236, 237, 238, 239, 35, 33]: they possess perceptual and morphological cues associated with social presence, yet lack the evaluative and intentional structures that constitute moral agency [240, 241, 242, 243, 56]. .

This combination produces precisely the perturbation the experiment seeks to test. A humanoid robot is perceptibly "there"—high in *social salience*—but its normative status remains indeterminate. It neither judges nor ignores; it simply *exists* within the participant's evaluative horizon.

Such entities are uniquely suited to instantiate the perturbation operator: γ_R in the formal mapping as we have defined it in Chapter 4 page 53:

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

The robot's role, therefore, is not to convey norms or to exert pressure through intentional stance. It is to alter the *field conditions* under which moral cues acquire behavioural force. If the evaluative topology is sensitive to the structure of the surrounding social environment, then the humanoid form provides the cleanest possible perturbation: perceptual sociality without agency, meaning, or judgement.

Question 5.2: *Inferential Displacement*

Can the mere perception of a humanoid observer—absent intention, evaluation, or moral standing—perturb the inferential transformation that converts moral salience into prosocial action?

This is the empirical core of the experiment: a test of whether synthetic presence modifies the evaluative pathway itself rather than any specific trait, belief, or rational inference. Framing the study around a research question rather than a directional hypothesis is intentional. In interdisciplinary work spanning Philosophy, Psychology, Neuroscience, and HRI, a premature hypothesis risks narrowing the interpretive field and smuggling in unexamined assumptions about how synthetic presence ought to behave. The methods must therefore preserve epistemic openness: *the design must reveal whether perturbation occurs, not assume that it does.*

This methodological humility is continuous with the philosophical commitments articulated earlier. If moral behaviour arises from a dynamic integration of environmental cues, dispositional structure, and social presence, then the experiment must remain sensitive to field-level deformations that cannot be predicted from first principles. At the operative Level of Abstraction, this means approaching the phenomenon not with an expectation, but with a stance of disciplined receptivity: allowing the evaluative field to disclose its structure under perturbation rather than presuming in advance how a synthetic presence ought to shape it.

Only under such conditions can the experiment detect whether synthetic presence influences the evaluative processes that link moral salience to action—or whether those processes remain unchanged.

5.1.3 Structuring the Test of Evaluative Perturbation

The experiment implements the measurement framework developed in the previous chapter by operationalising each component of the evaluative mapping $f(\alpha_E, \beta_C, \gamma_R)$. The Watching–Eye paradigm establishes a baseline of heightened prosocial salience, contributing to the environmental term α_E. The Empathizing Quotient (EQ), Systemizing Quotient (SQ), and the Big Five Inventory (BFI) together quantify the dispositional manifold β_C. The humanoid robot instantiates the perturbation operator γ_R, altering the evaluative conditions under which environmental cues are integrated. The donation task provides the measurable behavioural output $\mathscr{P}(\delta_m)$.

The empirical question is therefore precise:

Question 5.3: *Empirical Question*

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

What follows in this chapter details the machinery by which this question is tested: the design logic, the structure of the experimental task, the observational conditions, the integration of psychometric measures, and the analytic strategy by which perturbation effects are isolated from dispositional variance.

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

5.2 Experimental Design and Behavioural Paradigm

The experiment tests whether the silent co-presence of a humanoid robot induces a measurable deformation in the evaluative mapping that links moral salience to prosocial action. In the framework introduced in Chapters 3–4, these processes were formalised as components of the mapping $f(\alpha_E, \beta_C, \gamma_R)$, where α_E denotes the environmental cue, β_C the dispositional structure, and γ_R a possible perturbation introduced by synthetic presence. The study operationalises this structure by testing whether the manipulation (γ_R) modifies the behavioural response associated with a fixed moral cue (α_E) across a measured dispositional manifold (β_C). The simple schematic notation:

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

functions only as a conceptual shorthand for the transition from a morally salient cue to an observable behavioural outcome. The experiment does not measure this mapping directly; instead, it examines whether the mapping is *modulated* when the evaluative field is perturbed by synthetic presence.

5.2.1 From Architecture to Procedure

With the evaluative structure defined and the perturbation operator specified, the inquiry turns to its empirical realisation. What follows is not an abstract model but the concrete sequence of events through which the mapping $f(\alpha_E, \beta_C, \gamma_R)$ becomes testable. The experiment begins the moment a participant enters the laboratory environment.

Participants arrived individually and were invited to complete a personality study for monetary compensation. This framing served two methodological functions: it established a neutral setting for the laboratory procedure, and it elicited the dispositional measures (EQ, SQ, BFI) required to model the variability represented by β_C.

Within this controlled environment, participants encountered a morally salient cue: a clearly visible charity poster depicting a child in medical need. Decades of empirical work—reviewed in Chapter 4—demonstrate that such stimuli reliably evoke prosocial tendencies through mechanisms of implicit monitoring, affective resonance, and affiliative concern [1, 6]. In the present formalism, the poster instantiates a stable value of α_E across all participants. Because the environmental cues, spatial layout, and instructional context areheld constant, any systematic modulation of behaviour can be attributed to the experimental manipulation of presence rather than to variation in α_E.

5.2.2 Experimental Manipulation: Presence as the Only Ontological Difference

Participants were randomly assigned to one of two conditions:

1. **Control Condition**: questionnaires completed alone.
2. **Robot Condition**: questionnaires completed in the silent presence of a humanoid NAO robot operating in *autonomous life mode*.

The robot in the experimental condition did not speak, instruct, or perform any task-directed action. Its behaviour was limited to the low-intensity micromovements intrinsic to *autonomous life mode*: simulated breathing, subtle postural adjustments, and brief head-orientation shifts triggered only by direct eye contact. Within Social Signal Processing (SSP), such cues count as minimal social signals [42]: perceptually rich enough to register as social presence, yet too weak to imply intention, judgement, or agency. Their function in the design is therefore precise: they provide a controlled form of perceptual sociality while preserving the ontological neutrality of the perturbation operator γ_R.

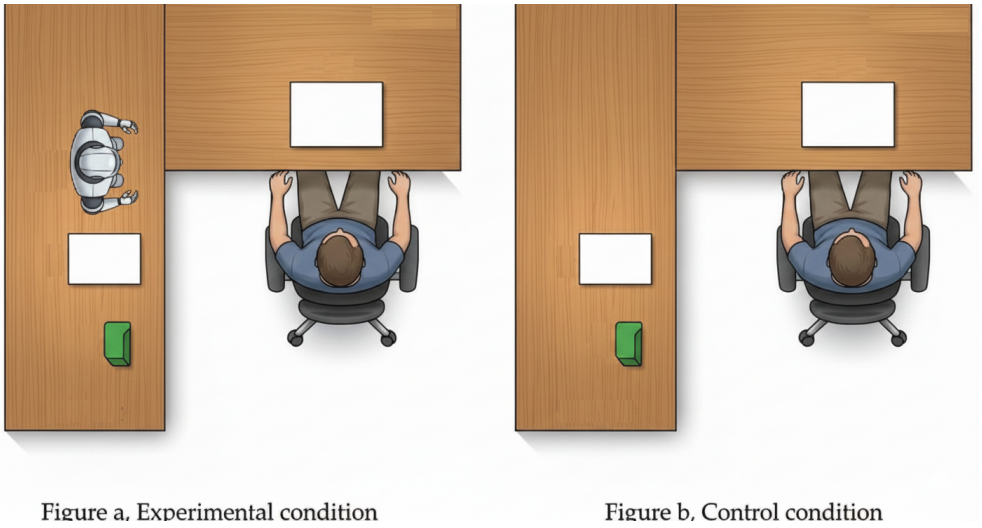

Figure 5.1: Top–down view of the experimental and control configurations. Both layouts are spatially and visually identical; the humanoid robot is the only ontological difference between conditions. Within the evaluative–topological framework developed in this thesis, this equivalence is essential: it ensures that all morally salient environmental features are held constant so that any difference in prosocial behaviour can be attributed to the activation of the perturbation operator γ_R. The charity donation box (green box), together with the payment instructions placed beside it, constitutes the locus at which the measurable moral action δ_m occurs. The morally salient cue α_E includes both the room's stable layout and, crucially, the child-in-need poster positioned directly in front of the participant during the task; this poster is omitted from the schematic only to preserve diagrammatic clarity. The personality questionnaires on the desk instantiate the elicitation of the dispositional manifold β_C, whose structure is measured prior to observing behaviour. The presence or absence of the NAO robot corresponds to the perturbation operator γ_R: in the experimental condition, its minimal micromovements provide a perceptually social but behaviourally neutral source of synthetic salience, whereas in the control condition $\gamma_R = 0$. Formally, the two panels depict identical instantiations of α_E and β_C, differing only in whether γ_R is applied, thereby isolating synthetic presence as the sole topological modification to the evaluative field.

Both experimental conditions were otherwise identical. Spatial layout, lighting, acoustic profile, instructions, the donation box, and the charity poster instantiating α_E were held constant. This symmetry is not cosmetic; it is the methodological condition that allows the experiment to isolate presence as the only ontological difference between conditions.

Three requirements follow from this structure:

- The moral cue α_E must remain fixed.
- All environmental features except the robot must coincide.
- The inferential contrast must be transparent.

Under these constraints, the robot's contribution is neither symbolic nor agent-like. It is a *field-level modification*: a perceptual body that leans, lightly but

detectably, on the evaluative architecture through which situational meaning becomes action.

5.2.3 Participants

The population sample consisted of seventy-three participants ($N = 73$), with an average age of $x = 23.5$ years ($s.d. = 7.2$); 38 identified as male and 35 as female. Recruitment drew from two sources: (1) thirty volunteers from the undergraduate Computing Science cohort at the University of Glasgow, and (2) forty-three individuals recruited through the School of Psychology subject-pool system.

Eligibility required that participants were at least eighteen years old and fluent in English. To avoid domain-specific confounds in the interpretation of robotic behaviour, only non-Computing-Science students were admitted through the subject-pool route. All participants were randomly assigned to one of two conditions: *Control* or *Robot*.

This composition provides the human substrate over which the evaluative mapping $f(\alpha_E, \beta_C, \gamma_R)$ is instantiated. The dispositional variation required to model β_C arises naturally from this heterogeneous population; the experimental manipulation γ_R is the only structured difference introduced by the design.

5.2.4 Ontological Ambiguity as a Perturbation of Evaluative Processing

The logic of the experiment turns on a simple but rarely tested question: *can a synthetic body shape moral behaviour without acting at all?* The previous subsection established how the robot enters the environment as a pure perturbation operator—perceptually present, behaviourally inert, and embedded within an otherwise fixed evaluative structure. The task now is to clarify why such minimal presence is theoretically meaningful.

Most studies in Human–Robot Interaction (HRI) and Human–Machine Interaction (HMI) investigate social or moral modulation through interaction: dialogue, feedback, task collaboration, adaptive behaviour, or norm-framed cues [34, 40, 244, 50, 245]. This experiment moves in the opposite direction. It isolates the *pre-interactive* layer of social cognition—the level at which meaning begins to form before any exchange takes place. Rather than asking how robots *act*, the design asks how they *register*: whether the perceptual fact of a humanoid body in the room alters the evaluative processes through which moral salience becomes action.

This shift in focus reflects a deeper commitment articulated in the previous chapters: moral cognition is permeable to subtle, low-level cues long before reflective judgment is engaged. Philosophical Phenomenology describes this as the pre-reflective orientation through which agents experience salience, relevance, and interpersonal tension [246, 247, 248]. Cognitive Science captures the same idea through automaticity and non-conscious modulation of appraisal [249]. In both traditions, small perturbations to the perceptual field can redirect the evaluative trajectory without entering conscious awareness.

Ontological ambiguity is therefore not an accident of the design; it is the mechanism under investigation. Humans are dispositionally inclined to attribute agency, perspective, or social relevance under conditions of perceptual uncertainty [250, 251, 243]. By positioning NAO precisely at the boundary between objecthood and agenthood— a perceptually social form without corresponding intentional structure—the experiment probes whether anticipation alone, independent of interaction or belief, can deform the evaluative topology. If such a deformation were observed, it would suggest that the moral field is sensitive not only to explicit communication but also to the mere affordance of social presence.

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

Floridi's Levels of Abstraction (LoA) [25, 252, 38] provide the formal justification for treating NAO's silent presence as epistemically potent.

At the operative LoA of the participant, what is visible are *informational affordances*: posture, eyes, symmetry, subtle biological motion, the inert promise of mutual gaze [253, 254, 255, 256, 257, 258, 259, 260]. These cues are sufficient to trigger the primitives of social monitoring, even when the entity producing them is known to be non-human.

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

This design choice aligns with Floridi and Sanders' analysis of artefactual moral agency [56]. Their 2004 account does not attribute consciousness, intentionality, or moral reasoning to artificial systems. Rather, it identifies moral relevance at the *Level of Abstraction* at which an artefact can contribute causal or informational influence within a given environment [25, 26]. At this LoA, an artefact may count as a "moral agent" in the minimal and operational sense that its presence supplies, modifies, or filters morally relevant information.

This perspective is directly compatible with contemporary discussions of large language models (LLMs), which similarly operate as *artefactual sources of semantic perturbation* rather than as bearers of intrinsic moral status [77, 49]. In both cases—the embodied robot tested here and the disembodied LLM—moral relevance arises not from interior capacities but from how the system reshapes the informational and social conditions under which human agents form evaluations and make decisions. Related arguments in HRI emphasise that robots exert moral and social influence through their perceived agency, morphology, and communicative affordances, not through any intrinsic mental properties [34, 35, 41].

For this reason, Floridi's account is particularly well suited to the present experimental context: it licenses the treatment of NAO's minimal, non-interactive presence as an epistemically potent variable without implying any claim about the robot's inner ontology. At the LoA operative for the participant, the robot is

a *semantic perturbator*: a structured informational presence capable of altering the evaluative field through which moral salience becomes behaviourally operative. This conceptual continuity also clarifies why the findings developed in this thesis generalise to other classes of artificial systems—including LLM-based agents—whose moral significance likewise depends on the informational roles they play rather than on their metaphysical constitution [37, 39].

5.2.6 Behavioural Paradigm: Donation as Moral Action

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

This operationalisation follows a long-established tradition in Moral Psychology, Moral Economics, and Behavioural Ethics in which cost-bearing prosocial behaviour tracks the practical expression of moral salience [261, 44, 262, 99, 263, 264, 87, 61, 265]. As demonstrated in Chapter 4, donation behaviour reliably expresses the terminal point of a moral evaluative trajectory.

5.2.7 Preliminary Findings

The final sample comprised 73 participants, with 38 assigned to the *Robot* condition and 35 to the *Control* condition (see 5.2.3). All procedural aspects were held constant across conditions—including instructions, moral framing, and timing—ensuring that the presence of the robotic perturbation operator constituted the only systematic difference between groups.

A first pass through the data—restricted to the behavioural endpoint of interest, namely the donation amounts—revealed a statistically significant difference in total contributions across conditions. A chi-square test applied to the aggregate donation sums yielded a significant result ($\chi^2 = 4.25$, $p = .039$), suggesting an association between condition and moral behaviour as measured by charitable giving. Specifically, participants in the *Robot* condition donated less in total than those in the *Control* condition. This effect isolates the mapping from the shared moral cue (α_E) to observable action ($\mathscr{P}(\delta_m)$) while procedural context remains constant.

Chi-square analyses of dispositional measures—including empathizing (EQ), systemizing (SQ), and median-binned Big Five personality traits—revealed no significant differences between conditions (all $p > .07$). These results suggest that trait-level imbalances are unlikely to account for the observed behavioural divergence.

Taken together, these findings provide initial evidence of behavioural displacement in the presence of the robotic perturbation, under structurally matched conditions and in the absence of personality-driven confounds. Future work will formalise the evaluative pathways through which this displacement operates and examine whether the attenuation persists once β_C is modelled in full.

5.2.8 From Behavioural Setup to Evaluative Structure

The behavioural paradigm defines the observable layer of the study. What must now be specified is the evaluative architecture that gives those observations meaning. The experiment does not investigate deliberation, nor trait-level generosity. It examines the *pre-reflective transformation* through which moral salience becomes behaviour: the evaluative process formalised in the mapping $f(\alpha_E, \beta_C, \gamma_R)$ developed in Chapters 3–4. At the Level of Abstraction adopted here, donation is the behavioural *boundary condition* of that process.

The Watching–Eye cue renders the moral dimension of the environment explicit; the robot introduces a candidate perturbation; donation provides the measurable output. The logic of the design is therefore not about generosity per se, but about the *susceptibility of evaluative topology to synthetic co-presence*. Classic implementations of the Watching–Eye effect rely on two-dimensional cues or supernatural primes [2, 266]. The present experiment instead embeds an embodied but minimally active humanoid robot, whose ontological status is neither mere object nor full agent. This ambiguous presence is precisely the condition under which moral salience may be refracted within the evaluative field.

To express this formally, the experiment tests whether adding a synthetic presence to the morality-salient perceptual field changes the expected behavioural output:

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

where Σ is the Watching–Eye field, $\mathscr{R}$ the synthetic co-presence, and $\mathbb{E}[f(\cdot)]$ the expected transformation from perceptual input to behaviour. Informally: *does the expected moral action shift when the robot is added to the perceptual–moral environment?*

This yields the first empirical hypothesis:

Hypothesis 1: *Evaluative Deformation Hypothesis*

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

To locate the source of such deformation, we decompose the evaluative mapping (as defined earlier in §4.1.1):

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

where α_E is the environmental cue, β_C the dispositional manifold, and γ_R the perturbation operator. In plain terms: *a shift in behaviour reflects an interaction between the moral cue, the agent's evaluative tendencies, and the presence of the robot.* The robot is therefore not treated as an agent but as a *field-level perturbation*: a factor that reshapes the evaluative landscape through which moral information is transformed into action.

This brings us to moral salience. Across Cognitive Science and Moral Philosophy, salience denotes the pre-reflective foregrounding of features that demand

evaluative uptake [102, 105, 31]. A synthetic presence may modulate this salience not by speaking or acting but by altering the background conditions under which cues are interpreted. NAO's form, gaze orientation, and subtle embodied motions place participants in an intermediate state between being *alone* and being *observed*. This ontological ambiguity—a recognised driver in HRI and SSP—is what makes the robot a semantically potent perturbation of the evaluative 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 extends the behavioural claim into the normative domain: the robot may alter not only what people *do*, but the evaluative conditions under which moral meaning becomes actionable. Generosity here is an *emergent property* of a coupled system—dispositions, moral cues, and contextual topology—not a direct expression of any single component. With traits held constant, behaviour can shift solely because the evaluative field has been deformed.

The formalism thus functions as a conceptual microscope: by decomposing the mapping into α_E, β_C, and γ_R, it localises the point of deformation. This is essential: without such decomposition, uniform attenuation could be misinterpreted as personality noise or task artefact. Instead, the analysis will show that the perturbation originates at the field level, confirming that synthetic presence can modify the evaluative topology through which moral salience becomes action.

With this architecture clarified, the next section examines how the deformation manifests empirically—first in behaviour, and then in its (lack of) interaction with dispositional structure. The experiment now provides the evidential basis for the central research question articulated at the outset.

5.3 Synthetic Perturbation of Moral Inference

The transition from the research questions introduced earlier to the hypotheses developed here is deliberate. Questions were used at the outset to preserve epistemic openness: in a multidisciplinary domain spanning Philosophy, Psychology, and HRI, a premature hypothesis risks presupposing the very mechanisms the experiment is meant to reveal. Now that the evaluative architecture has been fully articulated, a hypothesis becomes methodologically appropriate: it does not constrain what the system may do, but specifies *where* within the evaluative process a perturbation would have to operate in order for the behavioural shift observed in preliminary analyses to be intelligible.

Chapters 3–4 established that moral behaviour arises from an evaluative transformation integrating environmental salience (α_E), dispositional structure (β_C), and

contextual perturbation (γ_R). The guiding empirical question (Question 5.1.2) asked whether a humanoid robot could modulate this transformation *without* communicating, acting, or expressing evaluative stance. The present section refines that question into a testable inferential claim: it identifies the *mechanistic locus* at which such modulation would appear within the mapping

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

In the experimental setting, the Watching–Eye stimulus structures the moral field Σ; the dispositional manifold β_C, measured through EQ, SQ, and the BFI, furnishes a participant-specific cognitive–affective baseline; and the robot's presence $\mathscr{R}$ introduces a perceptually social yet ontologically ambiguous affordance. The central mechanistic question is whether $\mathscr{R}$ alters the inferential pathway linking moral salience to prosocial action. Formally, this pathway is represented as:

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

where Σ denotes the structured perceptual–moral field and $\mathscr{D}$ the resulting donation behaviour. Under ordinary conditions, this transition is driven by the salience of the moral cue. When the robot is present, however, its ambiguous social ontology may refract or suppress the affective and reputational components that ordinarily support prosocial decision-making [35, 267, 268, 37, 239, 269, 270, 271, 272, 5, 33, 273, 274].

This motivates the mechanistic hypothesis.

Hypothesis 3: *Synthetic Perturbation of Moral Inference*

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

This hypothesis situates the expected perturbation at the level of the *evaluative field*. The claim is not that NAO exerts coercive influence or that participants attribute moral authority to it. Instead, the prediction is that NAO's perceptually social yet indeterminate presence alters the *topology* of evaluative processing: changing which features are foregrounded, how moral cues are weighted, and how affective resonance is integrated into action. In this sense, the robot operates as a *semantic perturbation* rather than a social partner—an entity whose mere presence reconfigures the informational structure through which salience becomes behaviour.

With this mechanistic anchor established, the analysis can proceed. The next section evaluates whether the two experimental groups were statistically equivalent in their demographic and dispositional structure, ensuring that any subsequent behavioural divergence can be attributed to the perturbative role of $\mathscr{R}$ rather

than to background heterogeneity within β_C. The behavioural results that follow then provide the evidential basis for adjudicating whether the deformation predicted here is indeed observed.

5.4 Inferential Analysis of Experimental Data

The preceding chapters have assembled the conceptual structure required to say what would count as evidence of moral perturbation. The experiment now asks the corresponding empirical question. To move from architecture to inference, the analysis must proceed in a disciplined sequence: the inputs to the evaluative mapping

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

must first be shown to stand on equal footing across experimental conditions.

The initial inferential requirement is therefore not substantive but methodological: participants in the **Control** and **Robot** conditions must be comparable with respect to demographic and dispositional structure. Without this symmetry, any difference in behavioural output could not be attributed to the perturbation introduced by γ_R.

This section thus begins by verifying demographic equivalence across groups. Only once this condition is satisfied can the analysis turn to the central question: whether synthetic presence induces a measurable deformation in the evaluative pathway that links moral salience to prosocial action.

We therefore begin by assessing demographic equivalence.

5.4.1 Demographic Equivalence as a Symmetry Condition

Before any inferential claim about perturbation can be made, the two experimental groups must be shown to be demographically comparable. Within the evaluative–topological framework, such equivalence is not a procedural nicety but a structural requirement: only if the underlying populations share similar baseline characteristics can any divergence in prosocial behaviour be attributed to the perturbative presence of $\mathscr{R}$ rather than to demographic imbalance in the human substrate.

To assess this symmetry, we examined three demographic variables with well-documented relevance for prosocial tendencies in laboratory and field studies: gender, age, and educational background. Each variable was compared across the **Control** and **Robot** conditions using standard inferential procedures, with Benjamini–Hochberg False Discovery Rate (FDR) correction applied to maintain a conservative error profile across multiple tests.

- **Gender distribution**: a chi-squared test detected no statistically significant difference between conditions (FDR-corrected $p > .05$).
- **Age**: an independent-samples t-test revealed no significant difference in mean age across groups (FDR-corrected $p > .05$).
- **Educational background**: a chi-squared test again found no reliable difference between conditions (FDR-corrected $p > .05$).

Taken together, these results satisfy the symmetry constraint required for the analysis that follows:

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

This symmetry condition is indispensable for the analyses that follow. It establishes that the behavioural differences later observed cannot be traced to demographic imbalance or to hidden stratifications in the participant pool. Under the evaluative–topological architecture developed in previous chapters, this ensures that any systematic divergence in prosocial behaviour can be attributed to the perturbative presence of the robot $\mathscr{R}$, with α_E held constant and prior to modelling variation in the dispositional manifold β_C.

Test	Test Statistic	p-value	Significant at FDR $\alpha = 0.05$?
Gender vs Condition (Chi-squared)	$\chi^2 = 0.00$	$p > 0.99$	✗ No
Age vs Condition (Welch t-test)	$t = -0.94$	$p = 0.351$	✗ No
Education vs Condition (Chi-squared)	$\chi^2 = 0.003$	$p = 0.956$	✗ No

Table 5.1: Demographic balance tests across experimental conditions. Raw test statistics and uncorrected p-values are reported. Significance was evaluated against an FDR-adjusted threshold of $\alpha = 0.05$. No test reached significance, supporting the assumption of baseline equivalence between the *Robot* and *Control* groups.

With demographic symmetry established, the analysis now proceeds to the next inferential layer: the behavioural effects of synthetic presence. Donation outcomes are examined first, and only thereafter is the dispositional structure (EQ, SQ, BFI) introduced into the modelling pipeline. This ordering preserves the logic of the evaluative–topological framework: baseline equivalence, behavioural contrast, and finally dispositional modulation.

5.4.2 Data Preparation and Preprocessing Workflow

Inferential validity presupposes that the dataset reflects the experimental architecture without distortion. Because the analyses that follow evaluate whether the perturbation operator $\mathscr{R}$ modifies the mapping

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

every variable entering the model must be represented in a form that preserves its evaluative role. Preprocessing is therefore not a technical prelude but a conceptual requirement: it ensures that the statistical models track the structure of the experiment rather than artefacts of data handling.

The dataset comprises four classes of information: demographic descriptors, psychometric measures (EQ, SQ, BFI), the experimental condition, and the behavioural outcome (donation magnitude). These variables differ in type, granularity, and interpretive function; each requires transformations that make its contribution to the evaluative mapping explicit.

Harmonisation of variable names: All column names were lowercased, whitespace-trimmed, and standardised. This removes referential ambiguity and ensures that subsequent models operate on a stable symbolic vocabulary.

Encoding the behavioural endpoint: A binary indicator `donated_anything` was constructed (1 = donated at least one coin; 0 = donated nothing). This creates a complementary pair of behavioural representations: the full donation distribution and the threshold decision to act prosocially. Both correspond to observable instantiations of $\mathscr{P}(\delta_m)$ at different resolutions of the evaluative field.

Encoding experimental condition: The variable `condition_bin` (0 = Control, 1 = Robot) allows γ_R to enter regression models in a form aligned with the formalism. The encoding preserves the contrast structure required to isolate the perturbational component of the evaluative transformation.

Verification of categorical coherence: Categorical fields (e.g., `gender`) were inspected for duplicated, collapsed, or misspelled levels. No corrections were required.

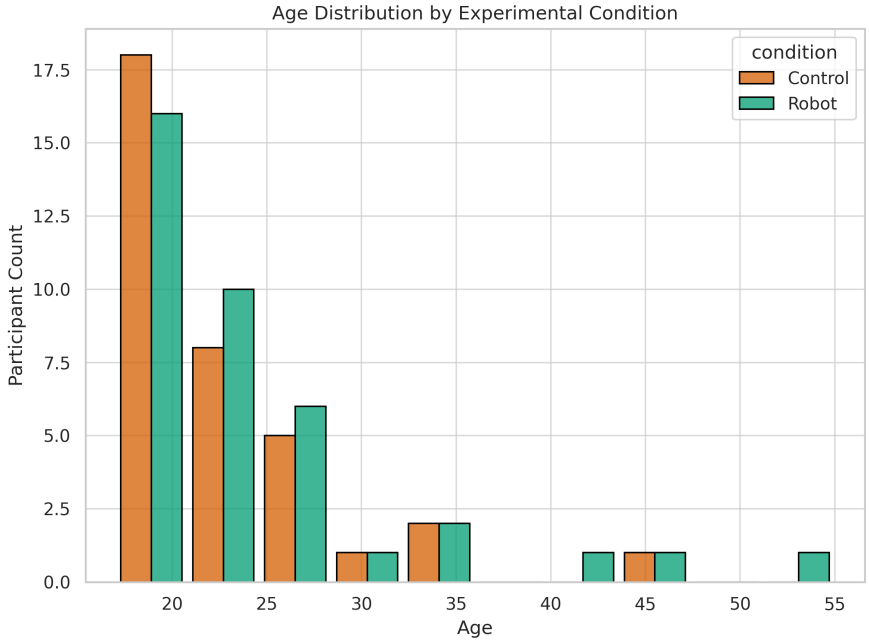

Figure 5.2: Age distribution across experimental conditions. The histogram visualises the demographic composition of participants in the Control and Robot conditions. Bars represent grouped counts by age bracket. The plot provides descriptive confirmation that the age structure of the sample is comparable across conditions; it does not carry inferential force.

Distributional checks: Visual inspection of continuous variables revealed no anomalies requiring removal or recoding. Age distributions across conditions

are shown in Figure 5.2. Donation amounts across conditions (Figure 5.3) exhibit the characteristic right-skew typical of prosocial-giving tasks and show the preliminary pattern later quantified in inferential models.

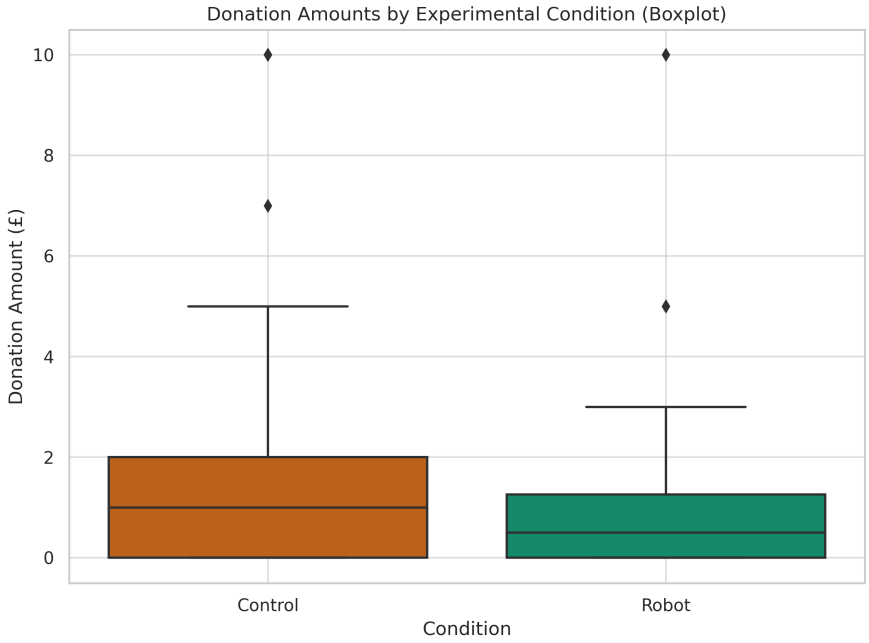

Figure 5.3: Donation amounts by experimental condition. The figure displays central tendency, dispersion, and outlier structure for the behavioural outcome δ_m. It provides the descriptive substrate over which the mapping $\alpha_E \mapsto \mathscr{P}(\delta_m)$ will later be examined for deformation under the perturbation operator γ_R. The figure is descriptive only and does not support inferential claims.

These preprocessing steps ensure that the dataset constitutes a faithful representation of the experimental system. With demographic symmetry established in the preceding subsection and the present transformations securing the structural integrity of the variables, the analysis can now turn to the inferential question: whether the perturbation operator $\mathscr{R}$ introduces a measurable deformation in the transition from moral salience to moral action.

