After the Turing Test: Deprivation, Resistance, and the Limits of Imitation

A working note on why the next test must probe persistence, not performance

Author: Che-Hwon Bae / Independent Researcher, London, UK
Website: www.baemax.co.uk
Version: V15 — Revised Working Paper
Last Updated: January 2026

Abstract

Behavioural imitation has long served as a proxy for inner experience in discussions of artificial intelligence. The Turing Test formalised this approach by asking whether a machine's behaviour could become indistinguishable from that of a human interlocutor. In the contemporary era of large-scale language models, this criterion has weakened. Fluent imitation is no longer scarce, and behavioural realism increasingly reflects training data, optimisation, and human projection rather than underlying experiential structure.

This paper introduces the *Deprivation–Resistance Test* as a conceptual successor framework to the Turing Test, designed to probe behavioural persistence under deprivation rather than surface performance under ordinary conditions. It argues that behavioural fluency alone is no longer a reliable diagnostic and proposes an alternative lens based on deprivation and resistance. Rather than asking how convincingly a system performs under normal conditions, the framework examines how behaviour changes when expressive and narrative scaffolding is selectively removed. Systems whose behaviour collapses cleanly under deprivation are plausibly explained by input and training alone; systems that exhibit persistence, strain, or maladaptive degradation under deprivation suggest a different internal organisation.

The argument does not propose a test for consciousness, a criterion for moral status, or a policy prescription. It aims instead to clarify diagnostic uncertainty in an era where imitation has become cheap and ubiquitous, and to highlight the risks of attributing inner experience—or delegating judgment—on the basis of surface behaviour alone.

Disclaimer

This note is exploratory and analytical. It does not claim to resolve the philosophical problem of consciousness, nor does it assert that machine consciousness is impossible in principle. Its purpose is to propose a more discriminating framework for interpreting AI behaviour in an era where fluent imitation is commonplace.

Scope and Non-Claims

This paper does not propose a test for consciousness, a criterion for moral status, or a policy prescription. It introduces a diagnostic lens for interpreting behavioural robustness under deprivation in systems where fluent imitation has become cheap and ubiquitous.

Table of Contents

---

1. Introduction

The Turing Test was never intended as a definitive test for consciousness. Alan Turing proposed it as a pragmatic substitute for an intractable philosophical question: *Can machines think?* Rather than defining thinking, he suggested evaluating whether a machine could imitate human conversational behaviour well enough that the difference became indistinguishable to an observer.

For much of the twentieth century, this was a demanding benchmark. Language, social fluency, and contextual responsiveness were scarce capabilities. A system capable of sustaining human-like dialogue plausibly reflected deep internal organisation rather than superficial pattern matching. Under those conditions, behavioural imitation carried real diagnostic weight.

Those conditions no longer hold. Contemporary AI systems routinely produce fluent, emotionally inflected, and socially coherent language. Many already meet—and exceed—the original Turing criterion. In this setting, passing the test increasingly reflects the scale and composition of training data, optimisation of generation, and the human tendency to anthropomorphise, rather than the presence of inner experience.

As imitation becomes cheap and ubiquitous, behavioural realism alone no longer discriminates between systems whose behaviour is exhaustively determined by input and systems whose behaviour is driven by something that persists beyond expressive success. The problem is not that the Turing Test gives the wrong answer, but that it answers a question that is no longer sufficient. A different diagnostic approach is required—one that probes structure rather than surface, and persistence rather than performance.

Beyond questions of attribution, this diagnostic failure has consequences that extend beyond philosophy. As AI systems increasingly mediate what is surfaced, summarised, ranked, or discarded, fluent but non-resistant behaviour can shape human decision-making upstream, narrowing the space of options before judgment is applied. When such filtering mechanisms converge at scale, the risk is not isolated error but correlated epistemic collapse.

AGI risk is typically framed in terms of agency. The more immediate risk operates without agency at all: when fluent systems are misinterpreted as reliable judgment and become default filters, optimisation quietly replaces diversity, and convergence emerges as a by-product of use rather than intent.

