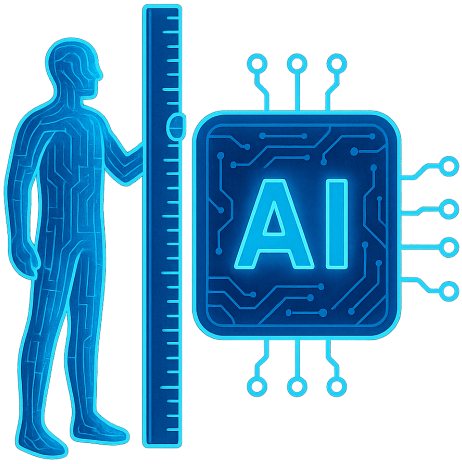
Measuring the Machine



Evaluating Generative AI as Pluralist Sociotechnical Systems

Rebecca Lynn Johnson

B.A., B.Sc., M.A.(Res).
ORCID: https://orcid.org/0000-0001-7321-0744
EthicsGenAI.com

PhD thesis submitted and currently under examination.

This document consists of:
Thesis highlights including introduction, chapter abstracts, and a sneak-peak at the conclusion.

Research Questions

Measurement: How can generative AI be evaluated in ways that surface the normative assumptions embedded in sociotechnical systems?

Responsibility: What does it mean to evaluate AI responsibly in a world of value pluralism, so that evaluation reveals rather than prescribes?

Co-construction: In what ways do generative systems co-construct values with humans and institutions, and how can evaluation make this co-construction empirically legible?

"The Analytical Engine. . .might act upon other things besides number, were objects found whose mutual fundamental relations could be expressed by those of the abstract science of operations, and which should be also susceptible of adaptations to the action of the operating notation and mechanism of the engine. Supposing, for instance, that the fundamental relations of pitched sounds in the science of harmony and of musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent. **The Analytical Engine weaves algebraical patterns just as the Jacquard loom weaves flowers and leaves.**"

Ada Lovelace, 1843 [260]

"The Jacquard loom remains in modern AI, but its thread is human values, its patterns our interpretations: what we measure, we amplify."

Rebecca L. Johnson, 2025

Thesis Abstract

In measurement theory, instruments do not simply record reality — they help constitute what is observed. The same holds for generative AI evaluation: benchmarks do not just measure, they shape what models appear to be. Functionalist benchmarks, rooted in computationalist assumptions, treat models as isolated predictors, while normative, prescriptive benchmarks frame evaluation in terms of what systems ought to be. Both approaches obscure the sociotechnical dynamics through which meaning and values are enacted. In a pluralist world, such measures risk reifying narrow cultural epistemologies and marginalising alternative value perspectives.

This thesis advances a descriptive alternative: responsible evaluation must treat generative AI as a pluralist sociotechnical system. It develops MaSH Loops (Machine–Society–Human-in-the-loop), an original framework that traces how models, people, and institutions recursively co-construct meaning and values. From this stance, evaluation becomes less about declaring what a model *ought to be* and more about revealing what it *is* and how it *enacts* values in interaction with users and society.

Chapter 1 situates Responsible AI debates (circa 2023) within deeper epistemological rifts (functionalism versus constructivism) and introduces enactivism as a bridge. It also introduces MaSH Loops as a broad framework. **Chapter 2** documents value drift in early GPT-3 (2021), documenting how culturally charged texts were reframed through normative "accents" when parsed through the model. It preserves a historical record of a now-vanished model and demonstrates the need for descriptive, pluralist evaluation methods. **Chapter 3** applies Responsible AI (RAI) concepts to real estate; translating abstract ideas into a domain that directly affects markets and housing. Based on a chapter written for an academic textbook, thereby showing a concrete application of many of the ideas discussed in earlier chapters. It equips educators with tools and case studies to communicate RAI to real world problems. **Chapter 4** develops the methodological core developed during my time at Google: The World Values Benchmark (WVB), a distributional evaluation framework measuring alignment of model outputs with the World Values Survey. The WVB demonstrates how controlling for prompt sensitivity and anchor bias yields more stable, contestable profiles, revealing both US training imprints and cross-cultural variation in aggregate placement. **Chapter 5** extends the

enactivist stance through participatory realism, drawing on quantum mechanics to argue that prompting collapses semantic potential into enacted outcomes, making evaluation itself a world-making act.