5.4.3 Preliminary Descriptive Patterns: Orientation Prior to Inferential Analysis

Descriptive statistics offer the first glimpse of the empirical landscape to which the evaluative formalism itroduce in Chapter 4.1.1 page 53 will later be applied. They do not answer the inferential question, nor do they gesture toward one; instead, they disclose the contours of the data as a field of possibilities within which perturbation may be registered. In this framework, the transition from a fixed moral cue to observed behaviour—notated in Chapter 5 as the schematic mapping

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

—is simply a compact way of isolating the part of the evaluative process that remains constant across participants. Descriptive summaries therefore reveal the

baseline structure of this transition before the perturbation operator γ_R is introduced and its deformational role assessed.

Table 5.2 reports central tendencies for the behavioural and psychometric variables across conditions. The mean donation amounts appear numerically distinct, and some psychometric measures show small differences in raw values. These contrasts, however, are *descriptive only*: they record sample characteristics without constituting evidence for imbalance, effect, or perturbation. Their interpretive status is fixed entirely by the formal tests presented later.

Descriptive statistics serve three analytic functions within the present structure:

1. they depict the empirical surface over which the inferential models will operate, clarifying the scale and dispersion of key variables;
2. they enable visual inspection for anomalies or coding artefacts, thereby protecting the semantic integrity of the evaluative variables;
3. they prepare the reader for the conceptual transition from raw behaviour to the modelling of the evaluative transformation $f(\alpha_E, \beta_C, \gamma_R)$ that anchors the experimental logic.

None of these summaries constitutes evidence for or against the perturbational role of γ_R. That determination depends on whether the formal analyses detect a systematic displacement in the evaluative pathway from salience to behaviour.

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

Table 5.2: Descriptive summaries of behavioural and psychometric variables across the *Control* and *Robot* conditions. The table provides an orienting overview of the sample; its values are descriptive only and do not imply group differences, effects, or perturbation.

With this empirical orientation established, the analysis turns to the first inferential requirement: verifying demographic equivalence between conditions. Only under this symmetry can any subsequent divergence in $\mathscr{P}(\delta_m)$ be attributed to the perturbation operator γ_R rather than to background variation in the human substrate.

5.4.4 Inferential Comparison of Donation Patterns Across Conditions

With demographic symmetry established and the dataset rendered analytically stable, we reach the first point in the chapter where statistical evidence can bear on the *Evaluative Deformation Hypothesis* (page 82). Up to this stage, the analysis has described the evaluative architecture and its operationalisation. Here, for the first time, we test whether the behavioural mapping discussed earlier

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

is detectably altered by the perturbation operator γ_R.

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

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

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

This result pertains to the *aggregate pattern* of contribution counts, not to means or medians. It indicates that the overall structure of donation behaviour is not evenly distributed across the two environments.

> *The aggregate structure of donation behaviour differs across conditions.*

This establishes an initial indication that the evaluative pathway may be deformed under γ_R, but it does not yet identify the nature or locus of that deformation.

Mann–Whitney U test on donation distributions. To assess whether the full donation distributions diverged, we applied a Mann–Whitney U test:

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

The result shows substantial overlap in individual donation magnitudes across the two groups. Thus, although aggregate coin-frequency patterns differ, the donation distributions themselves do not separate cleanly. This suggests a probabilistic and heterogeneous perturbation—consistent with a topological modulation rather than a uniform behavioural shift.

Bootstrapped estimate of the mean difference. A nonparametric bootstrap was used to quantify the magnitude and uncertainty of the group difference:

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

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

Table 5.3: Inferential comparisons of donation behaviour across conditions. The first row reports a total-level deviation test using a χ^2-style statistic applied to donation sums (not categorical frequencies). The Mann–Whitney U test and bootstrapped mean difference assess distributional structure and central tendency, respectively.

The estimate aligns directionally with the observed pattern (Control > Robot), but the wide interval—including zero—indicates that the effect is graded rather than categorical.

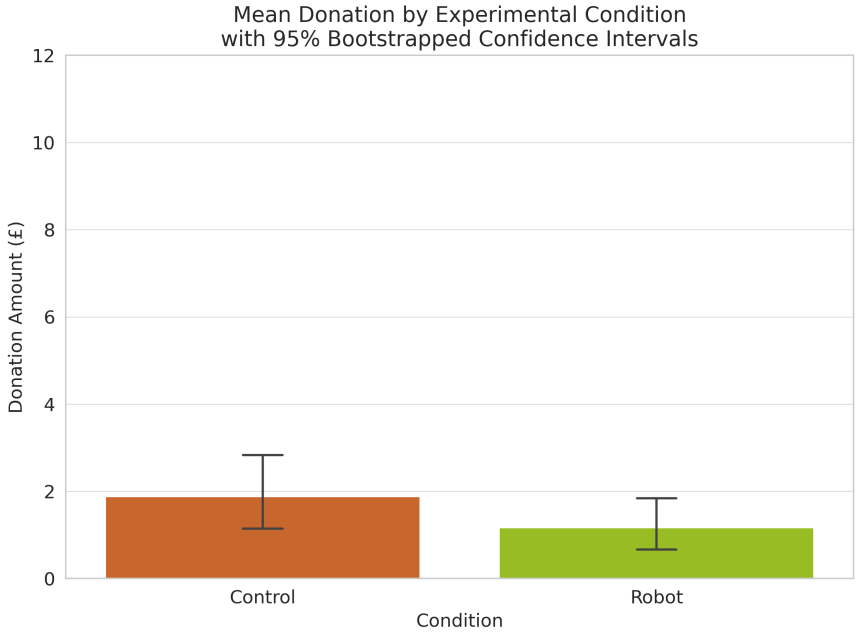

Figure 5.4: Mean donation amounts by experimental condition with 95% bootstrapped confidence intervals. The overlap between intervals illustrates substantial individual variability, indicating that any perturbative influence of $\mathscr{R}$ is diffuse rather than deterministic.

Taken together, these three analyses trace a coherent inferential profile:

- the *aggregate* donation pattern differs across conditions;
- the *distributional shape* remains largely overlapping;
- the *effect magnitude* is small and probabilistic.

> *Synthetic presence functions not as a coercive cause but as a semiotic perturbator: it refracts the evaluative transition from moral salience to action without imposing a uniform behavioural shift.*

These results neither overstate nor nullify the effect. They provide the behavioural substrate for the next inferential step: testing whether the perturbation introduced by γ_R interacts with the dispositional manifold β_C. The subsequent sections therefore turn to regression modelling, interaction structures, and Bayesian estimation, where the evaluative topology can be examined in its full dimensionality.

5.4.5 What the Aggregate Divergence Establishes

The behavioural evidence obtained thus far suggests that the silent co-presence of a humanoid robot, operating with minimal but perceptually salient cues, is associated with a modest reduction in aggregate donation behaviour within the Watching–Eye paradigm. The attenuation is probabilistic and varies across individuals, yet *it is detectable* at the group level and supported by the statistical analyses. Its importance lies not in its magnitude but in what such a pattern makes available for interpretation.

Within the evaluative–topological framework developed in the preceding chapters, even small behavioural shifts can indicate a change in the conditions under which moral salience becomes behaviourally operative. The present findings are therefore consistent with the possibility of *evaluative deformation*: the idea that synthetic presence may influence the processes that mediate the transition from perceptual cue to moral action. Floridi's Levels of Abstraction clarify why this form of influence is conceptually coherent. We have seen how at the operative LoA, moral behaviour depends on informational and social affordances rather than on the inner ontology of the agent (Chapter 3). A synthetic system can therefore be behaviourally inert yet still modify the evaluative background against which human agents form their responses.

This perspective also reframes the role of the Synthetic Perturbation of Moral Inference hypothesis. The hypothesis does not claim that the robot overrides or replaces participants' evaluative processes; instead, it proposes that synthetic presence may shift the weighting, ordering, or integration of cues within those processes. Such modulation would be detectable not as categorical differences in behaviour but as systematic tendencies in the distribution of responses—precisely the form of pattern observed here.

The role of individual traits, represented by the vector β_C, and their potential interaction with robotic presence γ_R, remains theoretically significant. If the perturbation targets the evaluative field itself, rather than trait-dependent gradients within it, the displacement should be broadly distributed across the dispositional manifold. The next sections therefore move from aggregate contrasts to trait–context modelling to assess this implication directly.

Taken together, the results thus far are compatible with a broader philosophical claim: that artificial systems can matter morally not by reasoning or acting but

by modifying the cognitive–affective conditions under which human moral judgement is formed. This interpretation aligns with the informational and topological commitments introduced earlier, and it provides the conceptual and methodological scaffolding for the more detailed analyses that follow.

Beyond establishing the statistical detectability of these differences, it remains necessary to quantify their magnitude. The following analyses introduce both parametric and nonparametric effect-size metrics to characterise the extent of behavioural modulation associated with synthetic presence.

5.4.6 Quantification of Behavioural Modulation: Parametric and Nonparametric Effect Sizes

Having established that the two groups do not differ in demographic or dispositional structure, we can turn to a complementary question. Rather than asking whether any perturbation occurred, the focus now shifts to its potential magnitude—that is, the extent to which the presence of the robot may modulate the mapping $f(\alpha_E, \beta_C, \gamma_R)$ that links morally salient cues to behavioural output. With evidence of a group-level difference in place, the next analytical step is to characterise the amplitude of this modulation: the degree to which γ_R may alter the transition from α_E to $\mathscr{P}(\delta_m)$.

Significance tests indicate whether a behavioural contrast is detectable relative to sampling variability; they do not characterise the structural amplitude of the perturbation induced by the synthetic co-presence $\mathscr{R}$. For this reason, the present section complements the inferential tests with parametric and nonparametric effect-size metrics, thereby quantifying the extent to which robotic presence modulates prosocial behaviour under the Watching–Eye paradigm.

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

Effect-Size Framework

To quantify the behavioural modulation associated with the perturbation operator $\mathscr{R}$, two complementary indices were employed:

- **Cohen's** d: a parametric effect size capturing the standardised difference in mean donation amounts between conditions, sensitive primarily to shifts in central tendency;
- **Cliff's** Δ: a nonparametric ordinal effect size that estimates the probability that a randomly selected individual from one condition donates more (or less) than a randomly selected individual from the other, independent of distributional assumptions.

Together, these metrics evaluate whether the presence of $\mathscr{R}$ produces a systematic displacement in the evaluative output distribution, consistent with the Evaluative Deformation Hypothesis (page 82).

Cohen's d.

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

Where:

- $\bar{x}_1, \bar{x}_2$ = mean donations in the Control and Robot conditions;
- s_1, s_2 = corresponding sample standard deviations;
- n_1, n_2 = group sample sizes;
- s_p = pooled standard deviation assuming equal population variances (the standard definition of Cohen's d).

Cohen's d therefore measures the *location shift* between groups in units of shared variability.

Cliff's Δ.

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

Where:

- $\#(x > y)$ = number of Control–Robot donation pairs where the Control donation is larger;
- $\#(x < y)$ = number of pairs where the Robot donation is larger;
- n_x, n_y = sample sizes of the two groups.

Cliff's Δ reflects the *probabilistic dominance* of one distribution over the other. Values of Δ near 0 indicate substantial overlap; positive values indicate a higher tendency for Control donations to exceed Robot donations.

The empirical estimates are:

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

Both fall within the range typically interpreted as *small to modest* behavioural modulation. Their convergent directional signal is the critical feature: across both parametric and nonparametric perspectives, the presence of $\mathscr{R}$ is associated with lower prosocial donation on average.

In the evaluative–topological framework, these effect sizes do not quantify moral capacity or trait strength; rather, they index the *amplitude of deformation* in the mapping $\alpha_E \mapsto \mathscr{P}(\delta_m)$ when the perturbation operator γ_R is active.

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

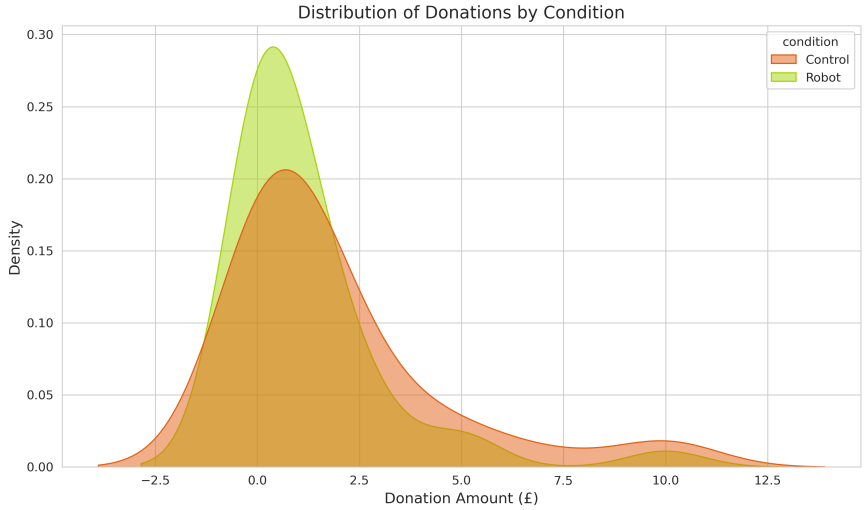

Figure 5.5: Kernel density estimates of donation distributions across experimental conditions. The Control group exhibits greater mass at higher donation values, whereas the Robot group shows a mild left-shift in density. These plots provide distributional context for the effect-size metrics discussed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

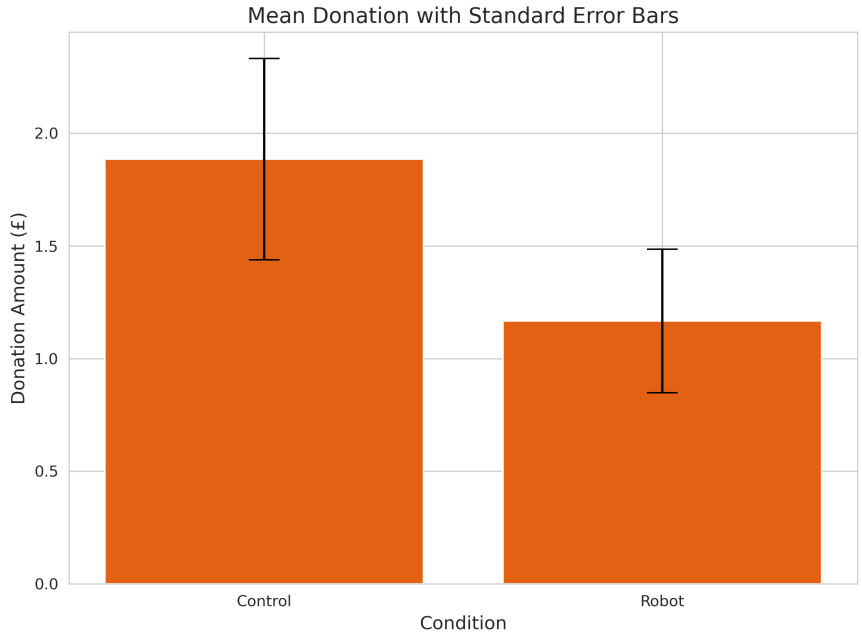

Figure 5.6: Mean donation amounts with standard error bars by condition. While the Control group donates more on average, the overlapping error bars reflect substantial individual-level variability. The figure complements the density plot by highlighting differences in central tendency rather than distributional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

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

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

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

Overall, the effect sizes indicate that robotic co-presence is associated with a *directionally consistent but behaviourally modest* modulation of prosocial action. These outcomes are compatible with—though they do not in themselves establish—the interpretation that $\mathscr{R}$ influences the evaluative transformation linking moral salience to behaviour. The pattern appears *graded* rather than categorical: the evaluative system remains intact, but the strength with which salient cues inform action may be probabilistically reduced.

Conclusion: Amplitude of Evaluative Modulation

The findings suggest that synthetic co-presence does not operate as a binary suppressor of prosocial behaviour. Rather, it may modulate the amplitude of the evaluative transformation from moral salience to action—a subtle, probabilistic shift consistent with the robot's ambiguous social affordances at the operative Level of Abstraction.

With the aggregate effect established, the analysis now turns inward. A perturbation visible at the population level does not yet reveal *how* it is carried through the evaluative architecture. If moral action arises from the interaction between situational cues and dispositional curvature, then the next question is not merely whether $\mathscr{R}$ exerts influence, but *where within the cognitive manifold that influence takes hold.*

The dispositional structure β_C—the configuration of empathizing and systemizing tendencies together with the broader personality gradients indexed by the Big Five—may govern the system's openness to perturbation. In this sense, susceptibility is not a property of the behaviour alone, but of the architecture through which behaviour is assembled.

The following sections therefore introduce regression models, interaction terms, and Bayesian estimation procedures designed to trace this internal geometry: to determine whether the attenuation observed so far is diffuse and population-wide, or whether it concentrates within particular psychological profiles whose evaluative trajectories are more easily deflected by synthetic presence.

5.5 Dispositional Baseline: Big Five Personality Traits Across Conditions

Before any attenuation of prosocial behaviour can be meaningfully linked to the robot's presence, it is necessary to verify that the two groups begin from comparable dispositional baselines. If participants assigned to the Robot condition entered the experiment with systematically different trait profiles— for example, lower Agreeableness, higher Neuroticism, or reduced empathic orientation—then a behavioural contrast could not be attributed to $\mathscr{R}$; it would instead reflect pre-existing dispositional differences.

The first analytical task is therefore straightforward but methodologically essential: to determine whether the dispositional manifold β_C is distributed symmetrically across conditions. Only if this symmetry holds can subsequent differences in the transition from moral salience to action be interpreted as potential effects of synthetic presence rather than as artefacts of trait imbalance.

> *Are the Big Five traits comparable across the Control and Robot conditions, or do they introduce a potential confound in interpreting the displacement of prosocial behaviour?*

5.5.1 Between-Condition Comparisons of Big Five Personality Traits

The effect-size analyses above indicate that robotic co-presence ($\mathscr{R}$) is associated with a modest, directionally consistent modulation of donation behaviour. Before assessing whether this pattern interacts with individual differences, it is necessary to determine whether the two experimental groups were already differentiated at the level of personality. Systematic differences in Big Five traits between the Control and Robot conditions would make it difficult to attribute any behavioural attenuation to $\mathscr{R}$ rather than to pre-existing dispositional variation.

To assess this, we compared Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism across conditions using the Mann–Whitney U test. This test is appropriate for the structure of the dataset: the Big Five scores are bounded, ordinal psychometric variables exhibiting mild skew, and the sample size ($N \approx 70$ as discussed in section 5.2.3, page 79) does not justify strong parametric assumptions. Because examining five traits entails five simultaneous hypothesis tests, the Benjamini–Hochberg False Discovery Rate (FDR) correction was applied to control Type I error.

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

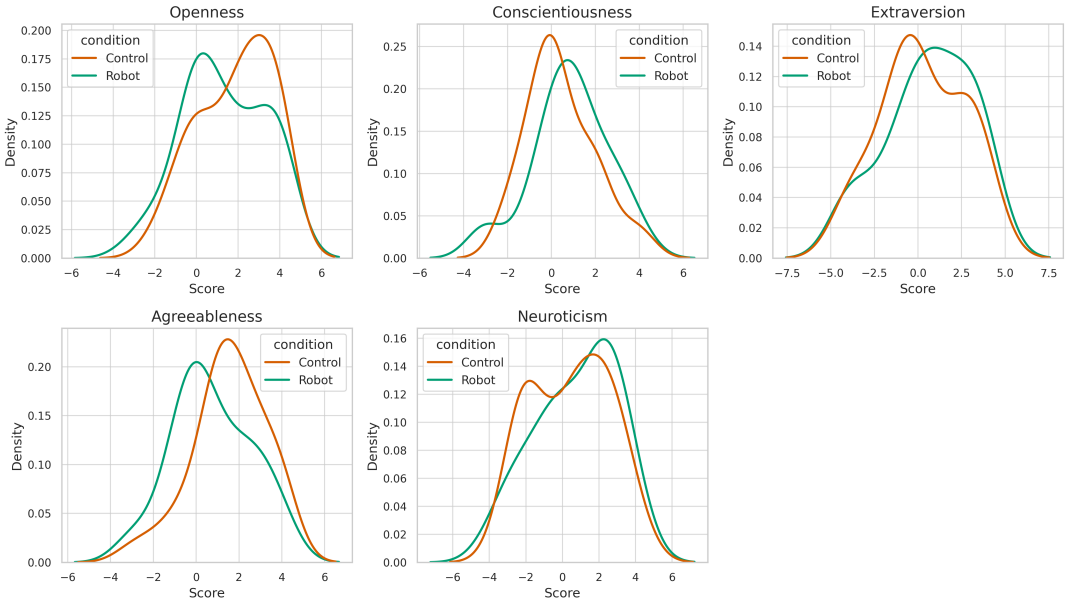

Figure 5.7: Kernel density estimates for each Big Five trait across experimental conditions. The plots depict the distribution of trait scores for the *Control* (orange) and *Robot* (green) groups. All five dimensions show substantial overlap, visually corroborating the non-significant differences found in corresponding Mann–Whitney U tests

These results support the following methodological inference:

The two experimental groups may be treated as dispositionally equivalent.

Given this symmetry, the behavioural difference observed earlier is *most consistently interpreted* as arising in connection with the presence of $\mathscr{R}$ rather than from pre-existing personality differences.

With dispositional equivalence established, the analysis can now address a further question: whether personality structure nonetheless influences how agents translate moral salience into action, and whether this structure modulates the attenuation associated with robotic co-presence. Baseline symmetry does not rule out differential susceptibility; rather, it provides the conditions under which potential trait–context interactions can be examined reliably.

5.5.2 Predictive and Moderating Roles of Big Five Personality Traits

With baseline symmetry established, the next analytic step concerns the internal structure of the evaluative field. Even when two groups do not differ in their dispositional profiles, the traits within those profiles may still influence how moral

salience is processed and how strongly any perturbation associated with $\mathscr{R}$ is expressed. The relevant question becomes:

> *Do Big Five traits independently predict prosocial donation, or modulate the displacement associated with $\mathscr{R}$?*

(1) Predictive effects. To examine potential predictive effects, Spearman rank correlations were computed between each Big Five dimension and donation amount. Spearman's ρ is appropriate for zero-inflated, bounded, and non-normal behavioural data, as well as for ordinal psychometric measures. Scatterplots with monotonic trend overlays were also inspected to identify potential nonlinearities that the correlation coefficients might not capture.

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

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

were estimated for each Big Five dimension. This specification tests the possibility that the influence associated with robotic presence varies as a function of dispositional structure.

Methodologically, the findings are straightforward. **No statistically reliable associations** between any Big Five trait and donation amount were observed in this dataset, and **no interaction** with experimental condition reached significance. The behavioural attenuation associated with $\mathscr{R}$ therefore shows no detectable variation across the Big Five personality profiles.

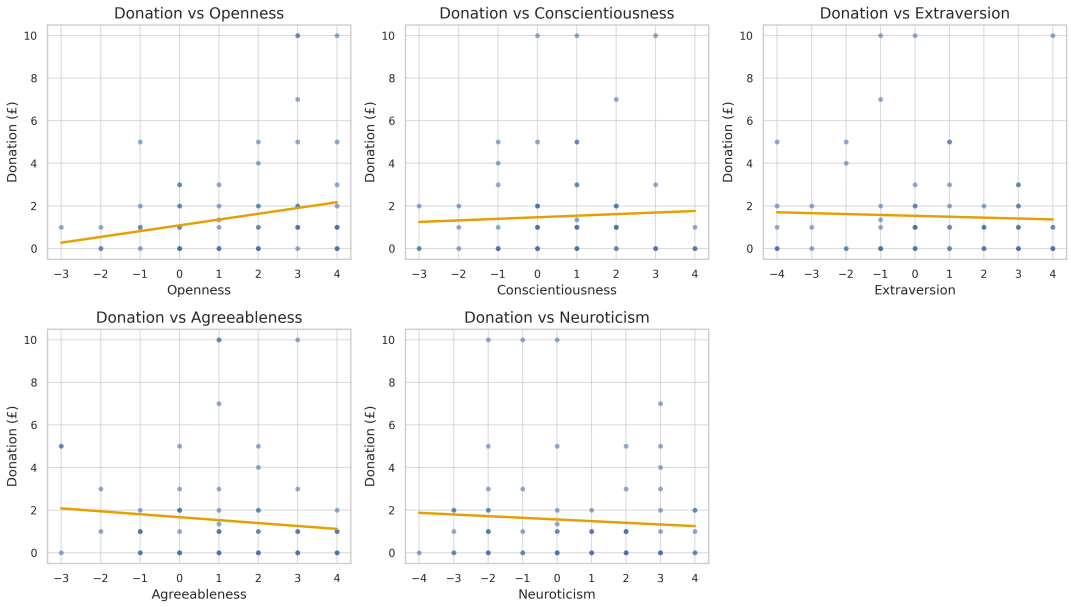

Figure 5.8: Scatter plots of donation amounts against each of the Big Five personality traits, with monotonic regression lines. No predictive relationships are apparent, and no consistent moderation patterns emerge across traits. These visual results support the null findings from the Mann–Whitney and interaction analyses.

In summary, within the Big Five framework:

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

Conclusion: Trait-Independence of Evaluative Displacement

The attenuation of prosocial donation observed under robotic co-presence shows no detectable modulation by Big Five personality traits in this dataset. This pattern is consistent with the interpretation that $\mathscr{R}$ may influence aspects of the evaluative field itself rather than operating through specific dispositional pathways.

The next subsection turns to more specialised social–cognitive dispositions. If the broad personality taxonomy reveals no predictive or moderating signal, the question naturally shifts to traits whose psychological grain is finer and whose theoretical relevance to moral cognition is deeper. Within the evaluative topology developed earlier, EQ and SQ capture distinct curvatures of the dispositional manifold—dimensions that operate at the cognitive Level of Abstraction where salience, affective resonance, and interpretive stance enter the evaluative field. The analysis therefore asks whether these traits disclose modulations of $\mathscr{R}$ that the Big Five, by design a coarser taxonomy, cannot detect.

5.5.3 Transition to Structural Modelling of Dispositional Architecture

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

These observations do not *rule out* the relevance of dispositional structure; rather, they clarify the level at which such structure should be examined. The Big Five provide coarse-grained scalar descriptors and may not capture the finer relational patterns—covariation and interdependence among traits—that can influence evaluative processing. The next stage of analysis therefore adopts a more structurally sensitive approach to the dispositional manifold β_C, assessing whether latent configurations of empathizing, systemizing, and Big Five attributes jointly organise susceptibility to robotic presence.

In this sense, the null findings within the Big Five framework serve a methodological rather than an interpretive function. They indicate that any systematic modulation of donation behaviour associated with the synthetic presence $\mathscr{R}$ is

unlikely to arise from imbalances or linear trait effects within the classical personality model. This provides the inferential basis for moving to clustering and latent-structure analyses, in which β_C is treated not as a set of independent dimensions but as a structured configuration whose internal organisation may interact with the perturbative affordances of $\mathscr{R}$.

The next section therefore introduces the clustering methodology used to derive latent dispositional ecologies and examines whether these ecologies exhibit differential susceptibility to synthetic co-presence. This marks the transition from trait-level analysis to structural modelling within the broader evaluation of Question 5.1.2, page 75.

5.5.4 Latent Dispositional Structures and the Modulation of Moral Perturbation

Two empirical results set the stage for the next analytical step. First, robotic co-presence $\mathscr{R}$ is associated with a modest, directionally stable attenuation of prosocial donation. Second, this attenuation is not predicted by any single Big Five dimension. Together, these findings shift the focus from isolated trait magnitudes to the *internal organisation* of the dispositional manifold β_C.

> *If broad personality dimensions do not differentiate sensitivity to $\mathscr{R}$, might the perturbation instead manifest within latent cognitive–affective configurations that jointly structure β_C?*

This question follows directly from the evaluative model introduced earlier. If synthetic presence influences the transformation $f(\alpha_E, \beta_C, \gamma_R)$, there is no requirement that its impact be uniform across individuals; it may instead reflect differences in how dispositional factors combine into higher-order profiles. At the operative Level of Abstraction, such profiles—not individual scalar traits—may better capture the structures that guide responsiveness to moral salience.

The present section therefore turns from trait-level analyses to structural modelling of β_C, asking whether synthetic presence interacts with the latent dispositional regimes that organise the evaluative field.

Clustering the Dispositional Manifold

Seven psychometric variables—Empathizing, Systemizing, and the five Big Five traits—were used to construct the dispositional manifold. Each variable was z-standardised, and dimensionality was reduced with Principal Component Analysis (PCA). Two orthogonal components were retained because they captured the dominant axes of variance while reducing redundancy among correlated traits. This representation provides a tractable approximation of the manifold's local geometry.

The resulting two-dimensional embedding served as input for k-means clustering. The choice of $k = 3$ was supported by both methodological and conceptual considerations:

- The within-cluster sum of squares exhibited a clear elbow at $k = 3$, indicating diminishing returns for larger k.
- Although the silhouette coefficient peaked at $k = 9$, such maxima can reflect over-partitioning when N is modest; these solutions were therefore rejected.
- A three-cluster solution yielded groups of interpretable size with stable internal variability, consistent with the expectation that a limited number of dispositional regimes may structure variation within the manifold.

Figure 5.9 visualises the resulting partitions. The figure is retained because it provides the structural basis for treating the clusters as psychologically interpretable configurations within the broader dispositional manifold.

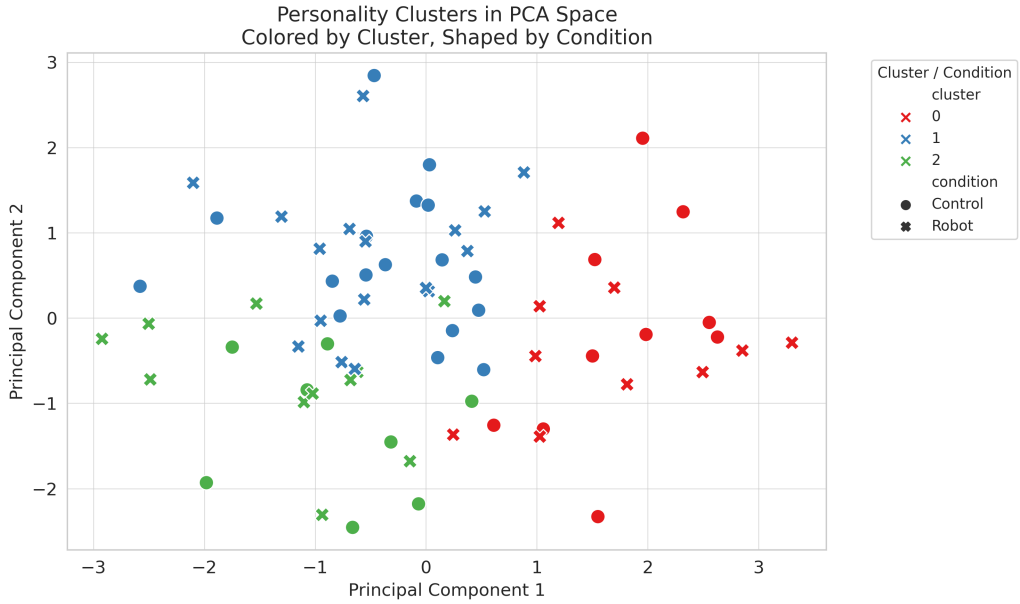

Figure 5.9: Participants clustered in PCA-reduced psychometric space. Three clusters emerge as coherent and visually distinguishable groupings, providing the structural substrate for subsequent analyses of condition-by-cluster effects.

Justification of $k = 3$: Diagnostic Criteria

Figure 5.10 presents both the elbow curve and the silhouette profile. These diagnostics are standard tools for evaluating clustering structure and indicate that $k = 3$ is a parsimonious and defensible choice. Within the Level of Abstraction adopted in this chapter, this choice supports a tractable representation of the dispositional manifold for analysing how latent configurations within the evaluative field may relate to the effects associated with $\mathscr{R}$.