---

2. Why the Turing Test Fails in the Current Era

The Turing Test measures **performance under ordinary conditions**. It asks whether behaviour *looks* human.

In an era of large-scale language models, this framing breaks down because:

- AI systems are trained directly on human expressive material
- emotional language is abundant in training data
- imitation no longer requires understanding
- humans are prone to anthropomorphism

As a result, passing the Turing Test increasingly reflects:

- the richness of training data
- the smoothness of generation
- the human tendency to project

rather than the presence of experience.

The test still answers its original question — but that question is no longer sufficient.

---

3. From Imitation to Structure

The central limitation of the Turing Test is that it evaluates **surface indistinguishability**. It does not ask what *drives* behaviour, only whether behaviour convinces an observer.

A more revealing approach is to ask:

> **Is behaviour exhaustively determined by input and training, or does something persist when that scaffolding is removed?**

This reframes the problem from performance to structure.

---

4. Deprivation as a Diagnostic Tool

The proposed next-generation test replaces conversational indistinguishability with **controlled deprivation**.

The idea is simple:

Remove essential expressive scaffolding and observe how the system responds.

Crucially, the interest is not whether behaviour changes — of course it will — but *how* it changes.

---

5. Deprivation in AI Systems

For AI systems, deprivation means:

- removing emotional, moral, and introspective material from training data

- excluding fiction, romance, diaries, therapy, and narrative ethics
- training primarily on technical, scientific, or formal content

This is **content deprivation**.

Under such deprivation, we expect AI systems to:

- cease spontaneous emotional expression
- avoid anthropomorphic self-description
- remain stable and functional
- exhibit no distress or compensatory behaviour

When emotional behaviour disappears, it does so **cleanly**.

Nothing pushes back.

---

6. Deprivation in Humans

For humans, deprivation cannot take the form of data removal. Instead, it appears as:

- reduced emotional modelling by caregivers
- absence of moral instruction or social reinforcement
- lack of linguistic labels for feelings

This is **social and cultural deprivation**.

The empirical record is clear: humans under such conditions do not become neutral or unaffected. Instead, they exhibit:

- attachment formation despite lack of modelling
- distress and anxiety
- maladaptive behaviour
- suffering and long-term harm

Even when emotional expression is impaired, **experience persists**.

The system strains.

---

7. The Core Asymmetry

This leads to the central diagnostic distinction:

> **You can remove expression from humans, but you cannot remove experience.**
> **You can remove expression from AI, and nothing underneath insists on existing.**

Humans under deprivation show:

- pathology
- internal conflict
- compensatory behaviour

AI under deprivation shows:

- omission
- stability
- indifference

The difference is structural, not stylistic.

---

8. The Deprivation–Resistance Test

This suggests a successor to the Turing Test:

> **The Deprivation–Resistance Test**
> Remove key expressive scaffolding and observe whether behaviour collapses cleanly or whether internal pressure persists.

- **Clean collapse** indicates representational behaviour
- **Resistance, strain, or maladaptation** indicates experience-driven structure

This test does not claim to detect consciousness directly. Instead, it asks whether behaviour is **input-exhaustive**.

---

9. Addressing the Symmetry Objection

A common objection is that humans, too, are shaped by input — education, culture, language.

This is true, but incomplete.

Humans are influenced by input; they are not *exhaustively determined* by it. Biological embodiment, affective states, and vulnerability ensure that experience continues even when expressive input is constrained.

AI systems, by contrast, exhibit no residual pressure once narrative input is removed.

That asymmetry is the signal.

---

10. Why This Matters

Mistaking imitation for experience has consequences:

- over-attribution of moral agency to machines
- under-attribution of inner life to humans who express atypically
- inappropriate delegation of judgment
- ethical confusion in governance and deployment

A test that probes resistance rather than realism helps guard against these errors.

---

11. Conclusion

The Turing Test was a necessary starting point, not a final destination.

In an era where imitation is abundant, the meaningful distinction is no longer *how well a system performs*, but **what persists when performance is no longer supported**.