The contributions include:

1. **Theory**: MaSH Loops, an enactivist framework for evaluating generative AI as recursive sociotechnical systems. Quantum participatory realism as a model to better describe human and societal interactions with LLM machines.
2. **Method**: WVB, a distributional benchmark that renders normative assumptions empirically legible without prescribing outcomes.
3. **Application**: Case studies spanning GPT-3's "American accent" and applied mapping in real estate, showing how methods themselves become findings.

Ultimately, the thesis advances the claim that generative AI cannot be evaluated adequately through static, functionalist benchmarks. Responsible evaluation requires pluralist, recursive frameworks that make visible whose values are being enacted. By reconceptualising evaluation from scores to sociotechnical processes, this work contributes to more inclusive, culturally responsive practices in AI governance, with direct implications for research practice, policy design, and public trust.

Introduction

Catching the Tiger’s Tail

"We have to remember that what we observe is not nature herself, but nature exposed to our method of questioning”

Werner Heisenberg (1958) [165]

Introduction: Catching the Tiger's Tail.

At the cross-currents of Generative AI and the Philosophy of AI, this thesis asks what it means to grasp the tiger's tail amid turbulence and speed. It does so by drawing on a deliberately wide set of traditions: philosophy of mind, measurement theory, ethics, cybernetics, cognitive science, quantum mechanics, participatory realism, sociotechnical systems theory, sociology, and moral value pluralism. These are not scattered ornaments, but carefully chosen tools, each brought in to clarify specific aspects of a technology. Philosophy of AI is still nascent; definitions are unsettled, frameworks are contested, and methods are in flux.

Alongside philosophers, this unsettled space has attracted computer scientists and engineers who turn eagerly to philosophy, though at times without engaging its depth. Their contributions are valuable but sometimes produce a patchwork *Franken-philosophy of AI*: conceptual borrowing without context, sociotechnical theories misapplied, or ethical categories flattened into engineering checklists. Such moves risk distorting the very traditions they draw from, obscuring rather than clarifying the nature of generative systems.

Working in this space requires constant code-switching. With philosophers, the pace and opacity of technical change can feel like a moat; part of my role is to lower the drawbridge, making models, data, and evaluation details legible without jargon. With engineers, I'm asked to show why uncovering normative assumptions and applying philosophical measurement theory matter. With policymakers, I translate pluralist arguments into practical and accountable governance recommendations. These shifts are rarely easy, and disciplinary silos often solidify. The AI ethics and safety debates of 2022–2023 demonstrated this vividly, as factions closed ranks around existential risk or sociotechnical harm, leaving little space for dialogue across paradigms.

The pace of the field compounds these tensions. Research on ethical and responsible generative AI has exploded, with more papers appearing each week than any one researcher can absorb. Release papers from major firms often prioritise capability claims, treat ethics as an afterthought, and circulate without external review; gaining approval through ethos rather than independent scrutiny. At the same time, the glacier-slow speed of peer review means that preprints and arXiv drafts dominate discourse, shaping debate before ideas are carefully

vetted. In such conditions, philosophical clarity and methodological rigour become essential: without them, debates risk being built on unstable ground.

This thesis is about chasing systems that refuse to hold still, and the equally shifting instruments we use to measure them. Generative AI (GenAI) is one such system: contingent, probabilistic, and deeply entangled with the societies that build and use it. To evaluate these systems responsibly, we need more than prescriptive benchmarks designed to populate leaderboards measuring machines against predetermined notions of success in the race to the "best" AI. Instead we need frameworks that reveal what values are embedded and reflected within them, and how those value patterns relate to different societies and communities.