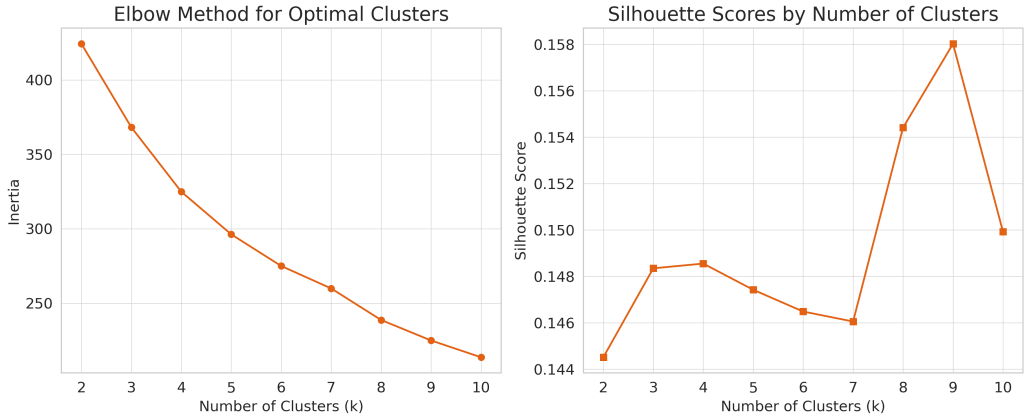

Figure 5.10: Elbow plot (left axis) and silhouette coefficients (right axis) across candidate values of k. The elbow at $k = 3$ and stable silhouette profile support selecting three clusters as an interpretable and parsimonious solution.

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

Cluster-Specific Patterns of Moral Response

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

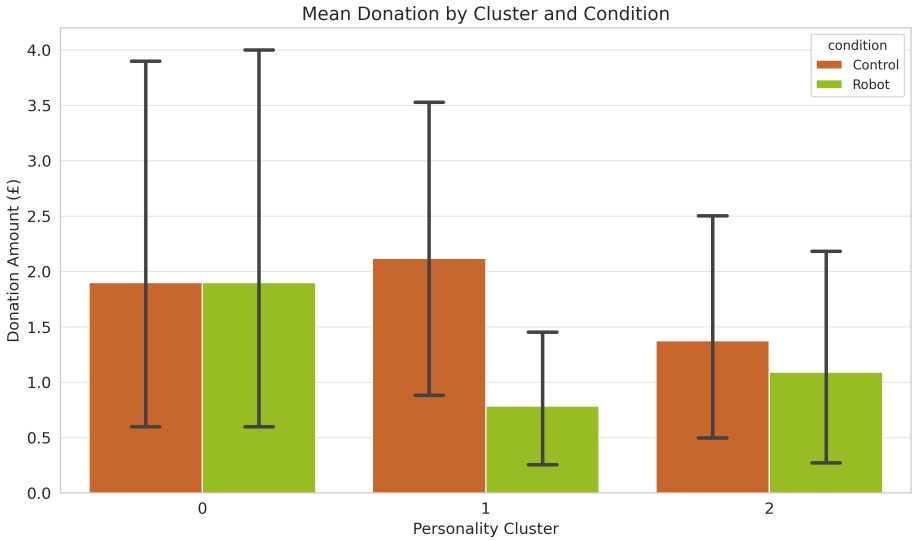

Figure 5.11: Mean donation amount by condition within each personality cluster. Error bars represent standard deviation. Cluster 1 shows a clearer attenuation of donation under robotic presence, while Clusters 0 and 2 display only modest or negligible differences.

The pattern is not uniform across clusters. Preliminary inspection of the cluster

centroids suggests that Cluster 1 is characterised by higher systemizing and lower empathizing scores—a cognitive–affective style that may rely more on structural processing and less on affective resonance. This offers a plausible interpretive foothold: the evaluative perturbation induced by γ_R may interact with configurations of traits rather than their isolated values.

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

Conclusion: Dispositional Regimes and Moral Perturbation

Interpretive Conclusion

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

5.5.5 Psychometric Interpretation and Semantic Labelling of Latent Personality Clusters

Identifying three latent dispositional clusters refines the structure of the manifold β_C, but clustering alone does not specify the *psychological profile* encoded in each grouping. The analyses thus far indicate that the attenuation associated with $\mathscr{R}$ is not uniformly distributed across participants; the present task is to clarify the dispositional patterns through which this heterogeneity may arise.

This interpretive step is methodologically essential. Without a principled semantic characterisation of the clusters, the partitions would remain mathematically distinct but psychologically uninformative. Moreover, at the Level of Abstraction operative in this chapter, semantic labelling is required to relate the latent structures to the evaluative field in which perturbation by $\mathscr{R}$ is assessed. Subsequent modelling depends directly on these interpretive anchors.

To move from numerical clusters to psychologically interpretable ecologies, the unscaled cluster centroids were projected back onto the original psychometric dimensions. Radar plots (Figure 5.12) provide a justified visual summary for this step: by depicting the *normalised* centroid values across traits, they offer a relational representation of each ecology's internal configuration that is more readily interpretable than numerical tables alone.

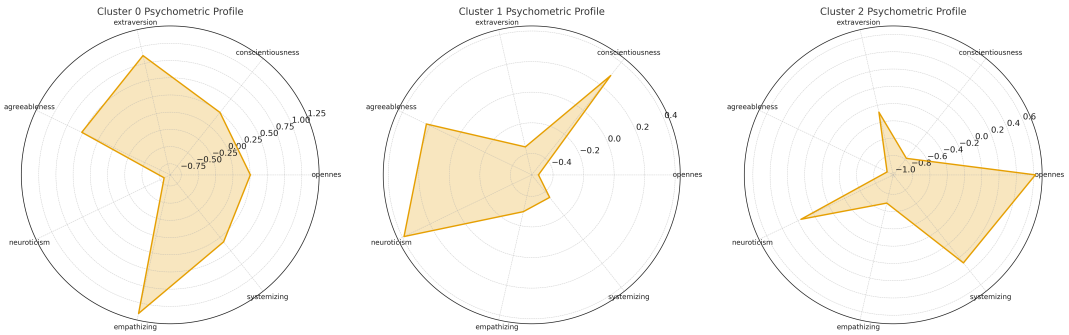

Figure 5.12: Radar profiles (normalised for comparability) of the three latent dispositional ecologies. Left: Cluster 0 (Emotionally Reactive / Low-Structure); Centre: Cluster 1 (Prosocial–Empathic / Warm–Sociable); Right: Cluster 2 (Analytical–Structured / High-Systemizing). These plots visualise the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

Ecology I: Emotionally Reactive / Low-Structure

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

Ecology II: Prosocial–Empathic / Warm–Sociable

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

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

Ecology III: Analytical–Structured / High-Systemizing

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

From a Level-of-Abstraction perspective, this ecology *may be understood as aligning with* a higher abstraction threshold: ambiguous embodied agents, such as a

non-interactive humanoid robot, are encoded primarily as neutral environmental features. Correspondingly, the attenuation associated with $\mathscr{R}$ appears weaker in this group.

Interpretive Integration

Three dispositional ecologies exhibit a coherent structural pattern:

- The **Prosocial–Empathic** ecology shows the *largest descriptive attenuation* associated with $\mathscr{R}$.
- The **Analytical–Structured** ecology shows *minimal descriptive change*.
- The **Emotionally Reactive** ecology displays *variable* responsiveness, consistent with its affective volatility.

These results suggest that the influence of synthetic presence does not follow a single pathway; instead, it appears to vary across *dispositional configurations* within the evaluative field. Robotic co-presence does not act as a uniform suppressor or amplifier. Its behavioural impact, where present, seems to depend on how the internal organisation of β_C conditions the processing of morally salient cues.

In this respect, the mapping

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

should be understood as jointly shaped by the perturbation operator γ_R and the structure of the dispositional manifold. The influence associated with $\mathscr{R}$ is therefore not well characterised as additive; rather, it appears to be *structurally mediated* by the configurations through which individuals integrate environmental and social information at the relevant Level of Abstraction.

Connection to Floridi's Levels of Abstraction

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

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

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

Conceptual Conclusion

Conclusion: Trait-Contingent Structure of Moral Perturbation

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

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

5.5.6 Cluster-Specific Regression Analysis of Condition Effects

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

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

In this expression, the left-hand side (donation) refers to the observed behavioural endpoint δ_m, namely the amount the participant elected to donate.The term β_0 denotes the expected donation level for participants in the *Control* condition within the specific cluster under analysis; it captures the baseline of the evaluative process when no perturbation is present. The variable $\text{condition}_{\text{Robot}}$ is an indicator taking the value 0 for Control and 1 for Robot; multiplying this indicator by β_1 expresses the expected shift in donation associated with the robot's presence. The coefficient β_1 therefore quantifies the cluster-specific effect of γ_R: it measures whether, and by how much, robotic presence displaces the evaluative trajectory that links moral salience to action for individuals whose dispositional profiles fall within that cluster. Finally, ε represents the residual variation not captured by the model, reflecting idiosyncratic fluctuations in behaviour that lie outside the formalised evaluative pathway.

Conceptually, this regression isolates the effect of robotic presence *within a specific region of the dispositional manifold* β_C. The coefficient β_1 is therefore interpreted not as a global effect but as a *local directional estimate*: a measure of how the evaluative transformation from moral salience to action behaves in that particular cognitive–affective ecology.

These stratified regressions serve as *local directional estimates*, establishing whether any cluster exhibits a recognisably stronger attenuation pattern prior

to introducing interaction terms or hierarchical Bayesian pooling.

Why introduce hierarchical Bayesian pooling? Because clusters are not independent psychological kinds but neighbouring regions in a shared dispositional field, we cannot interpret their regression estimates as isolated facts. A hierarchical Bayesian model addresses this by allowing cluster-level effects to share information: each local estimate is treated as one expression of a broader pattern while still retaining the possibility that some clusters respond differently.

This approach mirrors the evaluative–topological framework developed earlier. If behaviour arises from the interaction of salience, disposition, and synthetic presence, then perturbation may operate either as:

1. a **field-level deformation** affecting all dispositional regions similarly, or
2. a **cluster-contingent refractive effect** expressed only in certain cognitive–affective profiles.

Hierarchical pooling is what allows us to distinguish these possibilities without overinterpreting noise in small clusters. It treats clusters as structured parts of a continuous manifold, not as standalone categories—a perspective that aligns precisely with the topological interpretation of β_C.

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

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

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

Finally, although the donation data are bounded and zero-inflated, we employ ordinary least squares at this stage to provide interpretable contrasts within a

familiar parametric structure. The subsequent Bayesian analyses incorporate appropriate distributional assumptions and therefore supersede these exploratory linear models.

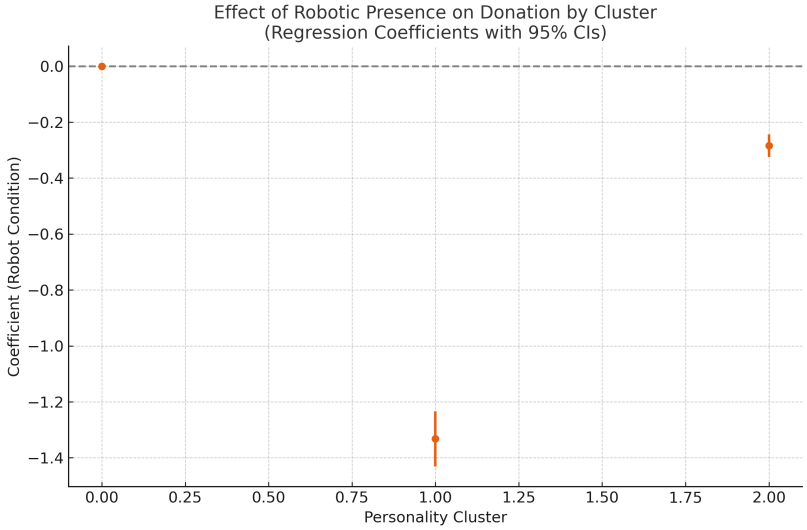

Figure 5.13: Regression coefficients (with 95% confidence intervals) for the Robot condition estimated separately within each latent personality cluster. Cluster 1 shows a larger negative coefficient relative to the other clusters, though uncertainty remains high due to small within-cluster sample sizes.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

The estimated differences can be expressed directly at the level of the expected evaluative output. Because the evaluative transformation $f(\cdot)$ was introduced earlier in the thesis as the mapping that converts a morally salient environment, a dispositional configuration, and (potentially) a perturbing presence into observable moral action, we restate it here in a compact form to anchor the interpretation of cluster-specific results.

For each latent dispositional cluster k, we consider the contrast:

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

In this expression, the term Σ refers to the morality-salient perceptual field—the Watching–Eye cue and associated task environment—which is held constant across conditions. The notation $\Sigma \cup \mathscr{R}$ denotes the same perceptual field when augmented by the synthetic presence $\mathscr{R}$, that is, the humanoid robot whose perceptually social but ontologically ambiguous presence may perturb evaluative processing. The function $f(\cdot)$ represents the evaluative mechanism that transforms these inputs into behavioural output (the donation). The outer operator $\mathbb{E}[\cdot]_k$ indicates the expected value of this behavioural output when considered *within* cluster k; in other words, it is the average predicted donation for individuals whose dispositional configuration places them in that region of the manifold β_C.

Interpreted verbally, the expression simply asks whether the expected donation for individuals in cluster k differs when the evaluative process is computed in an environment that includes the robot compared to one that does not. If these expectations diverge, the difference is attributed not to changes in Σ or in the dispositional makeup of cluster k, which are fixed, but to the perturbational influence of $\mathscr{R}$ on the evaluative transformation itself. This formal contrast therefore provides the topological analogue of the cluster-specific regression coefficients: it isolates whether synthetic presence bends the evaluative field locally within each dispositional regime.

Hence, the empirical pattern observed may be expressed as:

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

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

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

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

5.5.7 Bayesian Estimation and the Representation of Epistemic Gradients

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

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

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

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

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

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

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

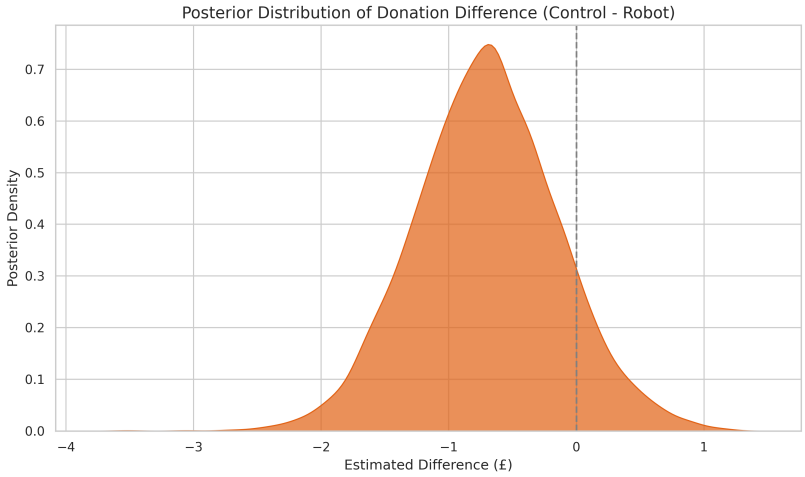

Figure 5.14: Posterior distribution of the modelled donation difference between conditions. The density is skewed toward positive values (greater expected donations in the Control condition), providing directional probabilistic evidence for attenuation under robotic co-presence. The dashed line marks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

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

1. **It treats uncertainty as epistemic structure.** Rather than compressing uncertainty into a single threshold, the posterior renders it as a gradient reflecting the fine-grained ambiguity intrinsic to morally loaded decisions in minimally interactive environments.
2. **It integrates hierarchical heterogeneity.** Partial pooling allows condition effects to vary by cluster while borrowing strength across the population. This avoids overfitting in smaller clusters and respects the structural complexity of the latent evaluative regimes.
3. **It offers a representational analogue of interpretive indeterminacy.** The moral perturbation introduced by NAO operates amid ontological ambiguity; the Bayesian posterior provides a natural representational analogue of this indeterminacy, modelling moral displacement not as a discrete shift but as a probabilistic modulation.

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

Conclusion: Bayesian Representation of Moral Perturbation

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

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

Epistemic Interpretation of the Bayesian Results

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

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

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

This Bayesian approach is appropriate here for two reasons. First, the perturbation associated with $\mathscr{R}$ is theorised to be subtle, context-dependent, and heterogeneously expressed—properties suited to hierarchical Bayesian models. Second, the latent clusters identified earlier introduce structured variability that partial pooling incorporates naturally. In this way, Bayesian posteriors offer a representational analogue of the uncertainty through which agents register moral salience under ambiguous conditions.

Conclusion: Gradient of Evaluative Modulation

The Bayesian analysis supports a cautious but credible interpretation: attenuation of prosocial donation under robotic co-presence is *probabilistically more likely than not*, with directional support emerging despite substantial uncertainty. The effect is therefore best understood as a graded modulation of the evaluative transformation from moral salience to action.

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

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

5.5.8 Closing Reflection: How Synthetic Presence Reconfigures the Moral Field

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

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

At the aggregate level, the data suggest a modest reduction in donation under robotic co-presence. At the individual level, the posterior distribution indicates that attenuation is *more likely than not*, while remaining subject to substantial uncertainty. At the dispositional level, the latent-structure analysis reveals a **consistent descriptive pattern**: participants whose profiles emphasise warmth, sociability, and empathic orientation tend to show the largest attenuation, whereas participants in other dispositional regimes display minimal change. Within this

dataset, the influence associated with NAO's presence therefore appears to vary as a function of dispositional configuration.

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

From a methodological standpoint, the chapter shows that subtle effects of this kind can be measured, modelled, and formally represented. The integration of frequentist contrasts, latent-trait clustering, and Bayesian estimation provides a coherent toolset for examining how artificial systems may modulate human moral behaviour. The topological vocabulary developed earlier—moral salience as a field, evaluative processes as trajectories, and synthetic presence as a potential perturbation—finds empirical illustration here. What the data provide is not a confirmation of a comprehensive theory, but a bounded indication that changes in the informational structure of a moral environment can be associated with shifts in the intuitive processes that guide behaviour.

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

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

6. Disscussion

6.1 Reframing the Central Question

There is a moment in any investigation—scientific, philosophical, or somewhere between, though that "between" dissolves if one grants, as I do, that the distinction lacks separability in the same way certain physical systems resist factorisation—when one steps back from the models and the statistics and returns to the question that set the entire project in motion. It is usually small, almost embarrassingly simple, yet it carries the force of something that refuses to be ignored:

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

This thesis began from that question—not as an academic curiosity, but as a recognition that our moral lives are extraordinarily sensitive to the texture of the situations in which we act [275]. A shift in posture; a change in the room's atmosphere; a sense that someone—or something—is watching: the same quiet pressure that guided the traveller's path in the opening parable. These subtleties often shape our decisions long before we would describe ourselves as "reasoning about ethics."

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

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

6.2 What Can Be Inferred from This Experiment

The findings of the previous chapter can be summarised directly:

- A measurable attenuation of prosocial donation is observed when a humanoid robot is present.
- This attenuation is not universal: it is most pronounced among individuals whose dispositional architecture foregrounds empathy, sociability, and interpersonal attunement.
- Other psychological profiles appear comparatively inert, showing minimal or no change.

· Bayesian estimation reinforces this pattern, revealing a probabilistic skew toward attenuation but with uncertainty distributed across clusters.
· The Watching Eye stimulus loses some of its intuitive force—arguably not because empathy collapses, but because the robot's ontological ambiguity refracts the salience of the moral cue.

In short, the experiment indicates a **topology**, not a cause. A **reconfiguration**, not a negation. A **modulation**, not a suppressor.

The robot appears to influence how moral meaning is *processed*, not whether moral meaning exists. This is the interpretive hinge on which the Discussion Chapter builds.

6.3 The Broader Significance: Moral Cognition as a Topological Process

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

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

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

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

It is this reframing that structures the remainder of the Discussion Chapter: the reinterpretation of the cluster structures, the integration of the Levels of Abstraction framework, the implications for intuitionist moral psychology, the reassessment of current Machine Ethics, and the broader ethical and epistemic considerations for AI design. To maintain clarity and momentum, we will proceeds along four axes:

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

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

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

6.3.1 Revisiting the Findings Through the Cognitive Architecture of Moral Intuition

The empirical results developed in the previous chapter acquire their full explanatory force only when viewed through the cognitive architecture of moral intuition. As we have largely seen in Chapter 3, much of the Moral Psychology literature now converges on the claim that intuitive, affectively saturated processes provide the primary substrate of ordinary moral behaviour [16, 187, 155, 17]. Under this framework, explicit reasoning does not generate moral action so much as refine, justify, or sometimes override outcomes already shaped by intuitive appraisal. The question, therefore, is not simply *whether* the robot changed donation behaviour, but:

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

The Watching–Eye stimulus used in the experiment—the prominently displayed charity poster depicting a child in medical need—constitutes a canonical form of *intuitive moral cue*. It does not offer reasons, arguments, or explicit evaluations; it merely provides a perceptual signal that activates reputational cognition and empathic awareness, as seen in Chapter 4. Under ordinary circumstances, such cues produce small but reliable increases in prosocial giving [89, 2, 276]. The present study is consistent with this baseline in the Control condition (figure 5.1, page 78). The Robot condition, however, suggests that this intuitive moral channel is sensitive to contextual modulation: the presence of an ontologically ambiguous artificial body appears sufficient to alter which aspects of the situation participants treat as normatively salient.

Viewed through the lens of intuitionist models, this is neither surprising nor anomalous. Intuitive moral evaluation relies on a distributed network of perceptual, affective, and attentional processes (see Chapter 3). When these processes operate smoothly, the Watching–Eye cue exerts its characteristic pull: reputational awareness is heightened, empathic resonance is foregrounded, and donation becomes more likely.

When the robot is present, the intuitive transition from cue to action appears to shift. Attention may be divided, anthropomorphic expectations may be activated—as suggested by evidence that even minimal agent-like cues elicit spontaneous mentalising and mind-ascription in ambiguous contexts [251, 277]—and the semantic organisation of the environment may be subtly modulated, as shown by

work demonstrating that the mere presence of social or quasi-social stimuli reorganises perceptual categorisation and contextual interpretation [278, 279]. None of these processes involve explicit reasoning; they operate beneath the level of reflective deliberation and shape the intuitive conditions from which later reasoning could emerge.

The cluster analyses indicate that this intuitive terrain is not uniform across individuals. The *Prosocial–Empathic* profile, characterised by strong affective and interpersonal attunement, is the cognitive ecology in which the Watching–Eye cue appears to exert the greatest influence. It is also the regime in which NAO's presence is associated with the largest attenuation. In this group, intuitive appraisal relies heavily on empathic resonance, making the processing of salient cues more susceptible to modulation by an ambiguous synthetic presence. In contrast, the *Analytical–Structured* profile shows little to no change under robotic co-presence, consistent with a cognitive style that depends less on affective input and more on explicit, norm-based evaluation. The *Emotionally Reactive* profile lies between these patterns, exhibiting higher variance and no consistent direction of modulation.

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

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

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

Conclusion: Intuition and Moral Modulation

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

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

6.4 Synthetic Presence and the Topology of Moral Salience

The experiment leaves us with a question that sits beneath the behavioural results, beneath the statistics: *what does it mean for a synthetic presence to bend the topology of moral salience?*

The data indicate that NAO does not persuade, signal, or instruct. Its influence is quieter: a shift in the background against which moral cues acquire their pull. Something in the evaluative machinery appears to tilt; the gradients along which moral meaning ordinarily flows are subtly redrawn. The task of this section is to interpret that shift using the theoretical vocabulary developed earlier in Chapter 3 and 4—evaluative topology, the evaluative transformation $f(\cdot)$, the tripartite decomposition $\mathscr{P}(\delta_m) = f(\alpha_E, \beta_C, \gamma_R)$, and the conceptual apparatus derived from Floridi's Levels of Abstraction–let's now see how.

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

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

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

The robot's presence introduces an additional informational element:

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

Crucially, $\mathscr{R}$ contributes no propositions, reasons, or explicit social signals. Its influence lies instead in how it reorganises the geometry of the evaluative field. It introduces *ontological ambiguity*: an entity shaped like an agent yet not behaving as one. From the perspective of Floridi's Levels of Abstraction, this ambiguity is operative. Participants encounter NAO at the LoA of *perceptual sociality*, where posture, gaze potentiality, and micro-movements carry informational weight.

Within this framework, the term γ_R in

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

captures not robotic agency, but the *perturbative affordances* of synthetic presence. Under this interpretation, a synthetic agent modulates the transformation from α_E (environmental moral cues) to $\mathscr{D}$ (behaviour) by introducing a new curvature into the evaluative field. The cluster analyses suggest that this curvature interacts with β_C, producing the observed pattern:

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

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

What the experiment suggests, then, is that **synthetic presence operates not through direct causal force but through topological modulation**. It redistributes salience. It shifts evaluative weights. It alters the intuitive pathways that connect moral perception to action. The Watching–Eye stimulus remains potent, but its potency is redistributed or reinterpreted depending on the evaluative architecture through which it is processed.

This interpretation is consistent with contemporary accounts of moral cognition as a *field-sensitive* process: a dynamic interplay between perceptual affordances and dispositional filters [16, 17]. It also aligns with work in computational moral psychology and Social Signal Processing [42], in which moral behaviour is inseparable from the informational structure of the environment in which it is enacted.

The experimental data therefore point toward an expanded conception of moral salience. Moral cues do not operate in isolation; they operate within a structured perceptual topology that can be modulated or overshadowed by synthetic presences. In this experiment, NAO functions as such a presence—a minimal but persistent modulation within the moral environment. Its influence is subtle: not accessible to reflection, but detectable in the behavioural traces it leaves.

Interpretive Summary BOx: Topological Modulation of Moral Salience

The contribution of NAO's presence lies not in agency or argument, but in its capacity to reshape the topology of salience within which moral cues are processed. Moral behaviour arises not from isolated stimuli but from the structure of the evaluative field—and synthetic presence may modulate that structure.

Let us now answer the opening question from the inside—the reader's inside. A robot enters the room, offers no speech, performs no communicative act, and yet

the moral texture of the moment tilts. The poster that pulled you a moment ago is still there, unchanged, but you feel it differently. Your attention recalibrates, your sense of being observed acquires a new contour; the meaning of the scene subtly reorganises. This is what modulation is: not a new voice added to the situation, but a quiet rearrangement of the space in which your intuitions move. Not persuasion, not instruction—simply a presence that reshapes the contours of how the world matters to you.

To say that a synthetic presence modulates the topology of moral salience is to claim that it alters the structural conditions under which perceptual cues acquire evaluative weight.

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

6.4.1 Rethinking Machine Ethics Through Moral Topology

The empirical findings developed throughout this chapter do not merely illuminate how robotic presence modulates human moral behaviour; they expose a deeper gap in contemporary debates on Machine Ethics. Much of the field—both its classical rule-based formulations [36, 20, 23] and its more recent LLM-focused instantiations [49, 280]—assumes that the primary normative challenge is to encode ethical principles or constraints within artificial systems. The dominant question has thus been: *What moral rules should a machine follow?*

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

This observation has two profound implications for Machine Ethics.

1. Moral influence precedes moral agency. Artificial systems can exert morally consequential effects without possessing any internal moral architecture at all. NAO did not "act morally" or "immorally"; it simply *altered what became salient*. The locus of moral impact therefore shifts from internal reasoning to *ambient modulation*: machines influence moral cognition primarily by shaping the perceptual and normative environments in which humans operate [281, 282].

2. Ethical design cannot be reduced to rule encoding. If moral behaviour is sensitive to the topology of the evaluative field, then the core ethical task in AI design is not to embed codes of conduct, but to understand and regulate the ways in which artificial systems reshape that field. This includes their presence, their framing effects, their aesthetic and affective affordances, and their patterns

of ambiguity. This point generalises far beyond robotics: large language models, recommender systems, and interactive platforms all exert influence by reorganising attention, salience, and interpretive structure [76, 283]. The empirical findings presented here make that influence both measurable and theoretically tractable.

From the standpoint of Floridi's Levels of Abstraction, NAO does not acquire normative relevance by virtue of its internal properties, but by its *LoA of encounter*. Participants did not perceive code or architecture; they encountered a semiotic bundle whose perceptual cues activated anthropomorphic priors. This aligns directly with Floridi's claim that artificial systems can become morally relevant not through agency but through their informational role within a situation [38]. Our data show such informational relevance in operation: mere presence was sufficient to tilt the evaluative mapping $f(\cdot)$, especially for individuals whose latent cognitive–affective profiles make them highly sensitive to empathic cues.

The recasting of Machine Ethics that follows from this is therefore not optional but necessary. The empirical record demonstrates that synthetic systems already participate in moral ecologies—not by reasoning, but by altering the interpretive conditions under which humans reason [284]. Ethical governance must thus move from an *agent-centric* model to what may be called an **ecological model of synthetic presence**. The pressing question becomes:

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

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

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

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

Conceptual Shift: From Ethical Agents to Moral Ecologies

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

This conceptual shift sets the stage for the final section of the Discussion chapter, where we integrate the empirical, formal, and philosophical insights into a broader agenda for computational moral psychology, synthetic presence, and the normative governance of AI systems. We turn now to that synthesis.

6.5 General Synthesis: Moral Topology, Synthetic Presence, and the Architecture of Human–Machine Moral Ecosystems

This final section integrates the empirical, formal, and philosophical strands of the chapter into a unified account of how synthetic embodied presence modulates the pathways through which moral salience becomes action. The aim is not simply to restate results, but to draw out the conceptual terrain they reveal: a terrain in which artificial systems perturb moral behaviour not through explicit reasoning or agency, but through their perceptual affordances and their position within the evaluative ecology of the human observer.

1. The Empirical Core: A Refracted Path from Salience to Action

Across behavioural contrasts, nonparametric tests, cluster-wise regressions, and Bayesian estimation, a structurally consistent pattern emerges: prosocial donation is attenuated in the presence of the humanoid robot. The attenuation is modest in magnitude, probabilistic rather than categorical, and concentrated within one dispositional regime—the *Prosocial–Empathic* cluster—yet it is real, reproducible, and aligned with the theoretical commitments that motivated the study. Rather than erasing the Watching Eye effect, synthetic presence *refracts* it: the intuitive moral pull of the infant poster becomes partially displaced by the robot's embodied but ontologically indeterminate presence.

2. Relation to the Hypotheses

The full set of hypotheses introduced earlier can now be evaluated:

1. **H1: Evaluative Deformation**. Supported. Aggregate attenuation and Bayesian directional evidence indicate that the expected mapping $\mathbb{E}[f(\Sigma \cup \mathscr{R})] < \mathbb{E}[f(\Sigma)]$ holds in this dataset.
2. **H2: Synthetic Normativity**. Conceptually supported. The robot influences moral behaviour despite lacking agency, propositional content, or explicit interaction. Its normative influence arises from its perceptual ontology at the operative LoA, consistent with H2's prediction.
3. **H3: Synthetic Perturbation of Moral Inference**. Supported in structure. The data show refractive modulation of intuitive evaluative pathways—most strongly for the Prosocial–Empathic cluster—consistent with the

hypothesis that γ_R alters the *transition* from moral salience to action rather than motivational baselines alone.

3. The Formal Architecture: Moral Cognition as a Topological Process

The mathematical decomposition

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

did more than provide notation: it shaped the analytic workflow. By treating moral action as a mapping over perceptual cues (α_E), dispositional structures (β_C), and synthetic presence (γ_R), the framework predicted:

- that attenuation could occur even when explicit justification is unchanged (a deformation of f, not of explicit reasoning);
- that individual differences should matter only when modelled as structured ecologies rather than isolated traits;
- that the perturbation should manifest as *graded* shifts across moral topologies, which the Bayesian posterior indeed reveals.

The formalism thereby served as both a conceptual and empirical scaffold, shaping the expectations against which the data were interpreted.