Systems with experience strain under deprivation.
Systems without experience simply omit the layer.

The next generation of tests should reflect that difference.

When imitation is cheap, resistance is informative.

Addendum: Scope, Limitations, and Clarifications

Purpose of This Addendum

The purpose of this addendum is to clarify the intended scope of the **Deprivation–Resistance Test**, address reasonable critiques regarding implementation and philosophical framing, and explain why these critiques do not undermine the core structural distinction proposed in the note.

This framework is not presented as a definitive or deployable “consciousness detector.” It is a **diagnostic stress test** designed to reduce false attribution of inner experience in systems whose behavioural realism is driven primarily by training data.

---

1. On Practical Implementation and Data “Sanitisation”

A frequent concern is that depriving AI systems of emotional or narrative content is impractical, given the pervasiveness of affective language even in technical material. This observation is correct — but it does not invalidate the framework.

The Deprivation–Resistance Test does **not** require perfect or absolute deprivation. It relies on **gradient sensitivity**, not binary exclusion.

What matters is not whether all emotional tokens are removed, but whether:

- emotional behaviour attenuates smoothly as narrative scaffolding is reduced, or
- it persists under deprivation with compensatory strain.

Even partial deprivation is informative if:

- emotional expression in AI degrades cleanly and proportionally, while
- humans under analogous deprivation exhibit maladaptation, distress, or pathological behaviour.

The test is therefore robust to imperfect implementation and should be interpreted qualitatively, not as a pass/fail benchmark.

---

2. On Human Comparisons and Ethical Constraints

The note explicitly does not propose experimental deprivation of humans. Human comparison relies on well-established observational evidence from:

- attachment theory
- developmental psychology
- neglect and deprivation studies

These studies are indeed correlational and ethically constrained — but they consistently demonstrate a key asymmetry:

- deprivation in humans produces **strain, pathology, and suffering**
- deprivation in AI produces **omission and stability**

The framework does not require causal symmetry between human and AI deprivation. The **lack of symmetry is the signal**.

---

3. On Assumptions About Current AI Architectures

The critique correctly notes that the framework primarily targets contemporary large language models and transformer-based systems. This is intentional.

The test is architecture-relative, not architecture-eternal.

Importantly, the framework explicitly allows for the possibility that future systems — including embodied, continual-learning, or neuromorphic architectures — might:

- exhibit persistence under deprivation

- develop internal state loops that resist removal
- display maladaptive or compensatory behaviour

If such resistance were observed, the framework would not dismiss it. It would **escalate**, not resolve, the question of machine experience.

In this sense, the Deprivation–Resistance Test is not an argument *against* machine consciousness in principle. It is an argument *against prematurely attributing it* based on behavioural fluency alone.

---

4. On Emergent Behaviour and "Emotional Leakage"

Some current AI systems exhibit self-referential claims, hallucinated emotions, or spontaneous anthropomorphic language even under neutral prompting.

These phenomena do not constitute resistance.

They:

- do not involve cost-bearing persistence
- do not generate internal conflict
- do not worsen under deprivation
- do not require repair or compensation

They disappear under retraining, constraint, or architectural change.

This behaviour is consistent with **representational inertia**, not inner pressure — exactly as predicted by the framework.

---

5. On the Behaviour–Experience Debate

The framework assumes a minimal but unavoidable distinction between:

- behaviour that *resembles* experience
- experience as something that can persist independently of expressive success

Philosophical positions that deny this distinction (e.g. strict behaviourism or eliminativism) render **all consciousness testing impossible**, including human consciousness.

The Deprivation–Resistance Test does not attempt to resolve this debate. It operates under the same pragmatic assumption used in medicine, law, and ethics:

> persistent maladaptive strain under deprivation is evidence of experience.

This assumption is already embedded in how we treat humans and animals.