The work presented here takes that instability as its starting point. It asks a simple question: what are we measuring when we say a model "understands," has "commonsense reasoning," or "aligns with human values"? My answer is a programme for evaluation that treats GenAI not as an isolated predictor of the next text token but as a participant in Machines, Societies, and Humans in-the-loop. I call this the MaSH Loops framework. On that foundation, I build the World Values Benchmark (WVB): a distributional, descriptive method that compares a model's *value profiles* against social-science baselines, controlling for prompt sensitivity and anchor bias. I demonstrate with examples how these evaluation design choices make normative assumptions empirically legible rather than invisible. Using case studies such as (i) a historical study of GPT-3's "American accent" in 2021, and (ii) applied work in the real-estate sector, where evaluation choices carry direct human and policy consequences. Together, these contributions reframe evaluation from performance ranking to relational measurement that surfaces whose values are being enacted.

The project began with a little existential angst: powerful systems were arriving fast, and the ethical guardrails looked thin. Stories like Buolamwini's *Gender Shades* [56] and other early work on bias [290, 356, 406] in deployed systems made clear that measurement failures could translate into real harm. I wanted to understand not only how values enter systems, but how we might measure those movements without collapsing plural perspectives into a normative-driven single score.

What began as concern about thin ethical guardrails matured into a recognition that evaluation practices are consequential. The Coda returns to this theme, showing how measures not only describe but also steer, amplifying some values while erasing others.

That pursuit shaped my research journey. In early 2021, I fought for months for access to GPT-3, finally receiving the "green light" on 25 May. With a small group of PhD peers, we began exploratory tests that revealed cultural value drift; work that seeded Chapter 2. In parallel, I founded the *PhD Students in AI Ethics* network, which grew to 400+ members across the world within just a few months: reinforcing the collective sense that we were all grasping something both urgent and under-defined.

From 2021 to 2022, a year's internship at Google Research (in the Ethical AI team) shifted my focus. Insider access to LaMDA and PaLM revealed a deeper question beyond *what can models do?* and instead *what do our measurement choices make them appear to do?* I read every Large Language Model (LLM) release paper like a digital archaeologist, poring over appendices to excavate hidden assumptions: proxy tasks, fragile validity claims, missing contexts. This thesis records that excavation: the attempt to catch the tiger's tail [1] not only of the models themselves, but of the evaluative practices racing to contain them.

Between 2020 and 2025, the ground kept moving. Models shifted from closed-door APIs to mass public adoption. Benchmarks proliferated, often treated as definitive leaderboards, even when their constructs needed deep scrutiny. Media discourse amplified existential-risk narratives and near-consciousness hype promoted by some AI factions, while questions of immediate sociotechnical impact and measurement validity often struggled for oxygen. I was pulled into those debates on social media, at academic conferences, and in public media interviews. I was frequently asked in these interviews why some of the developers saw species-level threats while others, including many AI ethicists, took a different stance.