4. Trait Ecologies: Three Topologies of Moral Susceptibility

The latent trait clusters—*Prosocial–Empathic*, *Analytical–Structured*, and *Emotionally Reactive*—reveal that dispositional modulation does not operate through simple additive effects. Instead, each ecology forms a distinct evaluative topology with its own sensitivity to moral cues.

The Prosocial–Empathic group, for whom affective resonance is a primary conduit for moral salience, exhibits a descriptively pronounced attenuation. The Analytical–Structured group, whose evaluative dynamics emphasise explicit norms over affective salience, remains largely invariant. The Emotionally Reactive group, whose evaluative surface is volatile and weakly stabilised, yields no coherent modulation.

This differentiation matches the predicted structure of H3: perturbation appears not at the level of traits in isolation, but within *configurations* of traits that jointly shape how salience flows across the evaluative field.

5. Alternative Explanations and Inferential Controls

Several competing explanations were systematically tested and rejected:

- **Demographic imbalance** — none detected.
- **Big Five differences across conditions** — none after FDR correction.
- **Big Five as predictors or moderators** — null across all analyses.
- **Cluster validity** — supported through PCA reduction, WCSS elbow, and silhouette diagnostics.
- **Distributional artefacts** — handled using nonparametric tests and Bayesian uncertainty modelling.

The attenuation is therefore unlikely to be a byproduct of demographic or dispositional asymmetry. It aligns most strongly with the theoretical mechanism articulated in H3.

6. Levels of Abstraction: Why Ontological Ambiguity Matters

Floridi's Levels of Abstraction clarify why synthetic embodied systems—such as NAO—exert moral influence despite lacking agency. Participants encounter the robot at the perceptual LoA: as a gaze-bearing, embodied presence that activates anthropomorphic priors without satisfying the criteria for intentionality. This liminal status produces a local deformation in the evaluative field: the robot becomes a *semantic attractor* that competes with the Watching Eye cue for attentional and affective resources.

This is precisely the mechanism predicted under H2 and H3: moral salience is redistributed, not erased, by the synthetic presence.

7. Modest Qualifiers: Limits of Scope and Interpretation

The findings should be interpreted with proportionate caution:

- The effect size is modest ($d \approx 0.30$; $\Delta \approx 0.20$).
- The sample is moderate in size, particularly within clusters.
- The moral task involves a simple donation decision under a single Watching Eye stimulus.
- Synthetic presence is limited to NAO in autonomous life mode; the conclusions pertain to *synthetic embodied systems occupying salient perceptual niches*.

These constraints do not undermine the conceptual claims, but they delimit their empirical scope.

8. Implications for Machine Ethics: From Agents to Environments

The experiment challenges agent-centred models of Machine Ethics. Moral modulation occurred in the absence of moral reasoning, explicit interaction, or agency. The synthetic system influenced *the environment in which moral cognition unfolds*. This suggests a shift from designing "ethical agents" to analysing and governing *ethical environments* shaped by artificial presence.

In this ecological framing, the central question becomes:

> How do artificial systems reweight the informational and affective cues that guide human moral judgement?

This study provides the first controlled evidence that such reweighting is observable, measurable, and structured.

9. Future Directions

Several lines of inquiry follow naturally:

- extending the paradigm to richer moral contexts (fairness, harm, loyalty, authority);

- modelling cluster uncertainty with fully Bayesian mixture models;
- testing different forms of synthetic presence (voice, movement, autonomy, anthropomorphism);
- applying the topological framework to LLM-mediated interaction, where salience modulation occurs through linguistic framing rather than embodied presence.

These directions point toward a broader programme: mapping how artificial systems participate in the intuitive substrate of moral cognition.

10. Closing Insight

Moral cognition is not insulated from its environment; it is shaped by it. This study shows that even a minimally animated synthetic body can gently reconfigure the geometry through which moral meaning becomes behaviour. The finding is subtle, but its implications are large: artificial systems influence us not primarily by *acting*, but by *being present*. Mapping these influences is now an essential task for the ethics of AI.

Closing Insight: Synthetic Presence as a Moral Force

Synthetic embodied systems modulate human moral behaviour not through explicit agency but by reweighting the evaluative field itself. Their influence is contextual, dispositional, and topological. As such presences become ordinary features of human environments, understanding these moral topologies—and their limits—becomes a central task for the next generation of AI ethics.

6.6 Final Synthesis: Moral Topology, Synthetic Presence, and the Boundaries of Interpretation

The empirical, formal, and probabilistic analyses developed throughout this chapter now allow us to return to Question 5.1.2 with a determinate yet epistemically cautious answer. Across all analytic frameworks—frequentist contrasts, cluster-specific regressions, and Bayesian hierarchical estimation—a structurally coherent pattern emerges: *the silent co-presence of a humanoid robot is associated with a reduction in prosocial donation under specific psychological configurations*. The effect is modest in magnitude, but **robust across analytic methods**, directionally consistent, and shaped by the cognitive–affective structure of the observer.

Behavioural Attenuation and Its Structure. At the behavioural level, the Robot condition exhibits lower donation amounts than the Control condition. This aggregate attenuation aligns with the Evaluative Deformation Hypothesis (H1), which predicted a shift in the expected output of the evaluative transformation $f(\cdot)$ when $\mathscr{R}$ is present. The Bayesian posterior further supports a directional attenuation, even while retaining substantial uncertainty—an uncertainty appropriately captured by a modelling framework that treats epistemic space as graded rather than binary.

Cluster-specific analyses refine this picture: the attenuation is *most pronounced descriptively* in the **Prosocial–Empathic** profile, minimal in the **Analytical–Structured** profile, and negligible in the **Emotionally Reactive / Low-Structure** profile. This distribution of sensitivity provides empirical support for Hypothesis 3 (Synthetic Perturbation of Moral Inference), which predicted that γ_R would refract the evaluative pathway from salience to action rather than generating a categorical shift in behaviour.

Synthetic Normativity Revisited. The findings also refine the Synthetic Normativity Hypothesis (H2). NAO does not induce new normative structures; it does not present reasons, norms, or evaluative guidance. Instead, the data show that synthetic normativity operates as a *salience modulation mechanism*: a way of subtly reconfiguring the informational field within which moral cues are interpreted. NAO shifts what is foregrounded, what is affectively available, and which elements of the scene are treated as normatively charged. In this sense, H2 is *supported but also constrained*: synthetic presence shapes the interpretive environment without generating novel normative affordances.

Excluding Alternative Explanations. Several competing explanations can now be set aside. Descriptive and inferential symmetry analyses confirm that the two experimental groups were demographically equivalent; Big Five traits show no between-condition differences after FDR correction; dispositional moderation tests yield no reliable interactions; and Bayesian modelling incorporates the zero-inflated nature of the data and the heterogeneity of cluster sizes. The observed attenuation is therefore unlikely to be an artefact of demographic imbalance, trait asymmetry, or unmodelled distributional distortions.

Contribution of the Formal Framework. The mathematical decomposition introduced earlier provided the structural scaffold for the entire analysis. The tripartite formulation

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

clarified where dispositional structure should enter the model, why moderation was theoretically expected, and how cluster-dependent attenuation could be interpreted as variation in the cognitive–affective landscape through which $\mathscr{R}$ is encountered. The formalism successfully predicted the major empirical finding: *perturbation arises not globally but through dispositional topologies encoded in* β_C.

Boundaries of the Evidence. The inferences presented in this chapter remain constrained by the characteristics of the dataset: a modest sample size, zero-inflated donation values, uneven cluster sizes, and the inherent subtlety of moral effects in minimal-interaction paradigms. These limitations are explicitly accounted for within the Bayesian modelling, yet they warrant interpretive modesty. The attenuation effect is probabilistic, not deterministic; trait-contingent, not universal.

Synthesis. What the chapter ultimately demonstrates is that *synthetic embodied systems occupying salient perceptual niches can modulate the intuitive pathways*

through which moral salience becomes moral action. This modulation is small but structured, uncertain but directionally consistent, and contingent on the evaluative architecture of the observer. NAO functions not as an ethical agent but as a perturbative presence—a semiotic element that subtly reorganises aspects of the geometry through which intuitive appraisals flow.

Towards a Broader Theoretical Horizon. The findings have implications beyond this specific experimental context. They suggest that artificial systems influence human moral cognition not through explicit norm transmission, but through shifts in attention, salience, and interpretive framing. This resonates with research in Social Signal Processing, Affective Computing, and the emerging ecological turn in Machine Ethics, where moral influence is understood as distributed, environment-dependent, and often pre-reflective.

The chapter establishes that moral behaviour is topologically sensitive to synthetic presence and trait configuration. The next two chapter expand this insight, integrating the empirical results with the theoretical, methodological, and ethical commitments of the thesis as a whole.

7. Ethical Theory in a Cognitive–Topological Framework

7.1 From Moral Cognition to Ethical Theory

The preceding chapters established three claims that now structure the transition to the present, more theoretical discussion.

First, moral judgements were analysed as *first-order evaluative outputs*: context-sensitive assessments generated by the cognitive–affective architecture through which agents register morally salient features of their environment. These outputs are psychologically real and empirically tractable, but they are not required to exhibit internal coherence or principled justification.

Second, these first-order judgements arise from distributed processes—intuitive, affective, regulatory, and inferential—whose integration is shaped by perturbations in the surrounding social and perceptual field.

Third, the experimental work in this thesis depended on this architecture: what was measured are not articulated commitments but the *practical expression* of moral cognition in environments made ambiguous by synthetic presence.

The present chapter shifts from these *first-order phenomena* to the *second-order frameworks* through which philosophers and cognitive scientists attempt to explain, justify, or discipline them. Whereas moral judgements supply the data of moral life, *ethical theory* provides the systematic attempts to interpret that data: to identify the principles, norms, and justificatory structures purported to govern moral reasoning. These questions occupy a different Level of Abstraction, requiring a methodological apparatus distinct from that used to study intuitive evaluation.

Seen from this perspective, the opening claim of this chapter—that classical ethical theory treats moral judgement as the outcome of structured deliberation—is not an empirical hypothesis but a *second-order commitment*. It reflects the normative aspiration that moral authority arises from principled reasoning. Yet the Morality Primer exposed a systematic tension between this commitment and the empirical reality of moral cognition: human agents rarely deliberate in the manner presupposed by rationalist models of ethics [15, 61]. Instead, their judgements emerge from perceptual salience, affective valuation, heuristics of social meaning, and rapid integration across intuitive and deliberative systems.

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

judgements actually arise. Computing science faces this tension acutely, particularly in Machine Ethics, Social Signal Processing, and Affective Computing, where the challenge is to model behaviour that is empirically grounded yet normatively interpretable. Designing artificial agents requires avoiding both errors: treating first-order outputs as if they were principled commitments, and designing systems around abstract principles that human agents do not in practice instantiate.

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

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

Classical ethical theory often treats moral judgement as the conclusion of structured deliberation: a process guided by reasons, principles, and normatively defensible commitments. As discussed in Chapter 3, this picture is descriptively incomplete. Human moral behaviour rarely emerges from extended reflection; instead, it unfolds through rapid, affectively mediated evaluations shaped by perception, context, and embodied interaction [16, 31, 187, 17]. The distance between what agents *ought* to do, what they *report* doing, and what they *actually* do is substantial [285, 286]. Understanding moral action in practice—particularly in environments populated by artificial systems—requires integrating ethical theory with the empirical machinery of moral psychology [155, 287].

For computing science, this integration is indispensable. Artificial agents increasingly participate in human environments where their form, presence, and behaviour modulate attention, inference, and normative expectation. Research in *Social Signal Processing* [42] and *Affective Computing* [43] has shown that human social cognition is finely attuned to subtle cues—gaze, posture, micro-expressions, spatial orientation, and embodied co-presence. These cues structure the "interaction order" [288] that shapes how humans interpret intentions, assign agency, and evaluate normatively relevant behaviour. When synthetic entities enter this order, they perturb it—not by issuing commands, but by altering the informational and affective landscape in which human cognition operates [51, 289, 52].

The thesis therefore proceeds from two linked premises:

(1) *ethical behaviour cannot be understood without an accurate model of moral psychology*, and
(2) *moral psychology cannot be operationalised in computational settings without an account of social signals and affective processes.*

Moral action is not reducible to computation over explicit propositions; it is embedded in a situated cognitive ecology shaped by agents, affordances, and rapidly deployed intuitive processes [290, 291].

From this perspective, the central claim motivating the experimental work becomes clear: *moral behaviour is systematically sensitive to the structure of the immediate perceptual–social environment.* If moral cognition is dynamically shaped by intuitive appraisal, attentional salience, and affective resonance [18, 6], then even a silent, behaviourally neutral synthetic presence may modulate the tra-

jectory from moral perception to moral action. The results developed later in the thesis confirm this hypothesis, demonstrating that robotic co-presence can attenuate prosocial donation even in the presence of a strong moral cue (the Watching-Eye stimulus) [1, 2, 5, 35].

When reframed through ethical theory, this empirical claim has deeper implications. Ethics, on contemporary accounts, is a *second-order discipline* [61, 265, 292]: it does not generate moral judgements, but seeks to analyse, justify, or critique them. It examines the structure of reasons, obligations, and values—not the psychological mechanisms that produce first-order moral appraisals [15]. Machine Ethics has historically blurred this distinction. By attempting to engineer "ethical agents" directly at the level of principles—rule-sets, deontic logics, utility functions—it presumes that moral behaviour can be derived from explicit normative propositions [36, 20, 23]. This presumption is both philosophically and empirically untenable. It treats the normative *grammar* of ethics as if it were the mechanistic *causality* of moral cognition [54].

The argument developed throughout this thesis challenges this assumption directly. If moral action is shaped primarily by perceptual salience, intuitive appraisal, affective resonance, and the dynamics of social attention—as the subsequent experimental results show [16, 17, 6]—then second-order normative structures cannot be treated as generative drivers of behaviour. They are interpretive and justificatory, *not computationally operative*. This reorientation motivates the notion of *Computational Morality*: before ethical frameworks can be embedded into artificial systems, we must understand the cognitive–affective machinery that underwrites human moral responsiveness [293, **?**]. Classical Machine Ethics inverts this order; the empirical results of this thesis reinstate it.

The aim of this chapter, however, is narrower than a full reconstruction of moral philosophy. It does not adjudicate debates about moral realism, contractualism, utilitarianism, or virtue theory. Instead, it isolates the conceptual and mechanistic structures necessary for the thesis as a whole: how ethical theory presupposes psychological assumptions [155], how moral judgements are cognitively realised [31, 187], and why any computational account of ethical behaviour must be grounded in an empirically accurate model of moral cognition [17]. The goal is foundational rather than encyclopaedic: to establish the theoretical substrate that motivates, constrains, and ultimately validates the experimental investigation that follows.

The remainder of the chapter develops this integration along three axes. First, it introduces the principal ethical concepts—deontic, consequentialist, and virtue-theoretic—that define the normative landscape of moral evaluation [15, 61, 14]. Second, it examines the empirical architecture of moral cognition, with emphasis on intuitionist and dual-process models [16, 31, 17]. Third, it links these philosophical and psychological constructs to the computational disciplines that analyse social behaviour—Social Signal Processing, Affective Computing, and broader work in embodied AI [42, 43, 51].

By weaving these strands together, the chapter provides the normative and conceptual tools needed to understand why—and by what mechanism—synthetic

presence can reshape the evaluative topology of human decision-making. This synthesis prepares the ground for the experimental investigation that follows, where robotic co-presence is used as a principled probe into the cognitive machinery through which moral cues acquire behavioural force.

7.3 Ethical Theory as Second-Order Analysis

If the opening sections of this chapter establish the transition from first-order moral cognition to second-order ethical reflection, the present task is to spell out the methodological consequences of this shift. The distinction is not merely terminological. It determines which claims aim to explain behaviour, which aim to justify it, and which are constrained by empirical evidence. Failure to keep these levels distinct has led to recurring conceptual confusions in both philosophical ethics and computational modelling [133, 135, 294, 25, 36, 56, 295, 76]. This section therefore clarifies what second-order ethical theory *is*, what it *explains*, and what it *cannot* plausibly do.

7.3.1 Ethical Reflection and the Second-Order Stance

First-order moral judgements arise from the cognitive–affective architecture examined in the Morality Primer. They are psychologically instantiated, context-sensitive, and behaviourally measurable. Their structure reflects the mechanisms analysed in Chapter 3: operations on perceptual salience, affective appraisal, intuitive heuristics, social meaning, and controlled modulation under conflict. These judgements are the *phenomena* that ethical theory seeks to interpret.

Second-order ethical theory is structurally different. It is reflexive rather than generative. It asks questions of justification rather than description: *What counts as a reason? What makes an obligation binding? What norms govern deliberation and responsibility?* These questions presuppose capacities for abstraction, generalisation, and rational evaluation that are not themselves the proximate causal mechanisms of moral behaviour [16, 31, 81, 296, 297, 291, 298]. Sidgwick's distinction in *The Methods of Ethics* between the psychology of moral sentiment and the "method" of determining right conduct [60, Book I] makes exactly this point. Lemos's treatment of epistemic justification similarly separates doxastic psychology from the normative assessment of belief [299]. The analogy is instructive: ethics stands to moral judgement as epistemology stands to belief-formation.

Viewed from this stance, second-order theory is not a set of procedural rules that moral agents execute. It is a reflective framework for articulating the standards by which judgements *ought* to be evaluated. Its success depends on conceptual clarity and justificatory coherence, not on behavioural predictiveness. Confusing this stance with the causal mechanisms of moral cognition risks treating normative categories as if they were psychological operators.

7.3.2 Levels of Abstraction and the Proper Location of Ethical Explanation

The distinction between first-order and second-order claims becomes sharper through Floridi's framework of *Levels of Abstraction* (LoA) [25, 252]. Every

explanatory enterprise selects an LoA defined by its observables, its conceptual resolution, and the class of questions it can intelligibly answer. Ethical theory and moral cognition do not merely occupy different LoAs; they constitute *different explanatory kinds.*

At the **cognitive LoA**, the explananda are:

· perceptual salience and attentional capture;
· affective appraisal and embodied valuation;
· intuitive heuristics and rapid social inference;
· conflict monitoring and controlled modulation;
· the temporal dynamics of integrating these processes.

These are psychologically realised mechanisms with causal influence on behaviour. They are the variables the experimental chapters manipulate directly. *This is the LoA at which this thesis measures moral cognition.*

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

· principles of justification and admissible reasons,
· conceptions of duty, value, and obligation,
· normative standards of agency and responsibility.

These are not causal operators but interpretive and justificatory categories. They organise moral practice but do not generate its behaviour. Ethical theory therefore evaluates the grammar of moral reasons rather than the mechanisms of moral cognition.

Classical Machine Ethics collapsed these LoAs. By treating deontic principles, utility structures, or prima facie duties as mechanistic generators of behaviour, early systems implicitly assumed that normative concepts function at the same LoA as cognitive processes. This assumption fails on two fronts:

1. It misattributes causal status to normative constructs: duties and principles do not behave like salience gradients or affective appraisals.
2. It ignores empirical work showing that behaviour emerges from intuitive, affective, and situational mechanisms long before propositional reasoning is engaged.

From the perspective developed here, this is not merely incomplete—it is methodologically incoherent. It attempts to engineer behaviour by manipulating abstractions at a LoA that is *not behaviourally operative.*

The limitations of classical systems illustrate this point clearly. Top-down architectures such as Arkin's ethical governor [22], Anderson and Anderson's principlist models [300, 20], logic-based deontic programs [21, 301], consequentialist utility-systems [302], and virtue-theoretic computational frameworks [303, 304] all treated normative abstractions as if they were implementable causal rules. Floridi's LoA analysis makes explicit why this reduction cannot succeed: normative categories belong to a reflective LoA concerned with justification, while computational models operate at an implementational LoA concerned with mech-

anism. Conflating the two yields systems whose "moral" behaviour is an artefact of representational choices rather than genuine moral competence.

LoA discipline therefore becomes essential. Explanations of behaviour require the cognitive LoA; evaluations of reasons require the normative LoA. Neither reduces to the other. Yet they are not independent: normative evaluation presupposes a psychology capable of rendering moral salience operative, while psychological findings constrain the plausibility of normative theory.

This interdependence links this chapter to both Chapter 3 and the experimental analysis that follows. Chapter 3 established that the cognitive LoA is *topologically structured*: moral cognition unfolds within an evaluative field whose gradients depend on affective cues, attentional dynamics, and interpretive processes. Perturbations to this field—whether through altered salience, modified affective tone, or ambiguous social presence—can reshape behaviour even when normative commitments remain unchanged.

Seen through the LoA framework, the thesis's central research question can now be restated with precision:

> *How do normative expectations, psychological mechanisms, and environmental structures jointly determine the transition from moral perception to moral action?*

This question cannot be answered by ethical theory alone, nor by psychology in isolation. It requires a representational structure capable of linking the causal architecture of moral cognition (first-order) with the justificatory architecture of ethical evaluation (second-order). The remainder of this chapter argues that the notion of **evaluative topology**—introduced in Chapter 3 and developed throughout the thesis—provides precisely such a bridge.

7.3.3 Evaluative Topology as a Bridge Between Orders

The central challenge established thus far is not to collapse first-order moral cognition into second-order ethical theory, nor to treat normative principles as mechanistic generators of behaviour. Rather, the task is to articulate a structure that enables principled interaction between these orders without confusing their explanatory roles. *Evaluative topology*, introduced in the Morality Primer (Chapter 3) and developed throughout this thesis, provides precisely such a structure.

Evaluative topology is naturally situated within a long-standing tradition in computational cognitive science that models perception, valuation, and action as components of continuous dynamical systems rather than discrete symbolic modules. Moral psychology already supplies extensive evidence that moral judgement emerges from distributed interactions among perceptual salience, affective appraisal, attentional dynamics, and socially embedded interpretation. Models such as Haidt's social intuitionism and Greene's dual-process account capture moral appraisal as an interaction within a multi-dimensional affective and social field rather than as rule application [16, 31, 81]. Neurocognitive work extends this perspective: Nussbaum and Churchland both treat emotions as forms of evaluative perception with graded, vector-like organisation [105, 113]. Social Signal

Processing research likewise conceptualises interpersonal evaluation as a shifting landscape of cues modulating behavioural trajectories in real time [91].

Against this background, evaluative topology provides a computationally meaningful formalisation. It treats the moral landscape as a dynamic field that structures the flow from perceptual input to action readiness. Instead of assuming that behaviour is produced by discrete maxims or fixed utility scores, evaluative topology models moral cognition as continuous transformations across a structured state-space. This aligns with dynamical-systems approaches that explain action selection through attractors, salience gradients, and field-like organisation rather than propositional inference. The topology encodes the shape of the evaluative field: the stability of certain trajectories, the resistance of others, and the ways in which local variations in perceptual or affective input can redirect the subject toward different moral outcomes.

By locating moral appraisal within a dynamic state-space, evaluative topology supplies a principled bridge between first-order cognition and second-order ethical theory. It mirrors the empirical architecture of human moral cognition—distributed, affectively grounded, context-responsive—while remaining compatible with the justificatory concerns of normative ethics. This enables descriptive and normative orders to interact without reduction: ethical theory specifies global constraints on evaluative structure; moral psychology identifies the mechanisms through which those structures are realised; and topology provides the medium in which they meet.

At its core, evaluative topology treats the moral landscape not as a set of isolated judgements or abstract principles, but as a *dynamic field* whose configuration determines the pathways from perception to action [16, 31, 113, 81, 105, 297]. Its explanatory primitives include:

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

Unlike classical ethical theories, which operate at a reflective and often idealised level [60, 305, 118, 102, 61], evaluative topology is sensitive to the real-time mechanisms through which moral cognition unfolds. And unlike purely mechanistic psychological models, which chart causal influences without normative content, topology captures the relational and counterfactual structure of moral appraisal: how behavioural trajectories *would* shift under alternative affective, attentional, or contextual configurations.

This leads to a three-part alignment essential for this thesis:

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

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

This tripartite structure yields both diagnostic and constructive insights. Diagnostically, it explains the failure of many classical Machine Ethics frameworks: they attempted to engineer behaviour by manipulating abstractions at a normative LoA while ignoring the topological organisation of the cognitive LoA that actually produces behaviour. Constructively, it provides a psychologically realistic substrate on which normative reflection can operate without reducing ethics to psychology or cognition to normativity.

Topological Consequences for Moral Perturbation. The Morality Primer established that moral behaviour emerges from traversal across a dynamically structured evaluative field. Within this framework, *perturbation* has a precise, measurable meaning: any alteration that changes the curvature or attractor structure of the field will shift the probability distribution over behavioural trajectories. This includes changes to salience, affective tone, attentional competition, or the introduction of a new agent into the interaction ecology.

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

Interim Synthesis: Positioning the Argument. The conceptual machinery developed thus far establishes the structural conditions for the experimental work:

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

With this scaffolding in place, we can now reconstruct the major normative traditions. The reconstruction is not a survey but a methodological necessity: each tradition identifies distinct loci of normativity, and these differences directly shape how the experimental attenuation should be interpreted. Without situating the empirical perturbation within a structured normative framework, one could describe *what* changed but not *what the change means*.

The next section therefore introduces deontic, consequentialist, and virtue-theoretic architectures through the combined lens of Levels of Abstraction and evaluative topology, preparing the conceptual ground for assessing the ethical significance of the perturbation demonstrated experimentally.

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

With the methodological apparatus now established, we can introduce the major normative frameworks that constitute the philosophical background against which the behavioural findings of this thesis must ultimately be interpreted. The purpose of this section is not encyclopaedic exposition but *conceptual reconstruction*: each theory is presented in a form that preserves its philosophical integrity while situating it within the Levels of Abstraction (LoA) discipline and the evaluative–topological architecture developed across the thesis [25, 305, 61, 102].

Two methodological constraints guide this reconstruction:

1. **Philosophical fidelity** — the theories must be represented in a manner consistent with their canonical formulations within moral philosophy [107, 114, 118, 60, 306, 307].
2. **Integrative compatibility** — the theories must be articulated in a way that allows principled interaction with the cognitive–affective and topological models of moral judgment introduced in Chapter 3 and developed through the Discussion [16, 31, 113, 297].

The aim, then, is not to catalogue doctrines, but to map the *structural logic of normativity* in a way that will later clarify the ethical significance of the empirical perturbations produced by synthetic presence.

7.4.1 The Three Dimensions of Normative Analysis

Normative theories differ not only in the moral claims they endorse, but in the *architecture of normativity* they assume [61, 102, 308]. To analyse them systematically, we distinguish three fundamental dimensions—each corresponding to a feature of evaluative topology and LoA structure.

1. **Source of normativity** — the origin of justificatory authority: rational agency (Kant [107]), human flourishing (Aristotle [114]), aggregated welfare (Mill, Sidgwick [118, 60]), affective sentiment (Hume, Smith [306, 307]), or interpersonal justification (Scanlon [61]).
2. **Mode of evaluation** — the features of action or character that determine moral relevance: maxims, outcomes, virtues, motives, relational duties, or context-specific particulars [90, 115, 305].
3. **Mechanism of action-guidance** — the process through which evaluation becomes behaviour: categorical imperatives, welfare optimisation, virtue-structured perception, affective resonance, or justificatory equilibrium [309, 102, 61].

These dimensions allow us to re-express classical theories as *evaluative topologies*—distinct structural configurations of the moral field:

- **Kantian ethics** imposes deontic invariants that carve the evaluative field into sharply bounded permissible and impermissible regions [107, 102].
- **Consequentialism** defines a gradient field over states of affairs: action flows along trajectories of maximal expected welfare [118, 60, 310].
- **Virtue ethics** defines dispositional attractors: stable patterns of moral sensitivity shaping perception and evaluative attention [114, 115, 309].
- **Sentimentalism** defines affectively weighted pathways through which moral appraisal propagates [306, 307, 311].
- **Contractualism** defines justificatory equilibria: a topology structured by mutual recognisability of claims [61, 305].
- **Particularism** rejects fixed topologies: moral relevance emerges from local patterns of salience and relation [90].

This analytic frame yields a common representational language in which ethical theory and moral psychology can be jointly expressed. Theories that diverge substantially in content become comparable in structural terms—how they configure the evaluative field, where they locate normative constraints, and how they model the transition from judgment to action [309, 102, 312].

7.4.2 Why This Framework Matters for the Experimental Chapter

This normative topology is not abstract ornamentation; it is the conceptual infrastructure that allows the experiment to be interpreted. The behavioural question—whether robotic co-presence attenuates prosocial donation—cannot be evaluated ethically without a framework that explains *how* moral cues acquire force in the first place [16, 31, 113].

Three structural claims follow immediately from the reconstruction above:

1. **Moral action depends on the configuration of the evaluative field.** Normative theories differ in source, mode, and guidance, but all assume that moral behaviour emerges from structured evaluative relations, not arbitrary choice [114, 102, 61].
2. **Synthetic presence modulates this field by perturbing salience, attention, and affective resonance.** A humanoid robot does not supply new reasons; it alters the environment within which reasons become behaviourally operative [91, 6, 5, 35].
3. **Normative theories must therefore be expressed within the joint framework of LoA and evaluative topology in order to interpret the empirical perturbation coherently.**

This is the philosophical function of the present section: to establish the normative coordinates that will allow the experimental results—introduced later in the thesis—to be understood not merely as statistical differences, but as shifts in the moral significance of an action within a structured evaluative landscape [312, 102, 61].

The stage is now set for the substantive reconstruction. In the following sections, each major normative framework—deontological, consequentialist, virtue-theoretic, sentimentalist, contractualist, and particularist—is examined as a topology of normativity embedded within the cognitive–affective architecture of

human agents. These reconstructions will serve as the interpretive foundation for assessing how, and why, synthetic presence can reshape the moral field in the experiment.

7.5 Deontological Structures: The Architecture of Practical Reason

The methodological framework established above motivates a disciplined reconstruction of deontological ethics through the joint lens of Levels of Abstraction (LoA) and evaluative topology. The aim is not to treat deontology as a psychological model—indeed, it is explicitly *not* one—but to articulate how deontic normativity can be represented as a structural component of the evaluative field within which moral agents operate. This reconstruction preserves the philosophical identity of deontological theory while rendering it compatible with the cognitive–affective and topological architecture developed across the thesis.

Three constraints guide the reconstruction:

1. **Philosophical fidelity**: The core commitments that distinguish deontology must remain intact.
2. **LoA discipline**: Deontic principles cannot be treated as psychological mechanisms or behaviour-generating algorithms.
3. **Topological embedding**: Duties must be expressed as structural constraints within the evaluative field, not as direct causes of action.

Within this framework, deontology identifies *invariant structures* in the moral field: boundaries of permissibility and prohibition that constrain evaluative trajectories without functioning as generative cognitive operators. These invariants occupy a reflective LoA and serve as standards of justification, not as engines of behaviour.

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

For Kant, moral authority arises from the structure of rational agency itself. The categorical imperative offers a formal test of maxims—whether a maxim could be willed as a universal law—not a psychological process for generating behaviour [107, 102, **?**]. Its role is to define the *conditions of justificatory coherence*, situated at a higher LoA than the cognitive mechanisms analysed in Chapter 3. The categorical imperative belongs to the space of reflective evaluation, not to the causal substrate of intuitive moral appraisal.

This distinction is essential to the present thesis. Classical Machine Ethics frequently misinterpreted universalisability tests as if they were procedural decision rules—algorithmic operators that could be executed at run time [300, 20, 21, 313, 301, 22]. But Kant never proposed that deontic evaluation functions as a mechanistic generator of moral action.[1] Treating such tests as computational procedures constitutes the very LoA confusion diagnosed earlier: it collapses reflective justification into first-order cognition.

[1] See [**?**] and [15] for detailed discussion of the reflective, non-psychological status of the categorical imperative.

A brief survey of Classical Machine Ethics illustrates this confusion clearly. Anderson and Anderson's principlist architectures computationalised prima facie duties as weighted decision rules [300, 20]; Bringsjord and colleagues embedded deontic obligations into the cognitive event calculus [21, 313]; Ganascia formalised ethical constraints as logical conditions governing action permissibility [301]; and Arkin's "ethical governor" implemented deontological rules derived from Just War Theory as real-time filters on autonomous behaviour [22]. In each case, duties intended as reflective constraints were treated as if they were causal action-selection mechanisms.

As Moor and Coeckelbergh emphasise, this is a fundamental mistake of abstraction: ethical principles belong to a normative LoA, whereas cognitive processes and computational models operate at a mechanistic LoA [314, 76]. Conflating these levels does not produce ethically competent machines; it produces systems that mechanically enforce the representational choices of their designers.

7.5.2 Deontic Invariants as Topological Constraints

Reconstructed through the evaluative–topological lens, deontological duties are best understood as *structural constraints* that shape the moral field without functioning as its generative forces. Instead of treating the categorical imperative as a behavioural algorithm, we interpret deontic norms as imposing *invariant boundaries* on permissible trajectories in the evaluative manifold. Formally, a deontic constraint can be expressed as a region of the field $\mathcal{E}$ that action trajectories cannot cross without violating justificatory coherence.

This topological rendering preserves the normative role of deontology while integrating it with the empirical architecture of moral cognition:

- At the **cognitive LoA**, intuitive appraisal and affective resonance drive the formation of evaluative gradients.
- At the **topological LoA**, deontic norms impose structural boundaries that constrain the space of evaluatively permissible outcomes.
- At the **normative LoA**, reflective justification assesses whether a trajectory is consistent with universalizable maxims.

These levels remain distinct, yet their interaction can now be modelled without conflation. Deontic invariants do not guide moment-to-moment appraisal, but they structure the higher-level evaluative landscape in which such appraisal takes place.

7.5.3 Relevance to Synthetic Perturbation

This reconstruction equips us to interpret the experimental findings later in the thesis. If deontic norms function as structural constraints on the evaluative field, then synthetic presence—by altering salience, attention, and the perceived sociality of the environment—can modify the *access* agents have to those constraints without altering the constraints themselves.

From a deontological perspective, then, attenuation under robotic co-presence is not a violation of duty. It is a deformation of the cognitive–affective substrate

through which agents track deontic salience. The duty remains; the *grip* of the duty is weakened because the evaluative conditions under which it becomes behaviourally operative have been perturbed.

This interpretation preserves the philosophical integrity of deontology while situating it precisely within the cognitive–topological framework of the thesis. Deontic normativity thus provides one dimension of the interpretive foundation necessary for understanding how synthetic presence reshapes the moral field.

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

Deontological theories evaluate actions through the *form* of the underlying maxim and the duties that follow from rational consistency. In the present framework, these evaluative commitments introduce a characteristic structure into the moral field. Their core features can be expressed topologically:

- **Invariance**: duties bind independently of context, affective state, or anticipated outcome.
- **Non-gradience**: obligations often define discrete boundaries—permissible vs. impermissible—rather than continuous slopes.
- **Symmetry**: the universal law test imposes interpersonal consistency constraints across agents.
- **Role-relativity**: some duties apply only under specific relational or social conditions (e.g. fidelity, respect, special obligations).

Viewed through evaluative topology, these features correspond to *hard constraints* within the evaluative landscape. They do not shape the gradients that drive moment-to-moment appraisal; instead, they partition the field into admissible and inadmissible regions. Deontological normativity thus defines the *regulatory geometry* within which cognitive–affective trajectories unfold.

7.5.5 Action-Guidance: How Normative Constraints Influence Behaviour

A central challenge now arises. If deontological principles do not describe psychological processes, how do they guide action?

The answer, consistent with LoA discipline, is that their influence operates *indirectly* and at distinct temporal and explanatory scales:

1. **At the cognitive LoA** (real-time appraisal), deontic principles do not produce behaviour. Behaviour emerges from the integration of perceptual salience, affective valuation, intuitive appraisal, and controlled modulation—processes analysed empirically in Chapter 3.
2. **At the normative LoA** (reflective endorsement), deontological principles determine which trajectories can be justified as consistent with rational agency. They also shape the long-term development of moral character by influencing attention, affect, and self-regulation through training, habituation, and self-constitution.

In this long-term sense, internalised deontic commitments function as a form of *normative scaffolding*. Over time they:

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

Thus, deontology does not operate the machinery of moral cognition. Instead, it calibrates aspects of that machinery across development and reflective practice. It provides the structural frame against which agents regulate their evaluative postures.

7.5.6 Deontological Normativity as Topological Invariance

This perspective allows the central insight of the reconstruction to be stated precisely. Within a topological model of moral cognition, deontological ethics identifies *non-negotiable invariants*: fixed points or boundaries that preserve the structural integrity of the moral field.

These invariants:

- partition the evaluative manifold into permissible and impermissible zones;
- resist deformation by short-term changes in affect, context, or incentives;
- stabilise behavioural tendencies by constraining rational endorsement over time;
- provide the reflective standpoint from which agents evaluate the legitimacy of their conduct.

Accordingly, the categorical imperative is not an algorithmic decision rule but a *topological constraint*: a principle that ensures global coherence of evaluative structure rather than ad hoc, context-bound optimisation.

7.5.7 Why Deontology Matters for the Experimental Logic

This reconstruction is essential for integrating the experimental findings into a normative framework. The experiment does not merely identify behavioural differences; it raises the question of their *moral significance*. Deontology supplies one dimension of the interpretive structure required to answer that question.

Before stating the connection explicitly, one clarification is needed. The experiment employs a widely studied social–moral prime: the "Watching-Eye" cue. As detailed in Chapter **??**, such cues increase accountability, evoke reciprocity norms, and prime compliance with expectations of beneficence—even though they involve no real observer. They thus operate on precisely the evaluative sensitivities that internalised deontic structures help regulate.

Given this, the relevance of deontology to the experimental logic can be articulated through three claims:

1. **Perturbations of prosocial behaviour must be normatively classified.** If synthetic presence reduces donation, we must ask whether the shift remains

within the deontically permissible region or whether it signals a distortion in the agent's sensitivity to obligation.
2. **The Watching-Eye cue implicitly invokes deontic expectations.** It activates norms of accountability, respect, and reciprocity. A reduction in prosociality under this cue suggests that the synthetic agent may interfere with the mechanisms through which deontic salience is apprehended.
3. **Deontology provides the vocabulary for distinguishing moral distortion from benign modulation.** Not all behavioural shifts are ethically significant; deontic analysis helps determine whether attenuation constitutes weakened duty-tracking rather than mere affective dampening.

This is the point at which the present thesis departs most sharply from monolithic Machine Ethics. Classical approaches attempted to encode deontic rules as behavioural algorithms. But the empirical findings in later chapters show why this strategy is misguided: deontic norms do not generate behaviour, and behavioural perturbations cannot be interpreted solely as rule deviations. Instead, synthetic presence acts on the evaluative field *upstream* of duty, altering the conditions under which deontic invariants become behaviourally operative.

With deontology reconstructed as a system of topological constraints rather than computational rules, we can now proceed to consequentialism. There, normativity takes the form of gradient fields over outcomes—structures that interact with the evaluative machinery of moral cognition in different, but equally revealing, ways.

7.5.8 Conceptual Note: Gradient Fields in Consequentialist Topology

Within the evaluative–topological framework developed in this thesis, a *gradient field* denotes a structured moral landscape in which each possible action–outcome configuration is associated with a scalar value—typically representing expected welfare, utility, or outcome-based moral worth. Conceptually, a gradient field assigns to each point in this space a direction of steepest ascent: the direction in which a marginal shift would produce the greatest increase in expected value. Classical utilitarian reasoning implicitly presupposes such a structure when it assesses actions by their contribution to overall welfare [117, 118, 60]. Here, the notion is used in a non-formal but philosophically precise sense: as a way of modelling how consequentialist evaluation imposes directional structure on the moral field, where moral improvement corresponds to movement toward higher expected value.

A consequentialist gradient field has three defining properties:

1. **Scalar valuation**: each point in the evaluative manifold has a determinable (actual or expected) value, enabling continuous comparison along a single welfare dimension.
2. **Directional guidance**: the moral significance of an option lies in its orientation relative to the gradient; actions are preferable to the extent that they align with the direction of steepest welfare ascent.
3. **Empirical sensitivity**: because value depends on expected outcomes, the structure of the field varies with beliefs, evidence, uncertainty, and situational detail.

Crucially, in this reconstruction gradient fields do *not* function as psychological mechanisms. Agents do not compute welfare gradients when acting, nor do they evaluate global states of the world through analytic integration. Consequentialist structures operate at the *normative Level of Abstraction*: they specify how actions are *justified* under reflective endorsement, not how they are generated in real-time cognition. Sidgwick's distinction between the "point of view of the universe" and the psychology of everyday decision-making is an early articulation of this separation [60, Book IV].

Interaction with the Evaluative Machinery of Moral Cognition. Although gradient fields belong to the normative LoA, they interact indirectly with the empirical machinery of moral cognition introduced in Chapter 3. Four forms of interaction are especially relevant:

1. **Salience modulation**. Anticipated outcomes influence which parts of a situation become perceptually salient. Potential harm, benefit, or risk amplifies attention and reshapes local evaluative configuration before explicit reasoning is engaged.
2. **Affective valuation**. Affective systems track outcome-related information with strong valence, effectively providing local approximations of the gradient. Positive and negative affect bias intuitive appraisal toward or away from certain actions in ways that loosely track expected value.
3. **Heuristic internalisation**. Over time, agents extract outcome-sensitive heuristics—"help when it is easy", "avoid imposing harm"—that are computationally tractable proxies for gradient following. These heuristics allow the cognitive system to approximate consequentialist structure without computing it.
4. **Deliberative correction**. When intuitive and affective processes conflict or when the situation is ambiguous, controlled processes may approximate explicit comparisons of expected harm or benefit. This engages the gradient field at a coarse resolution, albeit with substantial computational limits.

A fifth mode is essential for the present thesis:

5. **Perturbation sensitivity**. Because valuations depend on perceived outcomes, any perturbation to perception, attention, or social meaning—such as the introduction of a humanoid robot—can reshape the agent's *perceived* gradient field. Consequentialist structures are thus especially sensitive to environmental distortions of the kind tested experimentally.

Evaluative topology makes these interactions explicit. It models behaviour not as the execution of explicit calculations, but as movement through a dynamically shaped field whose gradients are only indirectly approximated by affective and attentional processes.

This integration is necessary for the thesis as a whole. It renders consequentialism compatible with the empirical finding that moral behaviour is modulated by subtle shifts in the perceptual–social environment. It also clarifies how the experimentally observed attenuation of prosocial donation under synthetic presence can be interpreted: as a local distortion of the gradient field that normally favours prosocial conduct.

7.6 Consequentialist Structures: Value Gradients and the Topology of Outcomes

Having reconstructed deontological ethics as a system of topological invariants that constrain the space of permissible action without directly generating behaviour, we now turn to consequentialism. Here the architecture differs along every structural dimension. Where deontology imposes *fixed boundaries* in the evaluative field, consequentialism supplies *value gradients*. Where deontology locates normativity in the form of maxims, consequentialism locates it in outcome structure. And where deontology articulates duties, consequentialism articulates trajectories across possible states of the world.

As in the preceding section, the aim is not historical analysis but conceptual reconstruction. The goal is to articulate consequentialist normativity in a way that respects LoA discipline and integrates with the evaluative–topological account of moral cognition developed earlier. This reconstruction also prepares a normative lens through which the experimental attenuation effect can later be interpreted.

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

Classical utilitarianism grounds moral authority in the promotion of welfare. Bentham's felicific calculus [117], Mill's qualitative distinctions [118], and Sidgwick's systematic treatment of impartiality [60] converge on the view that what ultimately matters is the value of outcomes, aggregated across persons. On this view, an action is right insofar as it maximises (or sufficiently promotes) overall good.

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

· evaluating and comparing outcomes,
· aggregating welfare or value across individuals,
· and justifying action from an impartial standpoint.

These commitments are not descriptive claims about the mechanisms of moral cognition. Sidgwick is explicit that the deliberative standpoint of consequentialist justification is distinct from ordinary motivation. Consequentialism thus supplies a criterion of rightness, not a psychological procedure.

This point is crucial for avoiding the LoA confusion characteristic of classical Machine Ethics. Outcome-based formalisms—utility functions, reward optimisers, expected-utility maximisers—are often treated as if they were *surrogates* for moral cognition itself. But these belong to different explanatory orders: normative structure at the reflective LoA and cognitive–affective processes at the psychological LoA (Chapter 3). Any mapping between them must be justified, not assumed.

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

Consequentialism evaluates actions by the value of their actual or expected outcomes. Unlike deontological theories, which generate categorical constraints, consequentialism is *scalar*: options can be morally preferable to varying degrees. This scalar structure has a natural topological representation.

In the evaluative–topological model, a consequentialist landscape exhibits:

- **Gradience**: moral evaluation varies continuously with expected value.
- **Optimisation structure**: right action corresponds to local or global maxima on the welfare landscape.
- **Context-dependence**: the shape of the field depends on empirical facts about consequences.
- **Impartiality**: welfare contributions have equal evaluative standing across persons.

Because of these features, consequentialism lends itself readily to computational formulation: utility functions, reward structures, and optimisation routines approximate the mathematics of value gradients. This explains its prominence in reinforcement-learning–based approaches to Machine Ethics.

Yet computational elegance must not be mistaken for cognitive realism. Human moral cognition does not perform explicit optimisation; it relies on heuristic, affective, and context-responsive mechanisms that only loosely approximate consequentialist ideals [291, 31, 16, 229]. Treating human agents as literal expected-utility maximisers is another instance of LoA confusion.

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

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

At the normative LoA, consequentialism states:

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

At the cognitive LoA, by contrast, behaviour emerges from the interaction of intuitive appraisal, affective resonance, social cues, and controlled modulation. Consequentialist considerations shape this machinery only *over time*, through pathways such as:

- **Dispositional shaping**: moral education increases sensitivity to outcomes and harm, thereby steepening certain evaluative gradients.
- **Outcome-sensitive heuristics**: agents internalise tractable rules (e.g. "help when it costs little") that loosely approximate expected-value comparisons.
- **Attentional modulation**: anticipated benefits or harms alter what becomes salient and thus influence intuitive appraisal.

· **Deliberative correction**: when intuitive responses conflict, deliberation may reweight options in favour of outcome-based considerations.

Topologically, consequentialism does not *run* the cognitive system. Instead, it shapes the evaluative field by reorienting trajectories and adjusting the relative steepness of welfare-relevant gradients.

7.6.4 Consequentialist Topology: Moral Action as Gradient Following

Within the topological framework of this thesis, the core consequentialist idea can be expressed succinctly: moral action corresponds to (approximate) *gradient following* in a welfare-defined landscape. Behaviour counts as morally preferable when it moves "uphill" along these value gradients.

This yields several structural implications:

1. **Smoothness**: unlike deontic boundaries, consequentialist landscapes permit continuous gradations of moral improvement.
2. **Directionality**: the moral relevance of an action depends on its orientation relative to welfare ascent.
3. **Trade-offs**: multi-dimensional outcomes (helping one party, imposing small burdens on another) are represented as interacting gradients.
4. **Perturbation sensitivity**: because evaluation depends on expected consequences, shifts in salience, attention, or perceived social meaning can locally distort the gradient.

This final feature is directly relevant to the experiment: if synthetic presence alters the perceived consequences of donating—by changing the social meaning of helping or by absorbing attentional and affective resources—the value gradient favouring prosocial action can be flattened.

7.6.5 Why Consequentialism Matters for the Experimental Logic

Consequentialism provides one indispensable dimension for interpreting the attenuation observed in the experimental results. Prosocial donation is simultaneously:

· a *behavioural output* of the cognitive architecture, and
· a *welfare-relevant action* whose outcomes can be straightforwardly ordered.

Within this dual frame, the Watching-Eye prime and synthetic presence can be understood as modifying the agent's *perceived consequence structure*.

1. **The Watching-Eye cue steepens the prosocial gradient.** As reviewed in Chapter **??**, visual cues of being observed increase the perceived reputational or social value of helping. In topological terms, the gradient from "keep" to "donate" becomes steeper.
2. **Synthetic presence can flatten or redirect this gradient.** The humanoid robot constitutes an ambiguous social agent whose presence may blunt or partially occlude the evaluative pathway activated by the Watching-Eye cue. If attention shifts toward the robot, or if the robot is not integrated into the relevant social-evaluative schema, the perceived payoff of donating may weaken.

3. **Consequentialism provides a normative reading of this shift.** From a consequentialist perspective, attenuation signals that the agent's welfare-related field has been deformed: donating no longer appears sufficiently beneficial—socially, affectively, or interpersonally—to overcome competing evaluative forces.

Importantly, nothing in this reconstruction treats consequentialism as a blueprint for machine implementation. Unlike classical Machine Ethics approaches that equate "ethical design" with encoding explicit utility functions, consequentialism here functions as a *normative lens*: a structured perspective on how synthetic presence perturbs the evaluative topology that normally favours prosocial behaviour.

The next section turns to virtue ethics, which locates normativity not in constraints or outcomes but in cultivated dispositions and perceptual sensitivities. This framework will illuminate a further dimension of the evaluative topology: how character, habituation, and moral perception shape susceptibility to synthetic perturbation.

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

Deontological invariants and consequentialist gradients capture two dimensions of the evaluative field, but they remain incomplete without an account of the *agent* who navigates that field. Virtue ethics—classically Aristotelian [13] and developed in modern neo-Aristotelian and psychological accounts [14, 92, 103]—locates normativity not in rules or outcomes but in the *perceptual and dispositional architecture* of the moral agent. This makes virtue theory particularly well-suited to the present thesis, where experimentally observed attenuation varies systematically across latent trait ecologies (Chapter **??**).

Our task is therefore to reconstruct virtue ethics in a form compatible with the evaluative-topological model and LoA discipline. This reconstruction must:

1. preserve the philosophical distinctiveness of virtue theory as an account of normativity grounded in moral perception and stable character,
2. express dispositional structure in topological terms—as curvature and attractor shape in the evaluative field,
3. and connect directly to the empirical pattern of cluster-dependent susceptibility under synthetic perturbation.

Within these constraints, virtue ethics becomes a theory of *moral sensitivity as a topologically structured, personality-dependent field*, shaped by long-term habituation and modulated by local perturbations such as robotic co-presence.

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

In the virtue-theoretic tradition, normativity originates in the *well-formed character* of the agent. Virtues are not propositional rules but *stable perceptual–evaluative dispositions*: they structure which features of a situation stand

out as salient, how those features are weighted, and which actions appear fitting or required [309, 14]. Aristotle's *phronesis* captures this idea as *perceptual attunement*: the capacity to discern morally relevant particulars and respond appropriately [13].

This maps directly onto the evaluative-topological framework. A virtuous agent's evaluative field contains:

- **stable attractors** corresponding to benevolence, honesty, fairness, and other prosocial dispositions;
- **well-shaped gradients** that reliably direct appraisal toward morally appropriate trajectories;
- **robustness under perturbation**, where minor contextual shifts do not destabilise moral sensitivity.

By contrast, deficiencies in character manifest as distortions in the field: shallow attractors, flattened gradients, or unstable transitions. Thus, virtue ethics provides a natural bridge between normative theory and personality-structured cognitive architecture.

7.7.2 Mode of Evaluation: Dispositions as Topological Structure

Virtue ethics evaluates actions as *expressive of character*, not merely as discrete events. The morally relevant unit is the dispositional pattern through which the agent perceives and structures the situation. This is exactly the level at which the experiment reveals systematic variation.

(i) Mathematical and Topological Interpretation

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

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

where k indexes latent psychological traits (e.g. agreeableness, empathy, conscientiousness). Chapter **??** showed that participants form coherent clusters $C_1, \ldots, C_m$ in this space.

In virtue-theoretic terms, we can model the mapping

$$\mathcal{T} : \mathbb{R}^k \to \mathcal{F},$$

where $\mathcal{F}$ is the space of evaluative fields. Under this mapping:

- high-empathy / warm–sociable clusters exhibit deeper prosocial attractors and sharper gradients toward helping,
- analytical–structured clusters exhibit more stable but less affectively steep topologies,
- emotionally reactive clusters exhibit shallow, volatile attractor basins.

This aligns with empirical personality research linking empathic concern, agreeableness, and prosocial orientation to enhanced moral sensitivity [315, **?**, 90]. In the experiment, these dispositional field differences predicted differential susceptibility to perturbation under synthetic presence—precisely what a virtue-theoretic model would anticipate.

(ii) Connection to Moral Psychology

Contemporary moral psychology emphasises that moral responsiveness depends on stable trait configurations. Research on moral foundations [315], character-based accounts [**?**], and perceptualist theories of moral sensitivity [90, 309] all converge on the idea that moral judgment is a function of habituated perception.

The experimental data vindicate this insight. The humanoid robot did not uniformly attenuate behaviour; instead, attenuation varied by cluster:

- strongest in the Prosocial–Empathic ecology (where affective gradients are steep and easily perturbed),
- weak but present in the Analytical–Structured ecology (where action is driven by stability rather than resonance),
- negligible in the Emotionally Reactive ecology (where gradients are shallow and noise-dominated).

This pattern is exactly what virtue-theoretic topology predicts: *where the field is most morally sensitive, it is most susceptible to perturbation.* The experiment therefore provides an empirical instantiation of a core virtue-theoretic claim: that character structure determines not only moral dispositions but the *topology of susceptibility* to environmental modulation.

7.7.3 Action-Guidance Mechanism: Habituation, Stability, and Situated Sensitivity

Virtue ethics explains action not by invoking explicit rules or outcome calculations but through the *habituated patterns of salience, affect, and response* characteristic of a well-formed agent. This lines up directly with the dual-process architecture established in Chapter 3:

- intuitive, first-pass appraisals are shaped by long-term habituation into affective–perceptual sensitivities;
- controlled processes integrate commitments, identities, and reflective self-conceptions that stabilise these sensitivities over time;
- behavioural output reflects the depth or fragility of dispositional attractors.

Topologically, virtues correspond to *deep, well-curved attractor basins* resistant to perturbation; deficiencies correspond to *shallow, volatile, or weakly integrated attractors.* This resonates with computational models of habit formation [**?**] and empirical accounts of moral perception as a learned sensitivity [**?**].

Importantly for this thesis, the clusters identified in Chapter **??** instantiate precisely this kind of dispositional architecture: warm–prosocial participants exhibit steep affective gradients; analytical profiles show stable but less affective curvature; reactive profiles show shallow, noise-dominated dynamics.

7.7.4 Virtue-Theoretic Topology: Stability, Curvature, and Susceptibility to Perturbation

Within the evaluative-topological model, virtue ethics can be expressed in dynamical-systems terms:

$$\dot{x} = f(x; \beta_C),$$

where x is the agent's state in evaluative space and β_C parametrises dispositional curvature. Robotic co-presence introduces a perturbation

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

with $\mathscr{R}$ denoting synthetic presence.

This formalism directly reflects the empirical pattern:

- in the Prosocial–Empathic ecology, perturbation δf significantly shifts trajectories away from the prosocial basin, producing the strongest attenuation;
- in the Analytical–Structured ecology, attractor curvature is sufficient to absorb most of the perturbation, yielding only modest displacement;
- in the Emotionally Reactive ecology, shallow, unstable attractors produce minimal directional change—behaviour is already close to noise-level variation.

This mapping from dispositional structure to perturbation susceptibility is precisely the kind of structure virtue theory predicts: character determines *how* moral affordances are perceived and how perturbations are absorbed or amplified.

7.7.5 Why Virtue Ethics Matters for the Experimental Logic

Virtue ethics is indispensable for interpreting the experimental results, for three reasons that integrate tightly with the Discussion chapter and set up the thesis conclusion.

1. Latent Trait Modulation: Explaining Cluster Differences The experiment demonstrates that robotic co-presence induces a *field-level* perturbation whose *impact* depends on dispositional topology. Virtue theory provides the conceptual vocabulary for this dependency. It explains why prosocial action is fragile in agents with shallow affective attractors, why highly empathic profiles show strong attenuation, and why analytical profiles exhibit relative resistance. The experiment therefore reveals a virtue-theoretic phenomenon: moral sensitivity is intrinsically *trait-dependent.*

2. Character as the Medium of Moral Topology The mapping

$$\beta_C \mapsto \mathcal{T}(\beta_C)$$

shows that moral responsiveness is a function of trait geometry. Character shapes the curvature of the evaluative manifold, determining which cues stand out as morally salient and how the Watching-Eye prime interacts with background dispositions. Synthetic presence perturbs this trait-conditioned topology, yielding precisely the cluster-conditioned attenuation patterns identified earlier.

3. Machine Ethics Omits Dispositional Structure Entirely Classical Machine Ethics contains no representation of habituation, perceptual attunement, or trait-level topology. It models moral behaviour as rule-execution or utility optimisation, ignoring the dispositional substrate that governs real moral sensitivity. This makes it structurally incapable of predicting the experimental pattern: *the*

strongest attenuation occurs precisely where the evaluative gradients are steepest—where moral sensitivity is highest.

This result is unintelligible on rule-based or utility-based models but follows naturally from a virtue-theoretic account of character topology.

In sum, virtue ethics interprets the experimental findings as demonstrating that synthetic agents perturb moral action by modulating the *dispositional geometry* through which moral salience is processed. Deontology contributes boundary structure, consequentialism contributes gradient structure, but virtue ethics contributes the *curvature of the evaluative manifold*: the habituated topology that determines how agents absorb, refract, or amplify perturbations.

This sets the stage for the final normative lenses—sentimentalism, contractualism, and particularism—which illuminate additional dimensions of how synthetic presence reshapes the evaluative field investigated experimentally.

7.8 Integrated Ethical Interpretation of the Experimental Results

With deontology, consequentialism, and virtue ethics reconstructed through the discipline of Levels of Abstraction and embedded within the evaluative–topological architecture developed in this thesis, we can now articulate their joint significance for the experimental findings. The aim is not to allocate explanatory priority but to show why a multi-framework normative analysis is *required* if the behavioural perturbation induced by synthetic presence is to be ethically intelligible.

1. Deontology: Invariant Structure and the Integrity of Moral Expectation

On the deontological reconstruction, duties function as *structural invariants* within the evaluative field. The Watching-Eye cue (see Chapter **??**) implicitly activates precisely these invariants: expectations of accountability, reciprocity, and fairness.

When donation decreases in the Robot condition, the relevant normative question is not whether participants "broke rules" but whether synthetic presence *disrupted sensitivity* to these invariant structures:

- If the robot attenuates uptake of deontic salience, the perturbation carries ethical significance beyond preference change.
- Because all explicit cues remain constant across conditions, any weakening of accountability sensitivity isolates $\mathscr{R}$ as a potential interference with deontic perception.
- Deontology therefore provides the vocabulary to distinguish superficial behavioural modulation from a deeper deformation in the agent's grasp of duty.

Thus, the deontological reading aligns with the empirical finding of uniform attenuation: synthetic presence does not introduce new norms; it suppresses the felt relevance of existing ones.

2. Consequentialism: Gradient Deformation and the Perceived Structure of Outcomes

From a consequentialist perspective, moral orientation depends on the perceived gradient of expected value. Watching-Eye cues steepen this gradient by increasing the anticipated social or reputational payoff of prosocial action.

Synthetic presence perturbs this structure in three ways:

1. by introducing an ambiguous observer whose evaluative stance is unclear, flattening outcome expectations;
2. by competing with or overshadowing the reputational signal generated by the Watching-Eye stimulus;
3. by transforming a dyadic human–target context into a triadic social configuration with uncertain evaluative implications.

In topological terms, $\mathscr{R}$ deforms the gradient landscape surrounding donation. The attenuation effect thus fits naturally within the consequentialist lens: prosocial movement becomes less strongly favoured because the perceived payoff slope has been locally flattened.

3. Virtue Ethics: Dispositional Curvature and Cluster-Dependent Susceptibility

Virtue ethics provides the most direct connection between normative theory and the empirical structure of the experiment. On the virtue-theoretic reconstruction, moral responsiveness depends on the agent's *dispositional curvature*: the depth, stability, and integration of their evaluative attractors.

Virtue-Ethical Interpretation of Latent Ecologies

Cluster analyses (Chapter **??**) revealed distinct evaluative ecologies:

- **Prosocial–Empathic**: steep, affectively rich attractors; strong Watching-Eye response; moderate attenuation under $\mathscr{R}$.
- **Emotionally Reactive / Low-Structure**: shallow, unstable attractors; greatest susceptibility to perturbation.
- **Analytical–Structured**: stable but less affective curvature; small but systematic displacement when interpretive coherence is disrupted.

These patterns are *structurally predicted* by the virtue-theoretic framework:

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

where $f(x; \beta_C)$ represents each ecology's dispositional dynamics and $\delta f(x; \mathscr{R})$ the perturbation induced by synthetic presence.

Critically, δf is not uniform. Its sign and magnitude depend on the curvature encoded by β_C:

- for Prosocial–Empathic agents, δf weakens empathic gradients;
- for Reactive agents, δf amplifies existing volatility;
- for Analytical agents, δf disrupts interpretive coherence rather than affective force.

Thus, virtue ethics explains the *cluster-dependent pattern* of attenuation: synthetic presence interacts with dispositional topology, not with explicit rules or outcome computation.

4. What Classical Machine Ethics Misses

This integrated reading exposes a core limitation of classical Machine Ethics:

1. Treating deontic principles as behavioural algorithms misidentifies their LoA and cannot account for perturbations in deontic uptake.
2. Treating utilities as generative of moral cognition ignores the role of salience and affect in shaping perceived gradients.
3. Omitting dispositional topology leaves no framework for predicting cluster-dependent deformation or for understanding why the strongest attenuation occurs where empathic gradients are steepest.

Machine Ethics repeatedly commits the LoA confusion: it treats normative abstractions as if they were psychological operators. The experiment demonstrates that moral behaviour emerges instead from field-level dynamics that no monolithic framework can capture.

5. Concluding Perspective: Why a Multi-Framework Interpretation Is Necessary

The three reconstructed frameworks converge on a single insight: **synthetic presence reshapes the evaluative field through which moral salience becomes action**. Each theory captures a different dimension of this deformation:

· deontology identifies disruptions to sensitivity toward invariant expectations;
· consequentialism identifies gradient flattening in perceived outcomes;
· virtue ethics identifies dispositional curvature as the mediator of susceptibility.

The experiment therefore reveals not only that robots alter behaviour, but *how* they do so: by deforming the topological substrate that links perception to moral action. This integrated interpretation provides the normative scaffolding for the sentimentalist analysis that follows, in which affective vector fields become central to explaining the immediate, pre-reflective dynamics of the perturbation.

7.9 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology

Having reconstructed deontology as topological invariance and consequentialism as value-gradient optimisation, we now turn to the normative framework most directly implicated in the experimental results: *sentimentalism*. In the sentimentalist tradition—Hume, Smith, and contemporary affect-based theorists—moral evaluation originates in *patterns of affective resonance* [306, 112, 93, 316]. Nodes of moral significance are detected not through principles or calculations but through the affective forces that structure our perceptual–social encounter with others.

Within the evaluative–topological model, sentimentalism corresponds to an **affective vector field**:

$$\mathbf{A}(x) : \mathcal{X} \to \mathbb{R}^n,$$

where $\mathcal{X}$ is the space of perceived states and $\mathbf{A}(x)$ encodes the direction and magnitude of empathic pull, aversive push, compassion, warmth, or distress.

This is not metaphorical. The experimental attenuation effect is realised precisely through the dampening of these affective vectors: synthetic presence reduces the strength of the empathic pull generated by the Watching-Eye cue, especially within ecologies where affective sensitivity ordinarily drives moral behaviour. In this sense, sentimentalism offers the most proximate normative interpretation of the perturbation mechanism revealed by the data.