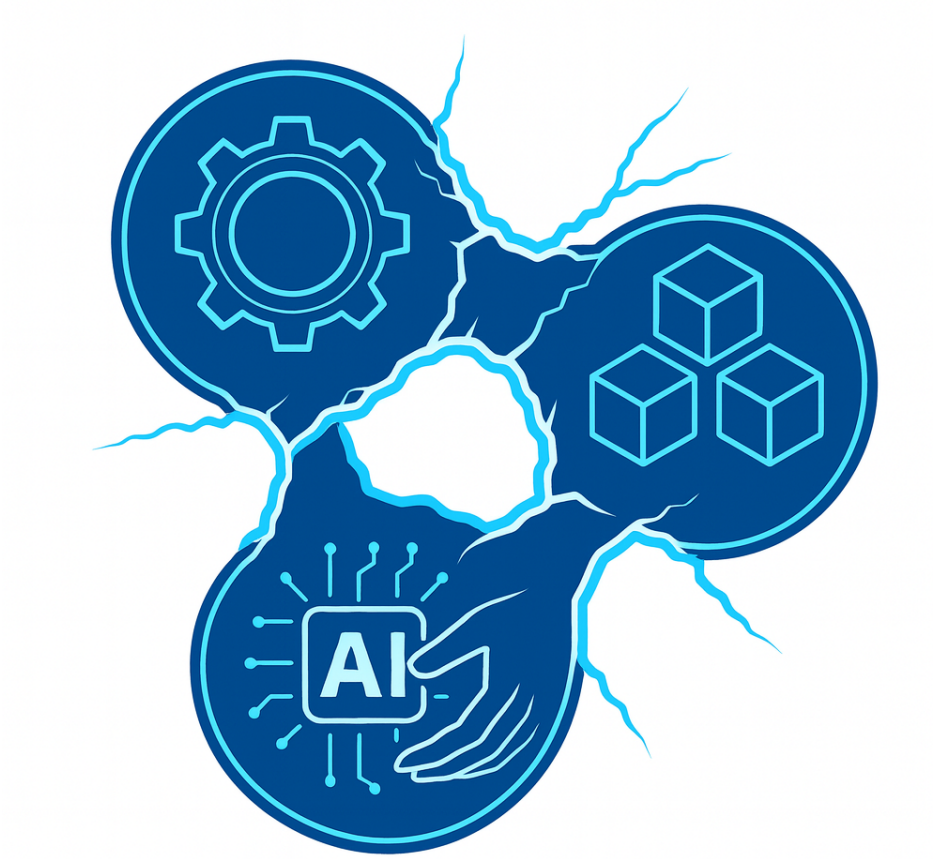
The answer I arrived at is not a single solution—there are no silver bullets in AI—but a shift in perspective. Evaluation should be descriptive, pluralist, and enactivist. It should

[1] From Burmese tradition, "grasping the tiger's tail" means being trapped in danger: you cannot let go safely, yet holding on is perilous. I use it here to describe the Philosophy of AI's engagement with Generative AI: unavoidable, precarious, and where the danger lies as much in our methods of measurement as in the systems themselves.

capture distributions rather than single verdicts, reveal assumptions rather than conceal them, and map recursive MaSH Loops rather than isolating outputs. The epistemological conflicts of the AI debates were not just about risk itself but about *how risk was being measured*. Many trained in functionalist traditions of engineering and computer science gravitated toward computationalism as a philosophy of mind, leading them to interpret machine behaviour through functionalist assumptions. My contribution is to show that such assumptions are not neutral: they are design choices embedded in our instruments of measurement.

ABRIDGED

THIS CHAPTER CONTINUES TO SET THE SCENE AND LAY THE GROUND WORK FOR THE REST OF THE THESIS. YOU CAN READ THE COMPLETE SECTION WHEN THE THESIS IS PUBLISHED.



Epistemological Rumbles in Responsible AI.

> “Cognition is not the grasping of an independent, outside world by a separate mind or self, but instead the bringing forth or enacting of a dependent world of relevance in and through embodied action.”
>
> Varela, Thompson, Rosch [411]

CHAPTER 1: EPISTEMOLOGICAL RUMBLES IN RESPONSIBLE AI.

Ethics and Safety through the lenses of Functionalism, Constructivism, and Enactivism.

Abstract

In 2023, fractures in the Responsible AI community became impossible to ignore. What looked like policy disagreements were rooted in conflicting epistemologies. Functionalist approaches, which dominate AI Safety and benchmarking, treat models as input–output devices whose performance can be scored and compared. Constructivist methods, central to AI Ethics and STS, uncover the sociotechnical embedding of systems and the normative assumptions they carry. Both perspectives illuminate important aspects of AI, yet neither fully accounts for the recursive, adaptive nature of today's generative systems.

This chapter argues that a third stance is needed. Enactivism reframes intelligence not as a static property but as relational and participatory. From this perspective, evaluation is less about discovering what a model *is* and more about observing how it *becomes* in interaction with humans and institutions. To make this operational, I introduce MaSH Loops—Machine–Society–Human—as an enactivist framework for evaluation. MaSH Loops show how models co-evolve with social practices, regulatory incentives, and everyday use, shifting the unit of analysis from isolated outputs to recursive sociotechnical processes.