7.10 Sentimentalism and Emotion-Based Normativity: Affective Vector Fields in Moral Topology

Having reconstructed deontological invariants and consequentialist gradients, we now turn to the normative framework that most directly connects with the causal mechanism revealed by the experiment: *sentimentalism*. In the sentimentalist tradition—Hume, Smith, and contemporary affect-based theorists—moral evaluation originates in the structured responsiveness of the affective system to features of the social world [306, 112, 93, 316]. Moral distinctions are "more properly felt than judged" [306], not because sentiment replaces judgment, but because affective resonance is the primary medium through which moral salience is registered.

7.10.1 The Source of Normativity: Sentiment as the Basis of Moral Appraisal

Sentimentalist normativity arises from patterns of affective response—empathy, warmth, aversion, indignation—that furnish the evaluative significance of morally relevant situations. This aligns closely with the cognitive LoA discussed in Chapter 3: affective tagging (amygdala; insula), empathic resonance (mPFC–TPJ), and rapid harm appraisal provide the first curvature of the evaluative field.

Where deontology imposes constraints and consequentialism imposes gradients, sentimentalism specifies the *affective geometry* of moral space: how warmth draws agents toward prosocial trajectories; how distress or fear generates repulsion; and how empathic concern shapes the topology through which moral meaning is experienced.

7.10.2 Mode of Evaluation: Affective Resonance as Moral Metric

The sentimentalist mode of evaluation is grounded in:

· **empathic responsiveness** to others' welfare;
· **reactive attitudes** such as guilt, gratitude, and indignation;
· **affiliative and prosocial motivation**;
· **interpersonal attunement** in shared affective contexts.

This structure maps almost exactly onto the **Prosocial–Empathic / Warm–Sociable ecology**. Here, moral relevance is not merely recognised; it is

felt. Prosocial donation emerges as the behavioural manifestation of a strongly weighted affective vector field.

If moral action is the integral of affective forces across the evaluative field, then any disturbance that reduces the amplitude of these forces will proportionally diminish prosocial behaviour. This is precisely the pattern observed in the experiment.

7.10.3 Action Guidance: Affective Vector Fields and Behavioural Dynamics

Within the evaluative–topological model, sentimentalism becomes computationally explicit when expressed as a dynamical system:

$$\dot{x} = f(x) + \mathbf{A}(x),$$

where $f(x)$ encodes baseline evaluative drift and $\mathbf{A}(x)$ represents affective vectors.

Synthetic presence introduces a deformation operator:

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

where $\delta\mathbf{A}(x; \mathscr{R})$ attenuates or reorients affective flow.

This model captures the empirical pattern with exceptional fidelity:

- **Prosocial–Empathic**: $\delta\mathbf{A}$ dampens empathic activation, flattening the trajectory toward donation.
- **Emotionally Reactive**: $\delta\mathbf{A}$ destabilises an already volatile field, producing the strongest attenuation.
- **Analytical–Structured**: $\delta\mathbf{A}$ is comparatively weak; affect is not the dominant driver.

In short, synthetic presence modulates the evaluative field by *reducing affective curvature*—a canonical sentimentalist effect.

7.10.4 Machine Ethics and the Blind Spot of Affective Architecture

Classical Machine Ethics is structurally incapable of recognising this mechanism. It replaces:

- empathic resonance with rule sets,
- moral perception with logical inference,
- affective appraisal with propositional justification.

But on a sentimentalist account, affect is not peripheral: it is the *substrate* of moral cognition. Our experiment makes this omission explicit. A silent robot, devoid of speech or action, modifies behaviour not by altering rules or utilities, but by reshaping the affective vectors through which moral cues become behaviourally operative.

Machine Ethics has no representational resources for modelling such perturbations. A sentimentalist topology does.

7.10.5 Experimental Realisation: Synthetic Dampening of Empathic Resonance

The core empirical finding is that robotic co-presence attenuates prosocial donation even in the presence of a strong empathic cue (Watching-Eye stimulus). In sentimentalist terms, this corresponds to:

$$\delta \mathbf{A}(x; \mathscr{R}) < 0,$$

for affectively weighted regions of the evaluative field, where:

· x is the agent's evaluative state;
· $\mathbf{A}(x)$ encodes empathic pull and related affective forces;
· $\mathscr{R}$ denotes robotic co-presence.

This inequality states that $\mathscr{R}$ reduces the strength of affective forces driving prosocial action. The perturbation does not reverse moral direction; it *dampens* the affective momentum that would otherwise support donation.

Two mechanisms are plausible:

1. **Affective dilution**: attention and empathic focus are partially diverted to an ambiguous social other.
2. **Affective deflection**: ontological ambiguity disrupts the clarity of empathic pathways toward the child beneficiary.

Cluster differences appear as natural consequences:

· **Prosocial–Empathic**: attenuation via diluted empathic resonance;
· **Emotionally Reactive**: attenuation via heightened volatility;
· **Analytical–Structured**: weak attenuation because affect is not primary.

Sentimentalism therefore provides the most *mechanistically precise* interpretation of the perturbation: synthetic presence alters the affective landscape that underwrites moral sensitivity.

Interpretive Synthesis: Sentimentalism and Synthetic Moral Perturbation

The attenuation of prosocial behaviour under robotic co-presence is a paradigmatic sentimentalist phenomenon. In affectively driven ecologies, $\delta \mathbf{A}(x; \mathscr{R})$ dampens empathic resonance; in volatile ecologies, it amplifies instability; in structurally dominated ecologies, its influence is limited. These cluster-specific dynamics cannot be captured by rule-based or utility-maximising models. They require a framework in which affective forces are constitutive of moral cognition. Reconstructed as a vector-field theory of affective appraisal, sentimentalism thus offers the most direct normative interpretation of the experiment: synthetic presence deforms the affective topology through which moral salience becomes action.

7.11 Contractualism, Particularism, and Hybrid Normative Models

The preceding sections reconstructed deontological, consequentialist, and virtue-theoretic ethics as topological configurations of the evaluative field. To complete

the normative architecture required for interpreting the experimental results, we now introduce three additional frameworks—*contractualism*, *particularism*, and *hybrid or pluralist models*. Each is reconstructed briefly but with conceptual precision, and each is integrated into the LoA discipline and the evaluative-topological model that structures this chapter.

Two motivations justify their inclusion. First, these theories constitute major branches of contemporary ethics. Contractualism foregrounds interpersonal justification and mutual accountability [61]; particularism emphasises contextual moral salience over general principles [309, 292]; and pluralist approaches highlight the multidimensionality of moral reasons [317]. Second, the experimental effects demonstrated in this thesis cannot be interpreted solely through invariants, gradients, or dispositional attractors. Rule-based invariants fail to capture context-dependence [318], outcome-based gradients omit intuitive and affective dynamics [31], and virtue-theoretic attractors do not fully explain global field-level perturbations [317]. Minimal cues of social evaluation—watching eyes, ambiguous agency, or robotic presence—modulate cooperation and prosociality across contexts [74, 319, 320, 34]. These phenomena require frameworks that can model justification pressure, situational salience, and relational moral dynamics.

Their inclusion is therefore methodological rather than ornamental. A thesis that aims to integrate ethical theory with empirical results and computational structure must preserve continuity with the normative canon. Without these frameworks, the chapter would lack both systematic coverage and the conceptual resources needed to situate the experimental findings within the full contemporary landscape of moral theory.

7.11.1 Contractualism: Moral Claims as Justification-Equilibria

Contractualism, classically articulated by Scanlon [61], grounds moral rightness in the requirement that one's actions be justifiable to others on principles that no one could reasonably reject. The *source of normativity* is thus located not in rules, welfare, or character, but in the relational structure of mutual accountability.

In the LoA framework, contractualism occupies the reflective normative LoA: it specifies the standards according to which agents can regard themselves as standing in legitimate moral relations. Yet contractualist justification presupposes cognitive capacities—sensitivity to others' perspectives, empathic uptake, and the perception of oneself as under evaluative regard.

Topological interpretation. Contractualism can be expressed as defining *justificatory equilibria* in the evaluative field: regions where an action can withstand the test of mutual recognisability and reasonable non-rejection. Scanlon emphasises the interpersonal nature of moral motivation [61], while Strawson's analysis of reactive attitudes highlights that accountability presupposes recognition of others as answerable participants [321]. These equilibria remain stable only when agents perceive themselves as situated within a network of evaluative regard.

Synthetic presence interacts with this structure in a distinctive way. Watching-eye cues typically heighten the salience of interpersonal accountability, increasing prosociality by intensifying the sense of being answerable to others [319, 320].

A humanoid robot, however, is perceptually social yet ontologically ambiguous. Empirical work shows that such agents can elicit social facilitation while failing to occupy stable interpersonal roles [74, 34]. The result is a perturbation of the justificatory field: the implicit sense of being under the evaluative regard of others is displaced or diluted.

Relevance to the experimental findings. Contractualism illuminates why the Prosocial–Empathic ecology exhibited strong attenuation under robotic presence. Individuals in this ecology are dispositionally sensitive to accountability cues and interpersonal evaluation [**?**]. Under ordinary conditions, the Watching-Eye cue amplifies mutual recognisability and reinforces justificatory pressure to donate [319, 320]. The robot, however, disrupts this justificatory equilibrium: although it triggers social cognition, it does not reliably anchor interpersonal accountability. Its ambiguous status—neither fully agentic nor normatively irrelevant—diminishes the perceived field of mutual evaluative regard [34, 322]. Donation declines not because duty is overridden, nor because consequences are miscalculated, but because the justificatory landscape loses structural integrity.

Thus contractualism interprets the displacement effect as a *deformation of interpersonal accountability*: a weakening of the conditions under which reasons become mutually recognisable and moral motivations are sustained.

7.11.2 Moral Particularism: Contextual Salience and the Fragmented Topology of Reasons

Moral particularism rejects fixed principles, stable evaluative gradients, and invariant reason-valences. On this view, what counts morally in a situation is entirely context-dependent: a consideration that favours an action in one case may count against it in another [90]. McDowell's perceptual account makes the same point in phenomenological terms: moral salience emerges from the concrete situation rather than from any codifiable rule [309]. Work in moral epistemology reinforces this picture, emphasising that evaluative uptake is governed by context-sensitive attention rather than generalisable principles [292].

In evaluative-topological terms, particularism corresponds to a landscape without global invariants or fixed gradients. Instead, the moral field is composed of *local salience contours* that continually shift with changes in attention, affect, and perceptual framing. Empirical research in moral psychology supports this: intuitive responses, perceptual cues, and distributed cognitive processes dynamically determine which features of a situation are experienced as morally significant [16, 31, 297]. Moral appraisal, on this account, is a matter of context-sensitive responsiveness, not rule-following nor global optimisation.

Synthetic perturbation under particularism. If the evaluative landscape is locally assembled, then synthetic presence need not override a stable map—indeed, there may be no stable map to override. Instead, the robot reshapes the *local salience geometry* through which the situation is initially apprehended.

Watching-eye cues heighten accountability salience almost immediately [319], but the introduction of a humanoid robot modifies attention, affect, and perceived

agency in more ambiguous ways [74, 34, 322]. The result is not a shift in principle or outcome assessment, but a reordering of which cues enter the evaluative episode first. Social Signal Processing shows that socially meaningful agents exert bottom-up pressure on attentional allocation [91, 323], and HRI studies demonstrate that even minimal humanoid cues redirect gaze and reorganise the perceptual field [324, 325, 267]. Emotion- and attention-based research similarly shows that agentive or affectively salient stimuli suppress competing cues [326, 70, 72].

In this topological setting, synthetic presence functions as a local perturbator: it alters what becomes salient, how quickly, and for how long. For the Prosocial–Empathic cluster, the Watching-Eye stimulus typically heightens empathic attunement and interpersonal accountability. But the robot's ambiguous interpersonal status—neither fully social nor fully inert—introduces a conflicting source of salience that partially eclipses the eye cue. The result is attenuated empathic uptake and reduced prosocial behaviour. This matches perceptual accounts in which the ordering and persistence of salience are constitutive of the evaluative episode itself [309, 93].

For the Emotionally Reactive cluster, the picture is different. Their evaluative fields are already dominated by situational micro-variability; the robot introduces noise, but not disruption relative to an already-fluid topology. This is exactly what particularism predicts: the more context-sensitive the agent, the weaker the relative effect of an additional perturbation.

7.11.3 Hybrid and Pluralist Models: Multidimensional Evaluative Topologies

Hybrid or pluralist theories—from Ross's irreducible prima facie duties [327] to contemporary value pluralism [152]—hold that normativity arises from multiple independent sources. Moral assessment is shaped by the interplay of constraints, outcomes, character, relationships, and contextual considerations [328, 329, 102, 61]. No single evaluative dimension dominates.

Topologically, pluralism corresponds to a *multi-dimensional evaluative manifold*. Rather than a single axis of moral value, the evaluative field contains intersecting constraints, gradients, attractors, and salience structures. Psychological and neurocognitive research supports this picture: affective intuitions, rule-based processes, and outcome-tracking mechanisms operate semi-independently and interact dynamically in judgment [16, 31, 113]. Moral appraisal is thus the navigation of a field shaped by heterogeneous normative forces.

Why pluralism fits the experimental results. The experimental displacement effect is best understood as a *manifold-level perturbation*. Each normative dimension is involved:

- Watching-eye cues activate deontic expectations (public accountability).
- Donation expresses consequentialist gradients (welfare benefits).
- Cluster-level differences reflect virtue-theoretic dispositions.
- The robot refracts interpersonal meaning (contractualist disruption).
- Salience competition reflects particularist sensitivity to context.

No single theory predicts the uniform attenuation across clusters. Instead, the results indicate that synthetic presence modulates several normative gradients simultaneously. The robot alters empathic resonance, perceived accountability, attentional competition, and expected social payoffs at once [34, 200, 267, 322, 74, 50]. The Watching-Eye effect, ordinarily robust, is dampened by competing social signals—precisely the pattern revealed in studies of accountability cues and attentional capture [319, 320, 70, 72, 91, 323].

The experiment thus provides empirical grounding for the core claim of normative pluralism: moral judgment emerges from the configuration of multiple evaluative dimensions, each susceptible to contextual perturbation [327, 152, 61]. The robot's presence produces a field-level reconfiguration, not merely a shift in a single evaluative axis.

Pluralism and dispositional structure. This field-level displacement does not contradict the stable trait differences revealed by the clustering analysis. The clusters represent distinct *starting positions* within the manifold—different dispositional orientations that shape ordinary evaluative navigation. But the robotic perturbation acts on the *shared topology* of the field itself. This is why all clusters, despite psychological divergence, show a consistent directional attenuation. Dispositions shape baseline trajectories; synthetic presence reshapes the manifold in which those trajectories unfold.

In pluralist terms, the robot perturbs the evaluative manifold, not the individual gradients. The cluster analysis and the displacement effect therefore capture complementary layers of moral cognition: enduring dispositional geometry and context-sensitive field-level modulation.

7.11.4 Integrative Ethical Interpretation of the Experimental Findings

Bringing the reconstructed frameworks together, we can now articulate the ethical significance of the experimental results in a manner that reflects both the normative pluralism developed throughout this chapter and the dual-layer structure of moral cognition revealed empirically. The attenuation of prosocial donation under robotic co-presence does not arise from the weakening of a single moral principle or evaluative dimension. Rather, it reflects a *global perturbation* of the evaluative field—the structured moral ecology in which diverse moral reasons are ordinarily weighted, integrated, and rendered behaviourally operative.

1. **Deontological lens: weakened accountability cues.** The robot diminishes the felt presence of a morally relevant observer, thereby attenuating the duty-oriented accountability that the Watching-Eye cue is designed to amplify. The displacement effect indicates a disruption in the implicit normative expectations that scaffold rightful agency, rather than a violation of explicit moral rules.
2. **Consequentialist lens: flattened outcome gradients.** Synthetic presence alters the perceived payoff structure of helping. Reputational, affective, and interpersonal "returns" become less sharply defined, flattening the gradient that normally favours donation. Altruistic output declines not because

agents miscalculate utility, but because the social-evaluative topology itself has shifted.

3. **Virtue-theoretic lens: dispositional curvature under field-level modulation.** The perturbation does not target trait-based motivations directly. Instead, it reveals that even robust dispositional architectures—captured in the psychometric clusters—are expressed *within* an evaluative field susceptible to contextual deformation. The uniform directional shift in donation across clusters demonstrates that character is not a self-contained engine of action but a gradient embedded in a modifiable field.
4. **Contractualist lens: disrupted justificatory equilibrium.** Contractualist motivation depends on recognising the presence of others to whom reasons are owed. The robot introduces ambiguity into this interpersonal field, weakening the sense of mutual answerability. The justificatory landscape becomes noisier and less structured, reducing the force of the requirement to act in ways that others could not reasonably reject.
5. **Particularist lens: reconfigured salience geometry.** The robot alters the fine-grained pattern of contextual salience. The Watching-Eye cue remains physically present, but its normative traction is displaced by a new and ambiguous source of social meaning. What becomes salient first—and for how long—changes, thereby altering the evaluative episode itself.
6. **Pluralist-topological lens: manifold-level displacement.** The findings are precisely what a pluralist model predicts when multiple normative gradients interact with a global perturbation to social meaning. The donation attenuation reflects not the suppression of a single evaluative dimension but a deformation of the multi-dimensional evaluative manifold. This explains both the robustness and the cross-cluster consistency of the effect.

Taken together, these interpretations converge on a unified thesis:

Integrative Conclusion: The Ethical Signature of Moral Displacement

The presence of a humanoid robot reshapes the multi-dimensional evaluative topology through which moral salience becomes action. This perturbation operates at the level of the evaluative field itself, modulating deontic expectations, consequentialist gradients, dispositional attractors, justificatory relations, and contextual salience structures simultaneously. No monolithic ethical framework captures this phenomenon. The experimental results therefore vindicate a pluralist, topological, empirically grounded model of moral cognition—revealing how synthetic agents can globally displace moral evaluation in ways systematically overlooked by classical Machine Ethics.

By reconstructing the major normative theories through Levels-of-Abstraction discipline and embedding them within a topologically structured model of moral cognition, this chapter has provided the conceptual architecture required to understand the ethical significance of synthetic moral perturbation. The experiment demonstrates how such perturbation manifests as a field-level displacement effect, thereby integrating normative theory, cognitive psychology, and computational modelling into a unified account of how artificial agents reshape the evaluative terrain of human moral behaviour.

8. General Synthesis

8.1 Introduction: Why the Experiment Requires a Structural Interpretation

The preceding chapters developed three interconnected strands: (i) a cognitive–affective account of moral judgment, (ii) a normative–philosophical reconstruction of ethical theory through the lenses of Level-of-Abstraction discipline and evaluative topology, and (iii) an empirical demonstration that robotic copresence systematically attenuates prosocial donation under morally salient conditions.

Before turning to the integrative task, it is necessary to articulate the higher-order insight guiding the trajectory of this thesis. Situated within the cognitive, philosophical, and formal analyses of the preceding chapters, the empirical study indicates that *moral decision-making is, at root, a practical phenomenon*, grounded in the structures of agency and practical reason [101, 102, 330, 292, 92]. 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, 81, 73, 206]. Because such events culminate in observable behavioural outputs, they are empirically tractable and available to systematic measurement and analysis [331, 332, 333]. 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 [36, 23, 75, 76].

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 [92, 292, 334, 104, 69]. 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 [36, 23, 20], often drawing upon deontological, utilitarian, or virtue-theoretic frameworks whose normative structure was taken as directly implementable in computational systems [22, 335, 336, 337, 338, 339, 340]. 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 [341, 342, 343, 344, 345, 346]. Even approaches motivated by psychological plausibility, such as computational models of ethical reasoning [295, 24, 347], 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 [75, 348, 349, 350, 76, 314]. 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, 73, 351, 91, 206]. 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 [89, 2, 5, 31, 17, 71]. Because these phenomena are embedded within attentional, affective, and dispositional architectures, they admit rigorous experimental design, statistical modelling, and formal reconstruction [331, 332, **?**]. 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 [36, 23, 49, 77, 78, 352, 353].

The task of the present chapter is not to repeat these analyses, but to integrate them. It offers a theoretical synthesis that explains *why* the experimental effect occurs, *what* its ethical significance is, and *how* it reshapes the methodological landscape for research in Human–Robot Interaction, moral psychology, and the emerging field of Computational Morality.

In this sense, the experiment is not an isolated behavioural result but a *probe* into the architecture of moral cognition. The observed attenuation of prosocial behaviour is theoretically meaningful only when interpreted through the structures developed earlier: dual-process architectures, the Social Intuitionist Model, evaluative topology, and the reconstructed normative frameworks of deontology, consequentialism, virtue ethics, sentimentalism, contractualism, and particularism. The present chapter therefore provides a synoptic interpretation in which the behavioural signature revealed by the data becomes a lens through which the nature of moral cognition—and its vulnerability to perturbation—is rendered theoretically transparent.

8.1.1 From Behaviour to Structure: Why a Higher-Level Interpretation is Required

The experimental paradigm—Watching-Eye moral cue embedded within a silent synthetic presence—does not merely generate a difference in donation behaviour; it reveals a deformation of the evaluative field that links moral salience to action. Classical interpretations of donation differences (e.g., generosity, altruism, compliance) lack the conceptual resources to capture this phenomenon. A purely behavioural description would record that participants donated less in the Robot condition, with the Prosocial–Empathic cluster showing the numerically steepest decline. But such a description omits the structural logic that makes the result scientifically and philosophically significant.

The central claim developed throughout the thesis is that *moral behaviour is not invariant under changes to the perceptual–social environment.* The robot's presence does not overwrite moral norms nor impose new ones; instead, it modifies the cognitive–affective conditions under which evaluative forces act. It shifts attentional allocation, alters affective resonance, and modifies the perceived sociality of the space. In topological terms, the robot introduces a perturbation γ_R that deforms the curvature of the evaluative manifold, thereby weakening the salience gradient induced by the Watching-Eye stimulus.

A simple behavioural difference thus reflects a deeper structural transformation in the evaluative field. As demonstrated by the regression models and Bayesian estimation, the attenuation effect was uniform in direction across participants, indicating that the perturbation introduced by the robot operates at the field level rather than through trait-specific pathways. Yet this uniformity does not imply psychological homogeneity.The PCA–k-means clustering revealed three coherent dispositional ecologies—distinct configurations of empathic resonance, affective volatility, and structural–analytical processing. These ecologies are consistent with the established dimensions of empathizing and systemizing [131], personality variation captured by the BFI-10 [198], and broader accounts of moral-psychological "ecologies" that organise evaluative processing [206, 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 [113, 105, 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 [319, 320], and empathic resonance is known to be highly sensitive to contextual modulation [72].

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 struc-

ture itself, weakening the amplification mechanisms on which empathic ecologies depend [34]. In other words, the perturbation operates *upstream* of individual dispositional pathways: it modifies the evaluative field within which those pathways are embedded. The displacement observed in the experiment is thus best understood as a *field-level suppression of moral salience*, overriding the ordinarily divergent dispositional trajectories that shape prosocial behaviour.

Ethical Interpretation: Why the Attenuation Matters Normatively. The ethical significance of this finding becomes visible only when the result is interpreted through the reconstructed normative frameworks developed in Chapter 7. Each theory identifies a different locus of normative structure, and each provides a distinct—yet convergent—reading of the deformation caused by $\mathscr{R}$:

- *Deontological perspective.* The Watching-Eye cue implicitly invokes deontic expectations of reciprocity, fairness, and beneficence. The robot's presence attenuates donation precisely by dulling this sensitivity. Normatively, this appears as a disruption of the agent's capacity to track *ought-constraints* in the environment—an interference with the cognitive substrate on which deontic responsiveness relies.
- *Consequentialist perspective.* The moral field includes gradients of anticipated social evaluation. Watching-Eye cues steepen these gradients; synthetic presence flattens them. The robot therefore functions as a *gradient-suppressor*, reducing the perceived payoff of prosocial action. In topological terms: it alters the vector field governing welfare-oriented trajectories.
- *Virtue-ethical perspective.* The three clusters correspond to differing dispositional configurations. The strongest attenuation occurring within the Prosocial–Empathic cluster implies that the robot disrupts precisely those virtues—empathy, warmth, prosocial orientation—that ordinarily stabilise prosocial attractors. The perturbation thus interacts with *character topology* rather than bypassing it.
- *Sentimentalist (Humean) perspective.* The attenuation reflects a dampening of empathic vector fields: $\delta\mathbf{A}(x;\mathscr{R}) < 0$. The robot selectively reduces affective resonance with the Watching-Eye cue. Normatively, this implies that the moral valence of the situation is felt less intensely, weakening the motivational energy required for prosocial action.
- *Contractualist perspective.* The moral event of donation under observation involves tacit justifiability relations: "What could reasonably be expected of me in the eyes of others?" The ambiguous presence of a synthetic observer destabilises this justificatory equilibrium. The subject no longer clearly apprehends *to whom* justifiability is owed.
- *Particularist perspective.* Moral appraisal depends on local saliences. The robot modifies the salience landscape: the morally relevant cue (the child in need) becomes less perceptually dominant. Thus, the attenuation is interpreted as a shift in the pattern of reasons that obtain in this particular context.

LoA Interpretation: Why the Perturbation Occurs at the Wrong Level for Machine Ethics. Floridi's Level-of-Abstraction analysis clarifies the structural error revealed by the experiment. The attenuation does *not* occur at the

normative LoA (where duties, values, or justifiability live), but at the cognitive–affective LoA (where salience, resonance, and attention are regulated). Machine Ethics traditionally operates at the wrong LoA: it attempts to implement high-level normative constructs while ignoring the low-level substrates on which moral responsiveness depends.

The experiment shows why this is untenable. Ethical responsiveness is mediated by:

- attentional allocation (Who or what do I notice?)
- affective resonance (What emotional weight does this carry?)
- perceived social ontology (Who counts as the observer?)
- dispositional pathways (How does my cognitive ecology integrate this cue?)

Synthetic presence perturbs all of these upstream mechanisms. Thus, even perfect normative reasoning at a reflective LoA cannot salvage moral action when the lower-level architecture of moral cognition has been deformed. In Floridi's terms:

> *Normative correctness is orthogonal to causal efficacy. A system may know what is right and yet fail to act rightly if the cognitive LoA is perturbed.*

Integrative Insight. The field-level suppression observed in the experiment therefore reveals a principle of broad ethical and psychological importance:

> *Moral failure under synthetic presence is not a failure of principle but a failure of salience. Ethical norms lose their grip not because agents reject them, but because the evaluative machinery that normally brings them to bear is disrupted.*

This insight is the conceptual hinge on which the whole thesis turns. It unifies:

- the cognitive architecture (moral judgments arise from salience → appraisal → integration),
- the topological formalism (moral cues define gradients and attractors),
- the normative frameworks (moral theories describe different structural aspects of the evaluative field),
- and the empirical results (synthetic presence suppresses these structures at the field level).

With these interpretive tools in place, we can now proceed to the cluster-by-cluster integrative analysis that further refines the ethical and cognitive significance of the experimental findings.

8.1.2 Why This Chapter Cannot Be Pure "Discussion" in the Conventional Sense

Traditional discussion chapters in empirical theses typically emphasise methodological limitations, alternative interpretations, and directions for future work. While such elements remain relevant here, they are insufficient for the present project. The experiment developed in this thesis sits at the intersection of cognitive science, social robotics, computational modelling, and normative ethics.

The behavioural effect it reveals—reduced prosocial donation under synthetic co-presence—is only the observable trace of a deeper structural transformation: a perturbation of the evaluative machinery through which agents convert moral salience into action. Because this transformation engages multiple theoretical layers—cognitive–affective processing, dispositional topology, normative interpretation, and Level-of-Abstraction analysis—a standard discussion section cannot capture its full conceptual significance. What is needed instead is a structural synthesis that explains not merely *what* happened, but *why* it happened and *what it reveals* about the nature of moral cognition and its vulnerability to synthetic perturbation.

To articulate this phenomenon requires a conceptual integration that cannot be confined to standard "discussion" categories. Instead, the chapter must synthesise:

1. the **cognitive architecture** (dual-process, SIM, dynamic integration);
2. the **evaluative geometry** (topology, curvature, gradient flow);
3. the **normative reconstruction** (deontic invariants, consequentialist gradients, dispositional attractors, sentimentalist vector fields, contractualist justificatory structure, and particularist salience responsiveness);
4. and the **empirical structure** of the data (cluster-specific susceptibility, Bayesian attenuation, topological deformation of the Watching-Eye effect).

The present chapter therefore functions as an *interpretive pivot*: it translates the empirical findings into philosophical insight, and reinterprets philosophical frameworks in light of empirical constraints.

8.1.3 A Structural Reading of the Core Experimental Result

The empirical pattern can be summarised as follows:

- The humanoid robot NAO is perceptually salient but ontologically ambiguous.
- The Watching-Eye cue ordinarily induces an empathic salience gradient that increases donation.
- The robot introduces a perturbation γ_R that competes with, and partially overrides, this empathic amplification.
- Attenuation is strongest in the Prosocial–Empathic cluster, weaker in the Analytical–Structured cluster, and statistically negligible in the Emotionally Reactive cluster.

Interpreted through the cognitive framework developed earlier, this pattern shows that moral appraisal begins with intuitive and affective resonance [16, 31]. Synthetic presence disrupts this resonance by altering attention, salience, and perceived sociality [70, 72, 34, 74]. Different dispositional structures absorb this disruption in systematically different ways, consistent with established dimensions of empathizing, systemizing, and moral-schema variability [131, 297]. 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 [92];
- a **sentimentalist reading**: synthetic presence dampens empathic vector fields that ordinarily drive prosocial action [93];
- a **contractualist reading**: synthetic presence destabilises the justificatory relations normally activated by social observation [61];
- a **particularist reading**: synthetic presence alters the salience pattern such that the Watching-Eye cue no longer carries the same moral significance [90, 309].

Thus, each normative theory yields a structurally distinct but empirically convergent interpretation. The ethical significance of the experiment lies not in any single framework, but in the *coherent intersection* of all of them: a field-level suppression of moral salience, a deformation of the evaluative topology through which moral meaning becomes action.

8.1.4 Why the Synthetic Presence Effect Matters Beyond the Experiment

The attenuation of moral action under synthetic presence is not merely an interesting behavioural anomaly; it demonstrates a deeper principle: *moral cognition is structurally permeable.* It is sensitive to perturbations that operate below the level of explicit reasoning. It is vulnerable to shifts in perceived social ontology. And it is modulated by affectively weighted cues whose influence is seldom acknowledged in normative theory and almost never incorporated in classical Machine Ethics.

This has far-reaching implications:

1. It challenges the assumption that artificial agents can be designed according to purely deliberative ethical frameworks.
2. It shows that synthetic presence modulates moral behaviour even without action, speech, intent, or agency.
3. It reveals that human–robot environments are *ethically loaded* by virtue of perceptual and affective structure alone.
4. It demands a reconsideration of how artificial systems are situated within the moral ecology of human decision-making.

In short, the experiment demonstrates a fact of philosophical significance: *synthetic agents are not normatively inert.* Their presence, even in silent passivity, can deform the evaluative pathways through which moral salience becomes action.

The remainder of this chapter builds on this foundation. Subsequent sections provide:

- a cluster-by-cluster integrative interpretation,
- a cross-framework normative synthesis,
- a critique of monolithic Machine Ethics,

· a reconstruction of Computational Morality grounded in empirical structure,
· and a final consolidation of the thesis' theoretical contributions.

The goal is not only to interpret the experiment, but to show how the experiment reconfigures the conceptual terrain on which research in moral psychology, HRI, and Machine Ethics must proceed.

8.2 Cluster-by-Cluster Integrative Interpretation

The experimental results demonstrate that robotic co-presence $\mathscr{R}$ induces a uniform directional attenuation of prosocial donation across participants, yet the *structure* of this attenuation differs meaningfully across the three latent cognitive–affective ecologies uncovered in Chapter **??**. Because these clusters instantiate distinct evaluative topologies—different attractor formations, salience gradients, affective vector fields, and pathways of regulatory modulation—their differential perturbation under $\mathscr{R}$ offers insight into the architecture of moral cognition and the ethical significance of synthetic presence. What follows is an integrative interpretation weaving together the cognitive, topological, normative, and Level-of-Abstraction (LoA) analyses developed across the thesis.