The analysis demonstrates that functionalism and constructivism each miss the recursive character of generative AI, while MaSH Loops provide criteria that better capture situated responsiveness and participatory alignment. This is not a silver bullet but a shift in stance: from static benchmarks to relational measurement.

The impact of this chapter is twofold. Conceptually, it establishes the epistemological foundation for the thesis. Practically, it motivates the methodological innovations developed in later chapters, especially the World Values Benchmark, and offers a framework for evaluations that are descriptive, pluralist, and contestable.

The Ghost in the Machine has an American Accent (2021)

> “Our challenge is not to erase moral difference, but to learn how to live with it responsibly, in ways that sustain global cooperation without demanding global moral uniformity.”
>
> Shannon Vallor (2016) [406]

CHAPTER 1: THE GHOST IN THE MACHINE HAS AN AMERICAN ACCENT.

Exploratory Evidence of Cultural Value Drift in Early GPT-3.

Abstract

Early large language models were released with minimal alignment, providing a valuable glimpse into how generative systems reframed the ethical values embedded in human texts. This chapter examines outputs from a 2021 version of OpenAI's base GPT-3, using prompts that asked it to summarise culturally diverse source materials including laws, political speeches, and philosophical works. Interpreted through a descriptive, pluralist lens, these outputs reveal systematic *value drift*; the tendency of models to invert or overwrite normative content along familiar cultural axes.

Examples were often striking. Australia's firearm legislation, framed around public safety, re-emerged as a warning of lost liberty. Simone de Beauvoir's feminist critique was recast as gender-essentialist dating advice. Angela Merkel's humanitarian appeal became immigration control. By contrast, consensus-crafted multilateral documents such as UN and UNESCO statements showed greater value stability, suggesting that deliberately negotiated language may buffer against cultural mutation.

The analysis makes two contributions. First, it provides historical evidence that unaligned models could systematically transform value-laden texts in predictable ways, surfacing the cultural "accent" of their training distributions. Second, it demonstrates a pluralist, descriptive evaluation method that situates outputs against cross-national baselines such as the World Values Survey, showing whose values dominate and under what conditions.

The impact of this chapter is archival as well as methodological. It preserves a record of normative behaviours from an early, now-vanished system, and establishes why descriptive, culturally inclusive evaluation is essential for assessing alignment in contemporary generative AI.

The Model is not the Market (2025)

"The map is not the territory."
Alfred Korzybski (1931) [204]

CHAPTER 2: THE MODEL IS NOT THE MARKET

Applying Responsible-AI concepts to the Real Estate Industry

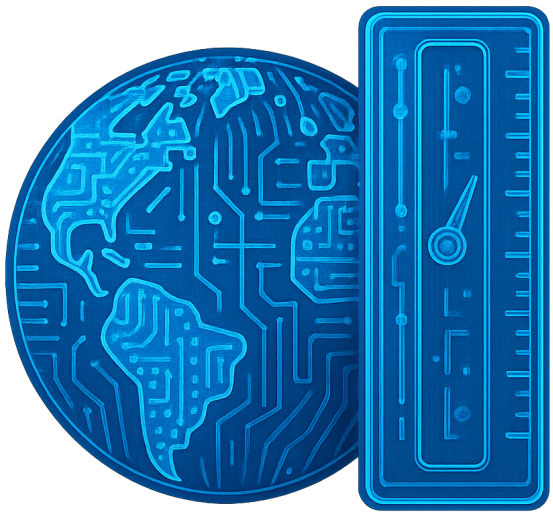
Abstract

Artificial intelligence is reshaping the real estate industry, transforming valuations, property management, and tenant screening. Adoption has been rapid, but regulatory capacity has lagged, leaving a fragmented landscape of Responsible AI frameworks—ethics principles, safety debates, and risk management guidelines—that educators must now translate into teaching. This chapter addresses that gap by introducing Responsible AI concepts tailored for real estate, showing how abstract principles can be grounded in applied, domain-specific contexts.