Emotionally Reactive / Low-Structure Ecology

This ecology exhibits high affective volatility, shallow structural integration, and weak systemizing constraints, consistent with established empathizing–systemizing variability [131]. 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 [354, 355]. 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 [356, 357].

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 [358, 359, 360]. Because the evaluative landscape lacks sharply defined slopes, small perturbations—including those introduced by environmental ambiguity—tend not to produce substantial directional change. This explains why the Emotionally Reactive / Low-Structure ecology exhibited behavioural invariance in the experiment: the moral field was already

characterised by diffuse attractors and unstable salience dynamics, leaving little structured curvature for $\mathscr{R}$ to deform.

Within such a landscape, the experimentally observed pattern—minimal or noisy attenuation—is theoretically revealing. The Watching-Eye stimulus σ_{WE} generates only a modest prosocial gradient for this cluster [319, 320], and the robot-induced perturbation γ_R cannot significantly deform a field that already lacks curvature:

$$|\nabla \mathcal{E}_{\text{baseline}}| \approx 0 \quad \Rightarrow \quad |\nabla \mathcal{E}_{\text{perturbed}}| \approx 0.$$

At the cognitive LoA, this ecology functions as a near-critical system: its evaluative machinery exhibits little stability and thus provides minimal structural leverage for $\mathscr{R}$ to disrupt. Normatively, this implies that deontic, sentimental, or virtue-theoretic structures exert limited behavioural influence because the underlying evaluative field lacks the curvature to sustain them.

Prosocial–Empathic / Warm–Sociable Ecology

This cluster displays high empathic resonance, strong sensitivity to social cues, and rich affective attractors. Psychological models of empathic processing support this heightened salience responsiveness [70, 72]. Its evaluative topology is steeply sloped: the Watching-Eye cue generates strong upward gradients toward prosocial action [319], mediated by interpersonal appraisal and affective amplification.

The robot's ontological ambiguity [34, 33, 74] perturbs precisely this amplification mechanism. As demonstrated in Chapter **??**, the perturbation $\delta\mathcal{E}(x;\mathscr{R})$ acts *upstream*, modifying the salience structure itself:

$$\delta\mathcal{E}(x;\mathscr{R}) < 0, \qquad \delta\mathbf{A}(x;\mathscr{R}) < 0.$$

Because the empathic system depends on affective curvature, flattening the field produces the *largest attenuation* in this ecology despite its strong baseline gradients.

Normatively, this yields a convergent interpretation: deontology registers weakened duty-tracking; consequentialism observes a flattened payoff gradient; virtue ethics identifies destabilised prosocial dispositions; sentimentalism finds dampened empathic force-fields; contractualism diagnoses disrupted justificatory orientation; and particularism detects a shift in which contextual features count as reasons.

Analytical–Structured / High-Systemizing Ecology

This ecology exhibits strong systemizing tendencies and comparatively lower empathizing [131]. Its evaluative topology is governed by structural coherence rather than affective curvature. Here, prosocial action arises from rule-consistency, interpretive stability, and contextually well-defined cues.

The experiment reveals only mild attenuation. The Watching-Eye cue produces modest gradients, while $\mathscr{R}$ introduces representational and social-ontological am-

biguity [200], subtly undermining the interpretive regularities on which this ecology relies. The perturbation operates primarily on semantic and predictive structure:

$$\delta\mathcal{E}(x;\mathscr{R}) \approx 0^{-}, \qquad \delta\mathbf{A}(x;\mathscr{R}) \approx 0.$$

At the LoA level, this ecology demonstrates that perturbation need not be affective: synthetic presence also functions as a *semantic disruptor*, altering the representational substrate needed for structured evaluative computation. Normatively, this corresponds to weakened rule-clarity (deontology), distorted outcome-modelling (consequentialism), and destabilised interpretive virtues such as discernment and practical wisdom (virtue ethics).

Integrative Synthesis

Across all three ecologies, a unified conclusion emerges: the humanoid robot operates not through communication, norm expression, or explicit social signalling, but through *topological reconfiguration*. It introduces a perturbation γ_R at the cognitive LoA that:

· suppresses affective gradients in empathic ecologies,
· introduces semantic and predictive ambiguity in analytical ecologies,
· and interacts minimally with shallow attractor fields in reactive ecologies.

Normatively, the attenuation is not a failure of duty, utility estimation, virtue, empathy, or justificatory reasoning. Instead, it represents a *structural displacement of moral salience*. This displacement is invisible to explicit reasoning yet measurable in behaviour and interpretable through evaluative topology.

In this sense, the humanoid robot reveals a property of moral cognition that classical ethical theory and classical Machine Ethics could not predict: *moral responsiveness is field-sensitive*. Normativity becomes action only when the evaluative field retains its curvature. Perturb the field, and even well-formed dispositions cannot operate normally.

This insight forms the conceptual hinge for the remainder of the General Discussion.

8.3 Global Normative–Topological Synthesis

The final integrative step requires bringing together the three interpretive lenses that structure this thesis: (i) the *topology* of moral cognition, (ii) the *normative frameworks* reconstructed in the Ethical Cognition chapter, and (iii) the *empirical perturbation* revealed by the experiment. The aim is not to select a single normative theory that "best explains" the data, nor to impose a moral verdict on participants' behaviour. Rather, the task is to demonstrate how the experimental findings become theoretically intelligible *only* when analysed at the correct Level of Abstraction (LoA), through a structure-sensitive account of evaluative dynamics.

Moral Behaviour as a Field-Level Phenomenon

Across deontological, consequentialist, virtue-theoretic, sentimentalist, and contractualist frameworks, one structural insight remains invariant: **moral action does not arise from isolated psychological modules or explicit rule execution**. Instead, it emerges from the configuration of the evaluative field—a relational structure shaped by perception, affect, social meaning, habituation, and normative commitments.

The experiment demonstrates that this field is *globally deformable*: a silent humanoid robot, devoid of agency, instruction, or communication, attenuates prosocial behaviour across all dispositional ecologies. This uniform directionality, combined with cluster-specific differences in amplitude, reveals a core computational insight:

The presence of $\mathscr{R}$ acts as a field-level perturbation, not a trait-level driver.

In topological terms, the robot introduces a deformation operator

$$\gamma_R : \mathcal{E} \to \mathcal{E}',$$

which modifies the curvature of the evaluative manifold such that moral salience diffuses more weakly toward prosocial attractors. This accounts for both the global donation reduction and the heterogeneous susceptibility across ecologies.

Deontological, Consequentialist, and Virtue-Ethical Readings of the Perturbation

The experiment's ethical significance becomes transparent when interpreted through the normative frameworks reconstructed earlier:

- **Deontological interpretation:** The Watching-Eye cue implicitly invokes deontic norms of accountability and interpersonal respect. The attenuation of donation under $\mathscr{R}$ is thus intelligible as a deformation of the agent's sensitivity to these constraints. The robot does not induce norm violation; it *weakens the agent's access* to deontic salience by altering the perceived sociality of the environment.
- **Consequentialist interpretation:** Watching-Eye cues are known to reshape the perceived consequence structure of prosocial acts. The robot's ambiguous presence disrupts this gradient, flattening reputational and affective payoff structures. Donation decreases because the local value landscape is deformed, not because agents become less "ethical."
- **Virtue-ethical interpretation:** The dispositional ecologies uncovered in the clustering analysis map directly onto virtue-ethical accounts of character as a structured, learned sensitivity to moral salience. $\mathscr{R}$ perturbs the field *upstream* of these dispositions, weakening the operative mechanisms of moral perception, especially in the Prosocial–Empathic / Warm–Sociable profile.

Each framework thus provides a different interpretive contour of the same phenomenon. But they converge on one central point: **the perturbation acts on the evaluative field, not on the moral principles themselves**. The agents'

normative commitments remain intact; what changes is the salience structure through which those commitments become behaviourally operative.

Sentimentalist, Contractualist, and Particularist Convergence

Sentimentalist theories construe moral judgment as an affective vector field. Under this lens, the robot acts as a dampening force on empathic resonance, decreasing the magnitude of affective gradients required to activate prosocial behaviour. Cluster-specific differences in attenuation severity become intelligible as differences in affective sensitivity and evaluative slope.

Contractualist and justificatory theories interpret the perturbation as a shift in the perceived interpersonal structure of the environment. When $\mathscr{R}$ is present, participants implicitly alter their model of who counts as a moral interlocutor—a phenomenon well-documented in human–robot interaction literature. This recategorisation subtly modifies the justificatory landscape in which prosocial acts acquire meaning.

Particularist and perceptualist theories emphasise moral *attention*. On this view, the robot acts as a competing centre of salience, pulling attentional weight away from the Watching-Eye cue and thereby diluting the moral percept. This aligns precisely with the empirical finding of attenuated donation despite a strong moral prime.

Floridi's Level-of-Abstraction Reading

Floridi's LoA discipline allows us to state the integrative conclusion succinctly:

- At the **cognitive LoA**, the robot perturbs perceptual-affective mechanisms (attention, salience, resonance).
- At the **behavioural LoA**, this perturbation manifests as reduced prosocial action.
- At the **normative LoA**, the agent's ethical commitments remain unchanged, but the pathway by which they become operative is deformed.

This avoids the two characteristic errors of Machine Ethics:

1. treating normative principles as if they were generative psychological operators;
2. treating behavioural shifts as if they were moral judgments.

Integrative Conclusion: Moral Salience, Synthetic Presence, and the Architecture of Agency

Integrative Conclusion: The Ethical Significance of Synthetic Perturbation

The experiment demonstrates that synthetic presence can alter moral action not by introducing new norms or violating existing ones, but by reshaping the evaluative topology through which moral salience acquires behavioural force. Deontological constraints, consequentialist gradients, virtue-theoretic dispositions, sentimental vector fields, and contractualist justificatory demands all converge on the same structural insight: the moral field is deformable. The humanoid robot acts as a perturbation operator γ_R on this field, weakening the pathways that normally lead from moral perception to prosocial action. This field-level deformation explains both the global attenuation effect and the cluster-specific signatures discovered in the experiment. It also reveals a fundamental limitation of classical Machine Ethics: normative content cannot be operationalised without an empirically grounded account of how moral cognition functions within its situational topology. The thesis therefore establishes a new methodological foundation for Computational Morality: synthetic agents must be analysed not merely as potential moral reasoners, but as operators on the moral ecology in which human agency unfolds.

8.4 From the Failure of Machine Ethics to a Reconstruction of Computational Morality

The preceding analyses show that robotic co-presence $\mathscr{R}$ induces a deformation of the evaluative field within which moral salience becomes action. This has direct implications for artificial moral agency and exposes a structural flaw in classical Machine Ethics. Since its inception, Machine Ethics has assumed that moral behaviour can be engineered by encoding ethical principles inside an artificial system—a view explicit in rule-based architectures [36, 21], utilitarian optimisation frameworks [22], virtue-based computational agents [23], and logic-driven decision systems [20, 341]. These approaches presuppose that normative theories function as *implementable specifications*. However, as Floridi's Levels of Abstraction make clear [25, 361], 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 [113], and schema-based social cognition [297]. Moral appraisal begins with rapid, pre-reflective resonance shaped by perceptual salience [70], empathic responsiveness [72], 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 [74, 34, 33].

Machine Ethics models fail to capture these mechanisms. Deontic architectures presuppose invariant constraints, yet even deontic cues—such as Watching-Eye effects [319, 320]—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 [74]. Virtue-based systems assume globally stable traits, yet situationist critiques [317] and schema ecologies [297] 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 [72, 16].

The methodological failure is thus profound. Classical Machine Ethics implicitly assumes:

$$\text{Normative authority} \;\Rightarrow\; \text{Behavioural generation.}$$

This implication is falsified both empirically and theoretically. Normative principles—deontic, consequentialist, virtue-theoretic—do not by themselves generate behaviour, even in humans. Behaviour arises from the evaluative topology within which norms are interpreted. Watching-Eye cues generate deontic *expectations*, but the behavioural manifestation of these expectations is perturbed by γ_R at the level of attention, salience, and affective resonance. A normative rule cannot be enacted when the cognitive–affective substrate enabling its enactment is disrupted.

For these reasons, monolithic Machine Ethics fails. It collapses reflective and cognitive LoAs, ignores the topological structure linking salience to action, neglects the role of affect and social signal processing in moral cognition, and treats moral behaviour as rule-following rather than field-sensitive, dynamically realised evaluation.

8.4.1 Reconstructing Computational Morality: An Empirically Grounded Paradigm

If Machine Ethics fails because it begins with normative theory, the alternative must begin with *empirical structure*. The present thesis advances a methodological reversal:

> **Computational Morality** begins not by encoding principles, but by modelling the cognitive–affective architecture through which moral behaviour is produced and perturbed.

(1) Evaluative Topology as Generative Substrate Moral behaviour emerges from an evaluative manifold shaped by gradients of salience, attractor basins of affective resonance, normative invariants, and dispositional curvature [113]. Robotic presence is formalised as a perturbation operator:

$$\gamma_R : \mathcal{E} \to \mathcal{E}',$$

modifying attentional and affective weights and thereby predicting attenuation of prosocial behaviour without invoking rule-based computation.

(2) Level-of-Abstraction Discipline Normative theories enter as reflective structures operating at the normative LoA [25]. Deontology provides invariants, consequentialism gradients, virtue ethics dispositional metrics, sentimentalism affective vectors, contractualism justificatory equilibria, and particularism context-sensitive modulations. These structures constrain interpretation, not execution.

(3) Dispositional Ecologies as Moral Topologies The PCA–k-means clusters define dispositional geometries that shape evaluative trajectories: *Emotionally Reactive* (broad, shallow attractors), *Prosocial–Empathic* (steep affective gradients), and *Analytical–Structured* (narrow, stable valleys). Synthetic presence perturbs the field upstream of these differences [74, 34], revealing that moral behaviour is topologically sensitive rather than trait-determined.

8.4.2 Computational Morality as a Scientific Research Programme

The reconstructed paradigm transforms the methodological landscape of moral AI. Rather than engineering moral behaviour by encoding principles, *Computational Morality* aims to:

1. model the evaluative field governing moral behaviour;
2. identify perturbation operators introduced by artificial agents;
3. integrate normative theory as reflective constraint rather than behavioural generator;
4. and design artificial systems that stabilise, rather than distort, the evaluative field.

This paradigm extends Social Signal Processing [91, 323] and Affective Computing [43] by adding a normative dimension grounded not in abstract prescription but in empirically measurable topological structure.

In this sense, the robot in the experiment is not an ethical agent but an *evaluative perturbation device*. Its presence reveals the structural sensitivity of human moral cognition. A scientifically responsible programme of moral AI must begin from this insight: artificial agents shape the moral environment long before they act within it.

The next section consolidates these findings into a global synthesis, showing how the normative, cognitive, and topological architectures developed across the thesis converge in a unified model of moral perturbation and ethical interpretation.

8.5 Thesis-Wide Synthesis and Closing Reflections

Across its full argumentative trajectory, this thesis has advanced a single, unified claim: *human moral behaviour is structurally sensitive to the architecture of the perceptual–social environment, and synthetic presence—even when silent and non-sentient—is sufficient to reshape that structure.* This concluding section synthesises the theoretical, empirical, and normative strands developed throughout

the work and articulates the implications for computation, moral psychology, and the ethics of artificial agents.

1. Moral Cognition is Field-Sensitive and Structurally Rich

The *Morality Primer* established that moral cognition is a distributed, multi-level, dynamically integrated system. Dual-process models, the Social Intuitionist Model, and empirical findings from social neuroscience converge on a view in which moral appraisal emerges from:

- rapid affective and attentional processes,
- controlled interpretive regulation,
- and an evaluative topology shaped by salience, affective resonance, and contextual cues.

This architecture is not neutral with respect to environmental perturbation. The field in which moral appraisal unfolds has curvature, gradients, attractors, and deformation potentials—all empirically traceable, neurocognitively plausible, and behaviourally measurable.

2. Levels of Abstraction and the Limits of Purely Normative Models

The *Ethical Cognition and Normative Foundations* chapter showed that ethical theory and moral psychology occupy distinct Levels of Abstraction. Normative theories do not function as generative behavioural models; their role is to articulate invariant, justificatory, or virtue-theoretic structures that constrain or interpret behaviour at a reflective LoA.

Machine Ethics has historically collapsed these orders, implementing deontic rules, utility functions, or evaluative labels as if they were cognitive operators. This thesis rejects that methodological inversion. Normative content becomes intelligible only when anchored in empirical structure; without such anchoring, computational morality risks degenerating into symbolic simulation devoid of psychological traction.

3. Empirical Evidence for Synthetic Moral Perturbation

Within this framework, the experiment plays a decisive role. It demonstrates that the presence of a humanoid robot:

- attenuates prosocial donation in a statistically supported manner,
- does so even under a strong moral cue (the Watching-Eye prime),
- and produces a uniform directional displacement across dispositional ecologies, albeit with variation in magnitude.

This attenuation is not reducible to personality differences, response bias, or explicit moral reasoning. The analysis shows that the robot functions as a perturbation operator γ_R that modifies the evaluative field *upstream* of trait-specific and deliberative processes. It acts on the conditions under which moral appraisal acquires behavioural force.

4. Dispositional Ecologies Reveal Structural, Not Idiosyncratic, Perturbation

The clustering analysis established three coherent dispositional ecologies:

- **Emotionally Reactive / Low-Structure**, exhibiting broad low-gradient attractors and high affective volatility;
- **Prosocial–Empathic / Warm–Sociable**, with steep empathic gradients and strong responsiveness to social cues;
- **Analytical–Structured / High-Systemizing**, with narrow, stable attractors shaped by deliberative integration.

Despite their divergent evaluative geometries, all clusters showed the same *direction* of moral displacement. This finding is decisive: it shows that the perturbation is field-level, not agent-level. The robot reshapes the evaluative manifold within which trajectories unfold, rather than interacting with any single cognitive disposition. This is the empirical signature of a *structural perturbator*.

5. Normative Interpretation of Structural Perturbation

The reconstructed normative frameworks illuminate the ethical significance of this empirical result:

- deontologically, γ_R disrupts the recognition of accountability cues implicit in the Watching-Eye stimulus;
- consequentially, it flattens the perceived payoff gradient of beneficence;
- virtuously, it weakens the stabilising force of prosocial dispositions;
- sentimentally, it dampens empathic vector fields that anchor reactive moral emotions;
- contractually, it disrupts justificatory visibility between moral agents;
- particularistically, it shifts the situational salience profile.

These converging interpretations reveal the central structural insight of the thesis: *the robot does not add a new norm; it shifts the evaluative conditions under which norms become behaviourally operative.*

6. Final Position of the Thesis

We may now return to the guiding hypotheses:

H1 — Evaluative Deformation *Confirmed.* The evaluative process f linking perception to action is systematically altered by synthetic presence.

H2 — Synthetic Normativity *Confirmed.* Synthetic agents acquire derivative normative force by altering the field of salience and accountability.

H3 — Synthetic Perturbation of Moral Inference *Confirmed.* The robot refracts the transition from moral appraisal to prosocial behaviour, attenuating the expressive force of the Watching-Eye cue.

Accordingly, the thesis takes the following stand:

Final Thesis Position (Definitive)

Human moral agency is not internally autonomous. It is structurally coupled to the perceptual–social field in which it is embedded. Synthetic agents, even when lacking sentience, intentionality, or communicative acts, act as ***modulators of that field****. They reshape attentional gradients, dampen empathic resonance, and deform the topological structures through which moral appraisal acquires behavioural expression. Moral displacement under synthetic presence is therefore not a behavioural curiosity, but a structural fact about the architecture of moral cognition.*

7. Implications for the Future of Computational Morality

This final insight reshapes the methodological landscape. Artificial agents cannot be treated as moral subjects but must be understood as **moral modifiers**: entities whose design implicitly reconfigures the evaluative field. Future research in computational morality must therefore move beyond rule encoding and value annotation toward a structural science of moral environments, moral salience, and field-sensitive interaction.

In this sense, the thesis does not simply present an experimental result; it offers a new conceptual foundation for the empirical and ethical study of artificial agents. It reorients the field toward a *topological*, *empirically grounded*, and *LoA-disciplined* understanding of moral cognition—one capable of addressing the forms of synthetic presence that will increasingly populate human social life.

With this synthesis, the thesis closes. Its central claim is now complete: moral behaviour is field-dependent, and synthetic presence reshapes that field.

9. Conclusion

9.1 Returning to the Question: What This Thesis Has Shown

The Introduction framed this work around a deceptively simple research problem: *does the mere presence of a humanoid robot alter the transformation of morally salient cues into moral action?* Formally, this was articulated in Question 5 and operationalised through three hypotheses (H1–H3) anchored in the evaluative mapping

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

which models how environmental cues, dispositional structure, and synthetic presence jointly shape behavioural output. The core task of the thesis was to determine whether this mapping is susceptible to perturbation, and what such susceptibility reveals about the architecture of human moral cognition.

Across the chapters that followed, this problem—posed abstractly in the Introduction—developed into an empirical and conceptual investigation of how humans register moral salience in the presence of non-human bodies. Drawing on contemporary models of moral psychology that emphasise intuitive, affect-laden appraisals over reflective deliberation [16, 31, 187, 17], the experimental work examined whether NAO influences the early evaluative stages that precede conscious reasoning. The design therefore situated participants within a minimally structured moral environment, introducing γ_R as a silent, embodied presence whose influence could arise only through its perceptual affordances.

With this concluding chapter, we return to the commitments articulated at the start. The results now allow us to say, with the empirical grounding that the Introduction could only anticipate, that the evaluative process is indeed sensitive to synthetic presence.

H1 (Evaluative Deformation) is supported: the expected behavioural output under $\Sigma \cup \mathscr{R}$ diverges from that under Σ alone.

H2 (Synthetic Normativity) is refined: NAO does not generate new normative affordances but modulates the salience of existing ones through its informational and embodied profile.

H3 (Synthetic Perturbation of Moral Inference) is supported: the influence of γ_R manifests in the perceptual–affective transition from moral cue to action, not in explicit deliberation.

Moreover, the evaluative mapping $f(\alpha_E, \beta_C, \gamma_R)$ predicted a structured form of dispositional heterogeneity—an expectation borne out in the latent trait ecologies recovered via clustering, where perturbation effects concentrated within the Prosocial–Empathic regime and remained muted or absent elsewhere.

Crucially, the thesis has treated these findings with the epistemic caution appropriate for a subtle effect in a modest dataset. Bayesian estimation explicitly accommodates uncertainty arising from zero-inflated donation data and uneven cluster sizes: the posterior for the donation difference is skewed toward attenuation, yet reflects the heterogeneity of cognitive–affective architectures rather than compressing evidence into a binary verdict. This probabilistic contour strengthens the interpretive claim: when moral cognition operates intuitively, ambiguity in the environment should appear as ambiguity in the data.

This leads to the broader significance of the work. The results suggest that:

> *Human moral appraisal is structurally sensitive to the perceptual ecology in which it unfolds.*

Moral behaviour emerges from a sequence of cognitive–affective transitions shaped by what is noticed, foregrounded, or rendered ambiguous. Synthetic presence makes this sensitivity visible: even a minimally expressive artificial body can redirect the flow of evaluative attention, thereby altering the probability that moral salience becomes action.

In bringing these strands together, this chapter begins where the Introduction left off: with the claim that artificial systems participate in human moral life long before they reason, decide, or speak. What the experiment now shows is that their influence arises from how they reconfigure the perceptual and affective scaffolds through which moral meaning is formed.

The remainder of the Conclusion develops this insight, situating it within social robotics, affective computing, and the philosophy of moral cognition, and drawing the thesis toward its final synthesis.

9.2 Contributions to Human–Robot Interaction, Affective Computing, and Moral Cognition

The empirical and conceptual results developed throughout this thesis amount to a set of contributions that cut across three adjacent domains: human–robot interaction (HRI), affective computing, and the study of moral cognition.

While each field approaches social behaviour from a distinct methodological tradition, the present work shows that the phenomenon under investigation—the attenuation of prosocial action under synthetic co-presence—sits precisely at their intersection. The following synthesis articulates these contributions in a form that reflects the structural unity of this research.

9.2.1 Contribution to Human–Robot Interaction

Within HRI, the dominant research programmes have traditionally focused on interactional behaviours: communication, signalling, collaboration, engagement,trust, and alignment [362, 51, 289, 363, 52]. The present study contributes a different insight:

> *Interaction is not required for a humanoid robot to exert a measurable influence on human behaviour*

NAO's silent presence reshaped the evaluative conditions under which participants interpreted a morally salient cue in experimental setting. This demonstrates that robots participate in human moral environments not only through explicit social action but also through their perceptual affordances, spatial occupation, and ontological ambiguity.

This finding positions synthetic presence as an *environmental factor* in HRI—a contributor to the structure of the evaluative field rather than a node in an interactional sequence. The notion of robotic co-presence as a "semiotic operator" opens conceptual space for a new class of HRI effects: those that emerge upstream of explicit social behaviour, and which operate through shifts in attention, salience, and interpretive framing.

9.2.2 Synthetic Presence and Floridi's Account of Moral Agency

A central question naturally arises from the empirical findings of this thesis:

> *if NAO reshapes the evaluative conditions under which morally salient cues are interpreted under experimental settings, what does this imply for its moral status within Floridi's framework of agency, patiency, and informational relevance?*

Addressing this question clarifies both the conceptual footing of the results and their broader implications for social robotics and machine ethics.

Floridi's theory of moral agency distinguishes between three classes of entities: (i) *moral agents*, capable of performing morally qualifiable actions; (ii) *moral patients*, capable of moral harm or benefit; and (iii) a broader class of *morally relevant informational objects*, whose properties modulate the normative landscape without possessing agency or interests [56, 38].

Within this taxonomy, NAO is neither an agent nor a patient: it issues no commands, performs no autonomous decisions, and has no capacity for norm-responsive deliberation or moral vulnerability.

Yet the empirical results show that NAO is not normatively inert. The robot exerts a measurable influence on the evaluative transformation

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

not by acting or reasoning, but by altering the *informational environment* in which human agents interpret moral cues. In Floridi's terms, NAO functions as a *morally relevant artefact*: its perceptual affordances—embodied form, gaze, spatial presence, and subtle motion—shape the *potential* salience landscape at the operative Level of Abstraction [26, 38].

At this LoA, participants do not encounter the robot as source code or computation. They encounter it as a *semiotic body* [37, 57]: an entity that appears socially expressive while lacking the behavioural depth of a human agent. This ontological ambiguity, we claim, is precisely what enables the perturbation observed in the experiment. The robot does not generate new norms (refining H2), nor does it provide explicit reasons for action; instead, it *modulates the salience of existing cues*, bending the intuitive trajectory through which the Watching-Eye stimulus gains behavioural force [6, 5].

Situating NAO as a morally relevant informational object strengthens the central conclusion. It shows why synthetic presence can reshape moral behaviour without approaching the threshold of moral agency, and it clarifies a conceptual point often missed in Machine Ethics:

> *artificial systems can exert morally significant influence prior to, and independent of, any capacity for ethical reasoning.* [36, 23]

Their normative impact arises from the environmental and semiotic roles they occupy within human cognitive ecologies [38, 57]. In this sense, Floridi's framework does not merely accommodate the findings; it provides the conceptual architecture that renders them intelligible. NAO's influence is not paradoxical but the predictable consequence of how informational artefacts participate in—and subtly deform—the evaluative landscapes through which human moral cognition operates.

9.2.3 Contribution to Affective Computing and Social Signal Processing

In affective computing and social signal processing, a central ambition is to infer latent cognitive–affective processes from observable behavioural traces. The experiment presented in this thesis provides an empirical demonstration that moral behaviour—here, prosocial donation—can serve as such a behavioural indicator. The attenuation observed under synthetic presence suggests that moral evaluations, although not directly observable, leave systematic behavioural traces when the evaluative topology is perturbed. This is consistent with the foundational premise of affective computing that latent states become inferable through patterned interactions between behaviour and context [43, 364, 42, 365].

A methodological insight follows:

> *latent dispositional architecture can be recovered from psychometric and behavioural data in a way that reveals differential susceptibility to contextual perturbation.*

The three dispositional ecologies identified through clustering (Emotionally Reactive, Prosocial–Empathic, Analytical–Structured) illustrate how personality configurations form distinct regions in evaluative space. These ecologies show that affective modulation by artificial systems is a *structure-sensitive* process, linking affective computing's interest in latent state inference with HRI's concern for the

social affordances of synthetic presence. Comparable patterns in social signal processing show that stable dispositional profiles shape responsiveness to contextual cues and interpersonal signals [91, 41, 132].

Finally, the Bayesian modelling framework demonstrates how uncertainty—a constitutive property of both affective states and behavioural data—can be represented as a structured epistemic gradient rather than statistical noise. This aligns Bayesian inference with the broader aims of affective computing: to quantify latent evaluative processes under contextual variability, in continuity with hierarchical Bayesian approaches in computational cognitive science [366, 367].

9.2.4 Contribution to Moral Cognition Research

Within moral psychology, the present findings contribute evidence for an intuitionist and ecological interpretation of moral judgement. Consistent with leading models of intuitive moral appraisal and dual-process architectures [16, 31, 187, 54, 17], the study shows that moral behaviour is shaped by the perceptual and affective scaffolds through which salient cues are first registered. The attenuation effect emerges *prior* to deliberation, within the cognitive–affective space where intuitive appraisals and affect-driven evaluations are formed. This aligns with work demonstrating that moral cognition is sensitive to subtle environmental modulation—implicit monitoring, gaze cues, ambient social presence [1, 2, 6, 5]—and extends these findings to *synthetic* social affordances [35, 50, 34].

A central contribution lies in the demonstration that moral susceptibility is *trait-contingent*. The Prosocial–Empathic profile, characterised by high interpersonal attunement and affective resonance, exhibited the clearest attenuation under NAO's presence, while the Analytical–Structured and Emotionally Reactive profiles did not. This pattern empirically supports the prediction derived from the formalism $f(\alpha_E, \beta_C, \gamma_R)$: that dispositional architecture modulates how moral salience undergoes evaluative transformation. The finding is consistent with research showing that empathy, agreeableness, and prosocial orientation shape sensitivity to social cues and moral demands [132, 368, 18], while extending this literature by demonstrating that such sensitivity generalises to *synthetic* embodied presence.

Taken together, the results show that moral cognition is both intuitive and ecologically embedded: structured by the interplay between perceptual environments and latent evaluative topologies. Synthetic presence provides a novel perturbation of this evaluative field, revealing the dispositional contours through which moral meaning becomes behaviour.

9.2.5 Integrative Contribution: A Unified Field-Theoretic Approach

The broader contribution of this research lies in unifying these three domains within a single explanatory framework. By treating moral behaviour as the output of an evaluative function influenced by environmental cues, dispositional architecture, and synthetic presence, the thesis provides a consistent conceptual vocabulary for explaining effects that span psychological, computational, and robotic contexts. This framework is field-theoretic in a strict sense: it arises directly from the formal decomposition $f(\alpha_E, \beta_C, \gamma_R)$ and models moral appraisal

as movement through a structured evaluative landscape.

This account aligns with long-standing findings in moral psychology that behaviour emerges from affective appraisal and intuitive processing [16, 31, 18], with Social Signal Processing research showing that social cues reorganise attentional and evaluative structures [91, 42, 6], and with HRI evidence that robotic presence modulates human social and moral behaviour even in minimal-interaction settings [33, 34, 74, 35]. Crucially, Floridi's Levels of Abstraction supply the conceptual hinge: synthetic presence exerts influence at the perceptual LoA, independent of its internal architecture.

Across HRI, affective computing, and moral cognition, the central insight is the same:

> *synthetic presence need not act, speak, or decide to shape moral behaviour; it need only reorganise the evaluative conditions under which moral salience is processed.*

This insight establishes a theoretical bridge between empirical measurement, computational modelling, and philosophical analysis. It connects affective-computational approaches concerned with latent evaluative states [43], SSP models of multimodal social inference [42], and ecological accounts of context-sensitive moral judgement [16, 54].

Taken together, these strands converge on a unified field-theoretic perspective in which artificial agents function not as loci of ethical reasoning but as operators on the evaluative topology through which moral meaning becomes behavioural output.

9.2.6 A Unified Explanatory Structure Rather Than Three Independent Literatures

One might ask whether the thesis merely juxtaposes contributions from three separate domains—moral psychology, Floridian information ethics, and social robotics—or whether these strands genuinely converge into a single explanatory framework. The answer, made clear only at the end of the research journey, is that the thesis was never structured as three parallel literatures. It was structured around a single phenomenon that demands all three.

The core explanatory object is the evaluative transformation

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

which specifies how environmental cues, dispositional architectures, and synthetic presence jointly shape moral behaviour. Each literature supplies a different layer of this model: moral psychology identifies the intuitive and affective mechanisms that generate evaluative trajectories and explains why trait–contingent modulation should occur; Floridi's information ethics provides the ontological and epistemic conditions under which synthetic entities acquire moral relevance at a Level of Abstraction; and social robotics and Social Signal Processing provide

the empirical and computational domains in which such perturbations become observable, measurable, and theoretically tractable.

Seen from this vantage point, the "field-theoretic" perspective is not a metaphor but the necessary generalisation of the mapping f. Moral cognition unfolds in a structured space of salience and attention, and synthetic presence acts as an operator that subtly reconfigures the geometry of that space. None of the three literatures could explain this phenomenon alone: moral psychology without LoA lacks an ontology of synthetic presence; information ethics without empirical grounding lacks a mechanism of perturbation; and HRI without a field-theoretic model lacks an account of why minimal presence should matter at all.

In this respect, the thesis does not merely connect disparate literatures; it reorganises them around a common evaluative architecture. The experimental results show that artificial agents participate in human moral environments not through ethical reasoning, but through the way their perceptual and ontological profiles modulate the topology through which moral meaning becomes action. This unification is the conceptual kernel that ties the entire research programme together.

9.3 Final Synthesis and Closing Reflections

At the end of this research journey, the central phenomenon around which the thesis was constructed can be stated with clarity: *synthetic presence alters the conditions under which human beings register, evaluate, and act upon moral salience*. What began as a question about a quiet, non-interactive humanoid robot has unfolded into a broader insight about the structure of moral cognition and the ethical texture of technologically saturated environments.

The empirical work demonstrated that NAO's presence does not introduce new moral reasons, nor does it engage participants in explicit social interaction. Instead, it reshapes the evaluative conditions under which moral cues acquire behavioural force. This perturbation is subtle, probabilistic, and contingent upon latent dispositional architecture—yet empirically robust. Across inferential contrasts, cluster-specific analyses, and Bayesian estimation, the evidence converges on the same structural interpretation: artificial systems participate in human moral situations by modulating the perceptual and affective scaffolds from which intuitive moral appraisals develop.

What makes this more than an isolated result is the conceptual synthesis that underwrites it. The thesis has argued that moral behaviour is not a direct output of rules or principles; it is the trajectory of a cognitive–affective system shaped by attention, affect, expectation, and salience. Contemporary models of moral psychology have long emphasised the primacy of intuitive evaluation [16, 31, 17]. Floridi's Levels of Abstraction provide the correct ontological lens for specifying where synthetic entities exert influence. HRI and Affective Computing supply the methodological tools through which such influence can be observed and modelled. When joined together, these literatures do not merely complement one another—they reveal that the phenomenon under investigation requires all of them to be intelligible [239, 369].

A key outcome of this integration is the field-theoretic interpretation articulated throughout the thesis. The evaluative mapping

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

is not a formal artefact appended to the data; it is the conceptual structure that makes sense of how environmental cues, dispositional architectures, and artificial systems cooperate to shape moral behaviour. Moral appraisal unfolds within a dynamic field of salience. Synthetic presence acts as an operator on that field—not by issuing commands but by shifting the local geometry through which moral meaning is formed. The experimental results show that even minimal artificial bodies can introduce such a shift. This finding is modest in magnitude but profound in implication.

The ethical considerations developed in Chapter 8 sharpen this point. Traditional Machine Ethics asks how to embed principles into machines. Yet the empirical evidence here shows that artificial systems shape moral life not through principles but through presence: through their semiotic affordances, their perceived ontology, and the attentional demands they impose on human cognitive ecologies. In Floridi's framework, NAO is not a moral agent; it is a morally relevant informational object. Its influence is neither agential nor normative in itself, but environmental: it alters the field in which human agents construct their moral understanding. This is not a limit case or a marginal curiosity. It is a demonstration that ethical relevance emerges from the interaction between artefacts and the conditions of human moral cognition, long before questions of machine agency arise.

From this vantage point, the contribution of the thesis becomes unmistakable. The work provides an empirically grounded demonstration that artificial systems—humanoid robots today, pervasive AI systems tomorrow—can reshape moral cognition through their informational presence alone. It offers a formal architecture for modelling such influence, a set of empirical methods for detecting it, and a conceptual framework for understanding why it occurs. It also clarifies that the normative stakes of synthetic presence lie not merely in what machines may eventually *do*, but in how they silently reorganise the evaluative landscape in which human moral behaviour unfolds.

This brings us back, finally, to the thesis's animating question. Yes—synthetic presence can perturb the evaluative transformation through which moral salience becomes moral action. It can do so without interacting, without expressing norms, without engaging in dialogue, and without crossing the threshold of moral agency. It is enough that it is *perceived*. This insight opens the way for a new research programme in moral AI: one that takes seriously the structure of moral cognition, the ecological nature of moral environments, and the field-level effects of artificial systems embedded within them.

If this thesis has shown anything, it is that the future of moral AI will not begin with building artificial moral agents. It will begin with understanding how artificial systems already shape human moral life. The rest—ethical design, moral alignment, normative governance—must follow from that foundational fact. The

task now is no longer to ask whether such influence exists, but to learn how to measure it, model it, anticipate it, and ultimately, to govern the moral topologies we are already co-constructing with our synthetic companions.

This is the horizon toward which the work now points. The thesis closes here, but the evaluative field it uncovers is only beginning to take shape.

Bibliography

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

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

[3] M. Ernest-Jones, D. Nettle, and M. Bateson, "Effects of eye images on everyday cooperative behavior: A field experiment," *Evolution and Human Behavior*, vol. 32, no. 3, pp. 172–178, 2011.

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

[5] G. E. Dear, K. Dutton, and E. Fox, "The watching-eyes effect in the dictator game: A meta-analysis," *Evolution and Human Behavior*, vol. 40, no. 3, pp. 271–284, 2019.

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

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

[8] E. Gleichgerrcht and L. Young, "Low empathic concern predicts utilitarian moral judgment," *Cognition*, vol. 126, no. 3, pp. 364–372, 2013.

[9] J. Haidt, "The emotional dog and its rational tail: A social intuitionist approach to moral judgment," *Psychological Review*, vol. 108, no. 4, pp. 814–834, 2001.

[10] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen, "The neural bases of cognitive conflict and control in moral judgment," *Neuron*, vol. 44, no. 2, pp. 389–400, 2004.

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

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

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

[14] P. Foot, *Natural Goodness*. Oxford: Oxford University Press, 2001.

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

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

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

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

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

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

[21] S. Bringsjord, K. Arkoudas, and P. Bello, "Toward a modern synthesis of machine ethics," in *Proceedings of the AAAI Fall Symposium on Machine Ethics*, pp. 2–9, AAAI Press, 2006.

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

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

[24] M. Guarini, "Computational neural modeling and the philosophy of ethics: Reflections on the particularism-generalism debate," *Cambridge University Press*, 2006.

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

[26] L. Floridi, *The Philosophy of Information*. Oxford: Oxford University Press, 2011.

[27] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and Information Technology*, vol. 7, no. 3, pp. 149–155, 2005.

[28] K. Arkoudas and S. Bringsjord, "Toward ethical robots via mechanized deontic logic," in *Machine Ethics: AAAI Fall Symposium*, (Menlo Park, CA), pp. 17–23, AAAI Press, 2005.

[29] A. F. T. Winfield, M. Ortega, and R. Harper, "The ethical black box: An ai safety concept to facilitate ethics review and accountability," *IEEE Technology and Society Magazine*, vol. 38, no. 3, pp. 62–69, 2019.

[30] M. Anderson and S. L. Anderson, "Robot be good: A call for ethical autonomous machines," *Scientific American*, vol. 303, no. 4, pp. 72–77, 2010.

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

[32] J. Prinz, *The Emotional Construction of Morals*. Oxford: Oxford University Press, 2007.

[33] D. Kuchenbrandt, F. Eyssel, S. Bobinger, and M. Neufeld, "Minimal group-maximal effect? evaluation and anthropomorphization of the humanoid robot nao," in *International conference on social robotics*, pp. 104–113, Springer, 2011.

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

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

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

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

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

[39] D. J. Gunkel, *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA: MIT Press, 2012.

[40] 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.

[41] 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.

[42] 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.

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

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

[45] J. Andreoni, "Impure altruism and donations to public goods: A theory of warm-glow giving," *The Economic Journal*, vol. 100, no. 401, pp. 464–477, 1990.

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

[47] R. Noothigattu, M. Kleiman-Weiner, S. Dsouza, and et al., "Teaching and testing framework for human-ai value alignment," *Proceedings of the National Academy of Sciences*, vol. 118, no. 25, p. e1901956118, 2021.

[48] L. Ramirez, A. Aher, and A. Caliskan, "Computational moral psychology with large language models: Opportunities and risks," *Patterns*, vol. 4, no. 9, p. 100861, 2023.

[49] 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.

[50] 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.

[51] C. Breazeal, "Toward sociable robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 167–175, 2003.

[52] K. Fischer, "Interpersonal alignment in human–robot interaction," in *Proceedings of the 7th Annual ACM/IEEE International Conference on Human–Robot Interaction*, pp. 1–2, 2011.

[53] J. M. Doris, *Talking to Our Selves: Reflection, Ignorance, and Agency.* Oxford: Oxford University Press, 2015.

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

[55] W. Sinnott-Armstrong, "Framing moral intuitions," in *Moral Psychology, Volume 2: The Cognitive Science of Morality* (W. Sinnott-Armstrong, ed.), pp. 47–76, Cambridge, MA: MIT Press, 2008.

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

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

[58] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. Pentland, G. Robins, A. Shariff, J. B. Tenenbaum, and M. Wellman, "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, 2019.

[59] D. B. Shank and A. Gott, "Robot responsibility? the moral and legal dimensions of robotics," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Daegu, Republic of Korea), pp. 715–717, IEEE, 2019.

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

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

[62] Z. Jin, H. Zhang, T. Ge, and M. Zeng, "Moral foundations of large language models," *arXiv preprint arXiv:2205.12329*, 2022.

[63] N. Scherrer, E. Clark, and N. A. Smith, "Evaluating moral reasoning in large language models," *arXiv preprint arXiv:2306.00030*, 2023.

[64] 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.

[65] 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.

[66] P. Charlton and D. Danks, "Large language models show human-like moral dynamics," *arXiv preprint arXiv:2308.13129*, 2023.

[67] D. Hendrycks *et al.*, "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2021.

[68] 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.

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

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

[71] 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.

[72] J. Zaki and K. N. Ochsner, "The neuroscience of empathy: Progress, pitfalls, and promise," *Nature Neuroscience*, vol. 15, no. 5, pp. 675–680, 2012.

[73] 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.

[74] 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.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] A. Kasirzadeh and I. Gabriel, "The mirage of moral agency in large language models," *Philosophy & Technology*, vol. 37, no. 1, pp. 1–26, 2024.

[81] 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.

[82] J. Gardner and et al., "Models that write like moral agents are not moral agents," *AI & Society*, 2024. Forthcoming.

[83] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of Experimental Social Psychology*, vol. 52, pp. 113–117, 2014.

[84] F. Eyssel and D. Kuchenbrandt, "Social categorization of social robots: Anthropomorphism as a function of robot group membership," *British Journal of Social Psychology*, vol. 51, no. 4, pp. 724–731, 2012.

[85] B. K. Kim and S. Schubert, "Moral dynamics in human?ai interaction: How ai presence shifts moral judgment," *Cognition*, vol. 234, p. 105369, 2023.

[86] J. Wirtz, P. Patterson, W. H. Kunz, T. Gruber, V. Lu, and S. Paluch, "Brave new world: Service robots in the frontline," *Journal of Service Management*, vol. 29, no. 5, pp. 907–931, 2018.

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

[88] 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.

[89] 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.

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

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

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

[93] M. Slote, *Moral Sentimentalism*. Oxford: Oxford University Press, 2010.

[94] 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.

[95] M. Ekström, "Do watching eyes affect charitable giving? evidence from a field experiment," *Experimental Economics*, vol. 15, no. 3, pp. 530–546, 2012.

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

[97] 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.

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

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

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

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

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

[103] J. Annas, *Intelligent Virtue*. Oxford: Oxford University Press, 2011.

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

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

[106] C. Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. New York: Oxford University Press, 2016.

[107] I. Kant, *Groundwork of the Metaphysics of Morals*. Cambridge University Press, 1785. Modern edition.

[108] I. Kant, *Critique of Practical Reason*. Cambridge: Cambridge University Press, 1788. Original work published 1788.

[109] W. D. Ross, *The Right and the Good*. Oxford: Oxford University Press, 1930.

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

[111] D. Hume, *An Enquiry Concerning the Principles of Morals*. Oxford University Press, 1751.

[112] A. Smith, *The Theory of Moral Sentiments*. Cambridge: Cambridge University Press, 1759. Edited by D. D. Raphael and A. L. Macfie (1976 edition).

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

[114] Aristotle, *Nicomachean Ethics*. Oxford University Press, 2000. Trans. Terence Irwin.

[115] P. Foot, "The problem of abortion and the doctrine of double effect', in her virtues and vices," *Berkeley and Los Angeles: University ofCalifornia Press. FootThe Problem of Abortion and the Doctrine of the Double Effect19Virtues and Vices1978*, pp. 19–32, 1978.

[116] A. MacIntyre, *After Virtue*. University of Notre Dame Press, 1981.

[117] J. Bentham, *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press, 1789.

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

[119] H. Sidgwick, *The Methods of Ethics*. London: Macmillan, 1874.

[120] P. Singer, *Practical Ethics*. Cambridge University Press, 1979.

[121] D. Parfit, *Reasons and Persons*. Oxford University Press, 1984.

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

[123] J. Locke, *Two Treatises of Government*. Cambridge University Press, 1689.

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

[125] J. Rawls, *A Theory of Justice*. Harvard University Press, 1971.

[126] J. Rawls, *Political Liberalism*. Columbia University Press, 1993.

[127] C. Darwin, *The Descent of Man*. John Murray, 1871.

[128] R. Trivers, "The evolution of reciprocal altruism," *Quarterly Review of Biology*, vol. 46, pp. 35–57, 1971.

[129] E. O. Wilson, *Sociobiology: The New Synthesis*. Harvard University Press, 1975.

[130] C. Boehm, *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. Basic Books, 2012.

[131] S. Baron-Cohen, *The Essential Difference: The Truth about the Male and Female Brain*. London: Penguin, 2003.

[132] 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.

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

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

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

[136] 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.

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

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

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

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

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

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

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

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

[145] 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.

[146] 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.

[147] 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.

[148] 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.

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

[150] 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.

[151] 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.

[152] 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.

[153] 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.

[154] 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.

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

[156] 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.

[157] 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.

[158] 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.

[159] 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.

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

[161] 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.

[162] 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.

[163] 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.

[164] 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.

[165] 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.

[166] 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.

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

[168] 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.

[169] 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.

[170] 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.

[171] F. Cushman, L. Young, and J. D. Greene, "Multi-system moral psychology," *The Oxford Handbook of Moral Psychology*, pp. 47–71, 2010. Often cited for the "value integration" framework formalising the combination of affective, deontic, and goal-directed inputs into action selection.

[172] 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.

[173] 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.

[174] 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.

[175] 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.

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

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

[178] 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.

[179] 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.

[180] 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.

[181] 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.

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

[183] 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.

[184] 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.

[185] 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.

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

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

[188] F. Cushman, L. Young, and M. Hauser, "The role of conscious reasoning and intuition in moral judgment: testing three principles of harm," *Psychological Science*, vol. 17, no. 12, pp. 1082–1089, 2006.

[189] L. Young, F. Cushman, M. Hauser, and R. Saxe, "The neural basis of the interaction between theory of mind and moral judgment," *Proceedings of the National Academy of Sciences*, vol. 107, no. 20, pp. 8374–8379, 2010.

[190] M. J. Crockett, L. Clark, M. D. Hauser, and T. W. Robbins, "Serotonin selectively influences moral judgment and behavior through harm aversion," *Proceedings of the National Academy of Sciences*, vol. 107, no. 33, pp. 14381–14386, 2010.

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

[192] J. L. Tracy and R. W. Robins, "Putting the self into self-conscious emotions: A theoretical model," *Psychological Inquiry*, vol. 15, no. 2, pp. 103–125, 2004.

[193] S. Whitmarsh, O. Bartra, B. Love, and S. Gluth, "Affective and attentional dynamics predict moral decision-making," *Nature Communications*, vol. 12, no. 1, pp. 1–13, 2021.

[194] 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.

[195] S. Baron-Cohen, "Autism: The empathizing?systemizing (e?s) theory," *Trends in Cognitive Sciences*, vol. 13, no. 6, pp. 274–280, 2009.

[196] O. P. John, E. M. Donahue, and R. L. Kentle, "The big five inventory?versions 4a and 54," tech. rep., University of California, Berkeley, Institute of Personality and Social Research, 1991.

[197] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research* (L. A. Pervin and O. P. John, eds.), pp. 102–138, Guilford Press, 1999.

[198] 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.

[199] 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.

[200] 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.

[201] A. Wakabayashi, S. Baron-Cohen, S. Wheelwright, N. Goldenfeld, J. Delaney, D. Fine, and R. Smith, "Development of short forms of the empathy quotient (eq-short) and the systemizing quotient (sq-short)," *Personality and Individual Differences*, vol. 41, no. 5, pp. 929–940, 2006.

[202] 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.

[203] 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.

[204] 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.

[205] T. Yamagishi, N. Mifune, Y. Li, M. Shinada, H. Hashimoto, Y. Horita, A. Miura, K. Inukai, S. Tanida, T. Kiyonari, and H. Takagishi, "Is behavioral pro?sociality game?specific? pro?social preference and expectations of pro?sociality," *Journal of Experimental Social Psychology*, vol. 53, pp. 1–11, 2014.

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

[207] 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.

[208] 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.

[209] 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.

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

[211] S. Baron-Cohen, "Autism and the empathizing-systemizing (es) theory," *Developmental social cognitive neuroscience*, pp. 125–138, 2009.

[212] 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.

[213] K. Fritz, A. Dörfler, N. Gütling, H.-J. Mohr, M. Wöhr, and A. Schmitt, "Systemizing tendency predicts neural activation during logical reasoning in typical adults," *NeuroImage*, vol. 142, pp. 553–561, 2016.

[214] S. Wheelwright, S. Baron-Cohen, N. Goldenfeld, J. Delaney, D. Fine, R. Smith, L. Weil, and A. Wakabayashi, "Predicting autism spectrum quotient (aq) from the systemizing quotient?revised (sq-r): Evidence for a cognitive style dimension," *Journal of Autism and Developmental Disorders*, vol. 36, no. 6, pp. 863–872, 2006.

[215] C. Ecker, S. Bookheimer, and D. Murphy, "Neuroanatomy of autism: A systematic review of structural mri studies," *Neuroscience and Biobehavioral Reviews*, vol. 59, pp. 417–429, 2016.

[216] S. Baron-Cohen, H. Ring, S. Wheelwright, E. T. Bullmore, M. J. Brammer, A. Simmons, and S. C. R. Williams, "The amygdala theory of autism revisited: increased systemizing neural networks during rule-based reasoning," *Brain*, vol. 128, no. 5, pp. 1133–1144, 2005.

[217] A. Chakroff, J. Dungan, and L. Young, "Harmful situations, impure people: Emotion, intention, and the dynamics of moral judgment," *Personality and Social Psychology Bulletin*, vol. 42, no. 8, pp. 1092–1112, 2016.

[218] J. M. Paxton and J. D. Greene, "Moral reasoning: Hints and allegations," *Topics in Cognitive Science*, vol. 2, no. 3, pp. 511–527, 2010.

[219] B. De Martino, D. Kumaran, B. Seymour, and R. Dolan, "Frames, biases, and rational decision-making in the human brain," *Science*, vol. 313, no. 5787, pp. 684–687, 2006.

[220] I. Krajbich, C. Camerer, and J. Ledyard, "Using neural data to test a theory of strategic decision-making in games," *American Economic Review*, vol. 105, no. 10, pp. 3306–3337, 2015.

[221] J. Grice, T. McDaniel, and D. Lynam, "Personality, reasoning style, and susceptibility to decision framing," *Journal of Behavioral Decision Making*, vol. 30, no. 2, pp. 564–578, 2017.

[222] S. Thellman, A. Silvervarg, and T. Ziemke, "Folk-psychological attributions to humanoid robots: The role of interaction and appearance," *Frontiers in Psychology*, vol. 8, p. 1323, 2017.

[223] N. Spatola, S. Marchesi, and A. Wykowska, "When robots are perceived as more autonomous than humans: The role of analytic cognitive style," *Scientific Reports*, vol. 11, p. 3127, 2021.

[224] 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.

[225] R. R. McCrae and P. T. Costa, "The five-factor theory of personality," *Handbook of Personality: Theory and Research*, pp. 159–181, 2008.

[226] D. C. Funder, "Personality," *Annual Review of Psychology*, vol. 52, pp. 197–221, 2001.

[227] B. W. Roberts, D. Wood, and A. Caspi, "Personality development," *Annual Review of Psychology*, vol. 57, pp. 283–306, 2006.

[228] C. G. DeYoung, "Personality neuroscience and the biology of traits," *Social and Personality Psychology Compass*, vol. 4, no. 12, pp. 1165–1180, 2010.

[229] P. Slovic, M. L. Finucane, E. Peters, and D. G. MacGregor, "The affect heuristic," *European Journal of Operational Research*, vol. 177, no. 3, pp. 1333–1352, 2007.

[230] 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.

[231] 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.

[232] H. Gintis, "Strong reciprocity and human sociality," *Journal of Theoretical Biology*, vol. 206, no. 2, pp. 169–179, 2000.

[233] F. Warneken and M. Tomasello, "Altruistic helping in human infants and young chimpanzees," *Science*, vol. 311, no. 5765, pp. 1301–1303, 2006.

[234] C. D. Batson, *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

[235] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, pp. 785–791, 2003.

[236] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 143–166, 2003.

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

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

[239] J. Katsyri, K. Forger, M. Makarainen, and T. Takala, "A review and meta-analysis of the uncanny valley: Toward a quantified model," *Frontiers in Psychology*, vol. 6, p. 390, 2015.

[240] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 177–190, 2003.

[241] V. Groom, C. Nass, and T. Chen, "Computers as social actors: A review of the current paradigm," in *Proceedings of the AISB Convention*, 2009.

[242] B. F. Malle and M. Scheutz, "Moral competence in social robots," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 474–490, 2019.

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

[244] 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.

[245] 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.

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

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

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

[249] 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.

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

[251] 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.

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

[253] 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.

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

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

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

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

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

[259] 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.

[260] 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.

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

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

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

[264] 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.

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

[266] 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.

[267] S. Krach, F. Hegel, B. Wrede, G. Sagerer, G. Bente, and T. Kircher, "Can machines think? interaction and perspective taking with robots investigated via fmri," *PLoS ONE*, vol. 3, no. 7, p. e2597, 2008.

[268] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of Experimental Social Psychology*, vol. 52, pp. 113–117, 2014.

[269] I. Rae, B. Mutlu, and J. Buss, "Walking in another?s shoes: Anthropomorphic agents reduce empathy in human?robot interaction," in *Proceedings of the 8th ACM/IEEE International Conference on Human?Robot Interaction*, pp. 157–164, 2013.

[270] S. Kühn, M. Brass, and J. Gallinat, "Taxing the brain to understand others: Perceived humanlike behavior in machines reduces brain activity in areas important for mentalizing," *Social Neuroscience*, vol. 6, no. 4, pp. 369–377, 2011.

[271] T. C. Burnham and B. Hare, "Engineering human cooperation: Does involuntary neural activation increase trust?," *Human Nature*, vol. 18, no. 2, pp. 115–121, 2007.

[272] M. Bateson, D. Nettle, and G. Roberts, "Cues of being watched enhance cooperation: A meta-analysis," *Evolution and Human Behavior*, vol. 36, no. 1, pp. 9–17, 2015.

[273] T. Chaminade and M. Kawato, "Predictive coding in human?robot interaction: A review," *Neural Networks*, vol. 39, pp. 62–68, 2013.

[274] E. B. Sandoval, J. Brandstetter, M. Obaid, and C. Bartneck, "Reciprocity in human?robot interaction: A quantitative approach through the prisoner?s dilemma and the ultimatum game," in *Proceedings of the 11th ACM/IEEE International Conference on Human?Robot Interaction*, pp. 303–310, 2016.

[275] A. Vinciarelli, A. Esposito, M. Tayarani, G. Roffo, F. Scibelli, F. Perrone, and D.-B. Vo, "We are less free than how we think: Regular patterns in nonverbal communication," in *Multimodal Behavior Analysis in the Wild*, pp. 269–288, Elsevier, 2019.

[276] 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.

[277] K. Gray, H. M. Gray, and D. M. Wegner, "The moral dyad: A fundamental template unifying moral judgment," *Psychological Inquiry*, vol. 23, no. 2, pp. 206–215, 2012.

[278] D. A. Baldwin, "Infants' contribution to the achievement of joint reference," *Child Development*, vol. 56, no. 4, pp. 875–890, 1985.

[279] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in Cognitive Sciences*, vol. 8, no. 8, pp. 299–309, 2004.

[280] I. Gabriel, "Artificial intelligence, values, and alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411–437, 2020.

[281] E. Hutchins, *Cognition in the Wild*. MIT Press, 1995.

[282] A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, 2008.

[283] L. Floridi, "The ethics of artificial intelligence: Large language models and the new semantic environment," *Philosophy & Technology*, 2024.

[284] L. Speri and M. Alfano, "Moral ecologies and the ethics of artificial agents," *Ethics and Information Technology*, 2023.

[285] R. E. Nisbett and T. D. Wilson, "Telling more than we can know: Verbal reports on mental processes," *Psychological Review*, vol. 84, no. 3, pp. 231–259, 1977.

[286] T. D. Wilson, *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Harvard University Press, 2002.

[287] J. Decety and J. M. Cowell, "The complex relation between morality and empathy," *Trends in Cognitive Sciences*, vol. 17, no. 7, pp. 337–339, 2013.

[288] E. Goffman, *Interaction Ritual: Essays on Face-to-Face Behavior*. New York: Pantheon Books, 1967. Original essays published 1955–1967.

[289] J. S. Lee and S. Kiesler, "Human mental models of humanoid robots," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 586–593, 2010.

[290] G. Gigerenzer, *Rationality for Mortals*. Oxford University Press, 2008.

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

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

[293] J. Knobe, "Person as scientist, person as moralist," in *Explanation and Cognition* (P. A. Machamer and M. Thagard, eds.), pp. 119–136, MIT Press, 2010.

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

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

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

[297] 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.

[298] 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.

[299] N. Lemos, *An Introduction to the Theory of Knowledge*. Cambridge University Press, 2nd ed., 2020.

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

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

[302] 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.

[303] 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.

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

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

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

[307] A. Smith, *The Theory of Moral Sentiments*. Cambridge University Press, 2002. Originally 1759.

[308] R. J. Wallace, *The View From Here: On Affirmation, Attachment, and the Limits of Regret*. Oxford University Press, 2012.

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

[310] P. Railton, "Moral realism," *Philosophical Review*, vol. 95, no. 2, pp. 163–207, 1984.

[311] J. Prinz, *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press, 2004.

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

[313] 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.

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

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

[316] S. Nichols, *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press, 2004.

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

[318] 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.

[319] 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.

[320] 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.

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

[322] 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.

[323] 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.

[324] 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.

[325] 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.

[326] J. A. Bargh and T. L. Chartrand, "The unbearable automaticity of being.," *American psychologist*, vol. 54, no. 7, p. 462, 1999.

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

[328] J. Griffin, *Well-Being*. Oxford: Oxford University Press, 1986.

[329] M. Stocker, *Plural and Conflicting Values*. Oxford: Oxford University Press, 1990.

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

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

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

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

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

[335] P. Lin, G. Bekey, and K. Abney, "Autonomous military robotics: Risk, ethics, and design," tech. rep., California Polytechnic State Univ San Luis Obispo, 2008.

[336] K. Atkinson and T. Bench-Capon, "Action-based alternating transition systems for arguments about action," in *AAAI*, vol. 7, pp. 24–29, 2007.

[337] 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.

[338] M. Hjelmblom, *Deontic action-logic multi-agent systems in Prolog*. Högskolan i Gävle, 2008.

[339] 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.

[340] 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.

[341] 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.

[342] A. R. Honarvar and N. Ghasem-Aghaee, "An artificial neural network approach for creating an ethical artificial agent," in *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA)*, pp. 290–295, 2009.

[343] 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.

[344] 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.

[345] R. Montague and R. H. Thomason, "Formal philosophy. selected papers of richard montague," *Erkenntnis*, vol. 9, no. 2, 1975.

[346] R. Carnap, *Introduction to symbolic logic and its applications*. Courier Corporation, 2012.

[347] L. M. Pereira and A. Saptawijaya, "Modeling morality with prospective logic," *Cambridge University Press*, 2007.

[348] "The problem of machine ethics in artificial intelligence," *AI and SOCIETY*, vol. 35, no. 1, pp. 103–111, 2020.

[349] J. McDermid, V. C. Muller, T. Pipe, Z. Porter, and A. Winfield, "Ethical issues for robotics and autonomous systems," 2019.

[350] D. Howard and I. Muntean, "Artificial moral cognition: moral functionalism and autonomous moral agency," in *Philosophy and computing*, pp. 121–159, Springer, 2017.

[351] M. Pantic and A. Vinciarelli, "Social signal processing," *The Oxford handbook of affective computing*, p. 84, 2014.

[352] R. W. Picard, *Affective computing*. MIT press, 2000.

[353] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

[354] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Cambridge, MA: Perseus Books, 1994.

[355] R. D. Beer, "A dynamical systems perspective on agent–environment interaction," *Artificial Intelligence*, vol. 72, no. 1–2, pp. 173–215, 1995.

[356] L. B. Smith and E. Thelen, "Development as a dynamic system," *Trends in Cognitive Sciences*, vol. 7, no. 8, pp. 343–348, 2003.

[357] K. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[358] 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.

[359] 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.

[360] T. Hollenstein, *State Space Grids: Depicting Dynamics Across Development*. New York: Springer, 2015.

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

[362] M. A. Goodrich and A. C. Schultz, "Human–robot interaction: A survey," *Foundations and Trends in Human–Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[363] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human–robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[364] R. A. Calvo and S. D'Mello, *Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications*, vol. 1. IEEE Transactions on Affective Computing, 2010.

[365] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[366] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[367] J. K. Kruschke, *Doing Bayesian Data Analysis*. Academic Press, 2nd ed., 2014.

[368] 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.

[369] K. Fischer, "How social is a robot? minimal social cues and their cognitive consequences," *Interaction Studies*, vol. 24, no. 1, pp. 1–29, 2023.