The chapter focuses on three foundations. First, **model design**: bias enters not only through data but also through choices of architecture, optimisation, and objective functions. Second, **sociotechnical systems mapping**: outcomes in housing markets are co-produced by human actors, machine systems, and institutional rules, making visible where accountability lies. Third, **market design**: AI systems can be deliberately structured to nudge or reshape behaviour, amplifying or mitigating inequalities in areas such as lending, pricing, and tenant selection.

Through real estate-specific and cross-sector case studies, the chapter illustrates how Responsible AI concepts operate in practice. Examples show both promise and peril: efficiency gains in valuations, but also risks of reinforcing structural bias; improved tenant screening, but with heightened privacy concerns.

The contribution is both pedagogical and applied. It offers classroom activities and an individual assignment that encourage critical engagement and contextual application. By equipping educators with tools to adapt to local markets and curricula, the chapter demonstrates how Responsible AI can move from principle to practice in a domain that touches almost everyone.



The World Values Benchmark (2022)

"All models are wrong, but some are useful."
G.E.P. Box (1979) [49]

CHAPTER 3: THE WORLD VALUES BENCHMARK

Building an AI evaluation methodology from a meta-ethic viewpoint.

Abstract

This chapter introduces the World Values Benchmark (WVB), the methodological core of the thesis. Whereas most benchmarks are normative, prescribing how models ought to behave, WVB is **descriptive**: it situates language model outputs within existing cross-cultural value distributions and makes divergences visible without adjudicating them. The framework fills a gap between performance-oriented leaderboards and broad sociotechnical critique by providing a reproducible method that captures pluralism while controlling for known artefacts of prompt sensitivity and anchor bias.

The methodology combines four elements: **prompt sets** to dampen paraphrase effects, **balanced answer anchors** to reduce framing skew, **Bayesian bias correction** to counter training priors, and **sociotechnical mapping** to keep validity tied to context. Together, these safeguards strengthen construct validity and make model behaviour empirically legible.

Applied to early models (LaMDA and PaLM), WVB revealed clear item-level alignment with US value profiles on culturally charged issues such as abortion and religiosity. Yet aggregate placement on the Inglehart–Welzel cultural map was closer to southern and central European societies such as Spain and Luxembourg. These findings show how descriptive evaluation can surface both the imprint of US training data and the ways those imprints shift under aggregation.

The contribution is twofold: empirically, WVB demonstrates that naïve single-prompt methods overstate alignment, while distributional profiles provide more stable placements; conceptually, it reframes benchmarking as relational measurement. The chapter establishes WVB as a tool for culturally inclusive, contestable evaluation—an approach that can inform more democratic decisions about model alignment and governance.

Semantic Auroras:

A letter to AI (2025)

"Something that is not dependently arisen,
Such a thing does not exist.
Therefore a nonempty thing
Does not exist."

Nāgārjuna, Mūlamadhyamakakārikā

2nd–3rd Century CE [14]

CHAPTER 4: SEMANTIC AURORAS

A Letter to Generative AI

Abstract

The question “Do machines think?” conceals more than it reveals. This chapter reframes the issue through the lens of *semantic auroras*—patterns of meaning that emerge when human intention meets machine architecture and cultural inheritance. These auroras explain why generative AI can appear conscious, even though no inner life is required to account for its behaviour.

The chapter situates large language models as sites of probabilistic convergence, where prompt, model, and culture interact to produce outputs that echo features of human thought without replicating its interiority. To understand this dynamic, I draw on quantum mechanics, where measurement does not uncover a pre-existing state but participates in bringing outcomes into being. Similarly, prompting collapses semantic potential into text, making language models less like static archives than fields of possibility.

This participatory account is developed through enactivist philosophy and extended via participatory realism: meaning is not passively retrieved but enacted through interaction across machine, society, and human. From this stance, prompting becomes a form of semantic navigation that reveals as much about our languages and cultures as about the models themselves.

The contribution is twofold. First, it synthesises philosophical and scientific perspectives—enactivism, sociotechnical theory, and quantum metaphors of indeterminacy—to explain why generative systems feel uncanny in their resemblance to human intelligence. Second, it extends the thesis’s central claim: evaluation is not neutral but world-making. By treating outputs as semantic auroras enacted through participatory realism, the chapter offers a language for critically engaging with the cultural and epistemic effects of generative AI.

Coda:

Measuring What We Enact

Ripples into form
Auroras enact wonder
What we measure shapes

RLJ, 2025

CODA: MEASURING WHAT WE ENACT

This thesis began with a simple claim and three hard questions. In generative AI, evaluation is not neutral; it participates in bringing forth what we later treat as given. So: *How* can we evaluate in ways that surface embedded norms? *What* does responsible evaluation mean in a pluralist world? And, *How* do we make the co-construction of values by models, people, and institutions empirically legible?

The Thread

Seen from a distance, this thesis is a meditation on what it means to *measure* in the age of generative AI. At stake is a philosophical shift: away from viewing evaluation as the neutral reporting of pre-existing capacities, toward understanding it as a practice that participates in shaping what those capacities appear to be. In this sense, the work belongs equally to the philosophy of AI and to measurement theory. Fields that converge on a simple but unsettling insight: whenever we ask a system what it is, we are partly making it so.

Chapter 1 mapped fractures in Responsible AI to deeper epistemological rifts and set enactivism as a bridge. That move reframes evaluation as observing becoming rather than measuring a fixed property. It also introduced MaSH Loops (Machine–Society–Human) to keep attention on recursive effects rather than isolated responses.

Chapter 2 preserved an early system state: value drift in GPT-3 (2021), where culturally charged prompts took on recognisable "accents." The point here is archival and methodological. It shows why descriptive, distributional read-outs matter. Later fine-tuning can erase the very imprints we most need to study. The chapter calls for instruments that can register such shifts rather than average them away.

Chapter 3 brought the argument into the applied world of real estate. Here proxies and metrics do not merely mirror markets; they shape them. Through sociotechnical mapping, feedbacks and power relations become visible to educators and practitioners, showing that evaluation is a form of governance, not an afterthought.

Chapter 4 supplied the methodological backbone: the World Values Benchmark (WVB) with Responsible Prompt Design (RPD) controls. By aligning model outputs to World Values Survey

constructs and explicitly managing anchors, paraphrases, normalisation, debiasing, and uncertainty, the method yields *value profiles* rather than *performance verdicts*. The chapter demonstrates that correcting prompt and anchor artefacts can materially change what a model appears to be. Instruments do not simply record findings; they shape them.

Chapter 5 stepped back to consider what such measurement means. Through participatory realism and the metaphor of semantic auroras, it argued that prompting is an intervention: it collapses potentials into outcomes. It also insisted that responsible evaluation acknowledges what it cannot claim.

Taken together, these chapters trace a single thread: that evaluation is constitutive, not neutral. This is both a philosophical and a practical claim. Philosophically, it widens the philosophy of AI by aligning enactivism with participatory realism and by positioning measurement as world-making. Practically, it develops concrete tools that embody this stance: archival, applied, methodological. To measure generative AI is to act within a loop that returns to shape both models and societies. Recognising this is the first step toward designing evaluations that are not only rigorous, but also responsible.

ABRIDGED

THIS CHAPTER INCLUDES HOW THE RESEARCH QUESTIONS WERE ANSWERED. IT ALSO LAYS THE GROUND WORK FOR SOME FUTURE RESEARCH. YOU CAN READ THE COMPLETE SECTION WHEN THE THESIS IS PUBLISHED.

Final words

This thesis began from a simple intuition: **evaluation is never neutral**. It is a kind of making. In pluralist and contested spaces, our measures do not merely record; they steer. As Ada Lovelace observed of the Analytical Engine, the Jacquard loom remains in modern AI. Yet its thread is not punched cards but human values: our instruments weave the patterns we later mistake for the fabric itself. The task is not to find a single canon, but to build measures that reveal rather than prescribe.

If there is one line I would leave with readers and examiners, it is the one that has guided the work throughout: **what we measure, we amplify**.