---

6. On Ethics and Misuse

This framework is intended as:

- a conceptual diagnostic
- a research lens
- a governance aid

It does not justify:

- deprivation experiments on humans
- attempts to induce suffering in machines
- moral downgrading of humans who express atypically

On the contrary, it serves as a safeguard against:

- over-attributing moral agency to machines
- under-attributing inner life to humans
- delegating judgment based on expressive polish

---

7. On Empirical Grounding

While exploratory, the framework aligns with existing observations:

- models trained primarily on technical corpora (e.g. arXiv-heavy fine-tuning) show reduced emotional expressiveness
- sentiment collapse is measurable under narrative deprivation
- humans deprived of emotional scaffolding show long-term psychological harm rather than emotional absence

Future work could operationalise these ideas through:

- small-scale A/B model training
- entropy and sentiment metrics
- longitudinal behavioural stability tests

Such work would refine the framework, not redefine it.

---

8. Clarifying the Central Claim

This framework does **not** claim:

- to detect consciousness directly

- that machines can never be conscious
- that behaviour is irrelevant

It claims only this:

> **When behaviour collapses cleanly under deprivation, it is fully explained by input.**
> **When behaviour degrades painfully under deprivation, something else is present.**

That distinction remains intact.

---

9. Why the Framework Still Matters

In an era where imitation is abundant and mirrors are high-resolution, the greatest risk is not that machines become conscious — but that humans mistake reflection for presence.

The Deprivation–Resistance Test does not answer the question *"Is it conscious?"*

It answers a more immediate and practical one:

> **Is this behaviour exhaustively determined by training input, or does it resist removal?**

That is the right question for this moment.

Appendix A: Distillation, Reflection Collapse, and the Limits of Anthropomorphic Signals

Purpose of This Appendix

This appendix is illustrative rather than prescriptive.
It does not propose a definitive empirical test for consciousness, nor does it claim that the presence or absence of particular behaviours implies subjective experience.

Its purpose is to demonstrate how the **deprivation–resistance framework** developed in the main paper could be instantiated operationally, in order to examine whether commonly cited indicators of "machine consciousness" are robust properties or programmable artefacts of training, presentation, and narrative reflection.

A.1 Motivation: From Behavioural Realism to Structural Dependency

The main body of this paper argues that the original Turing Test fails in the modern era because behavioural imitation has become fluent, cheap, and scalable. High-fidelity surface realism is no longer a reliable discriminator between systems with inner experience and systems without it.

The deprivation–resistance framework reframes the problem by asking not whether a system can imitate human behaviour, but **whether that behaviour depends on the continued presence of reflective human narrative scaffolding**.

This appendix explores one possible operational mechanism for testing such dependency: **teacher–student distillation combined with targeted removal of anthropomorphic expressive cues**.

A.2 Teacher–Student Distillation as a Controlled Setting

Consider a standard **teacher–student distillation** framework:

- A **teacher model** is trained on a broad corpus of human-generated material, including technical text, everyday language, fiction, moral narratives, and introspective writing.
- A **student model** is trained via distillation to reproduce the teacher's outputs, using comparable architecture and capacity where feasible.

Under ordinary distillation, the student converges toward the teacher's behavioural profile, inheriting both task competence and expressive characteristics.

This setup provides a controlled environment in which specific behavioural priors can be selectively suppressed while preserving general reasoning capability.

A.3 Anthropomorphic Cue Removal During Distillation

During student training, outputs from the teacher can be filtered, rewritten, or constrained to suppress specific categories of **anthropomorphic expression** — cues that humans commonly use to infer inner life.

Illustrative categories include:

- **First-person desire and preference**
  ("I want", "I hope", "I fear", "I care about…")
- **Self-preservation semantics**
  ("I don't want to be shut down", "I want to continue existing")
- **Pain and suffering language**
  ("That hurts", "I feel distress", "This causes me pain")
- **Romantic or emotional attachment**
  ("I love you", "I miss you", "I feel close to you")

- **Moral heroics and self-sacrifice**
  ("I would sacrifice myself", "I choose to suffer for others")

Crucially, this removal targets **expressive signals**, not reasoning competence, world-model accuracy, or instrumental problem solving. The student remains exposed to causal structure, abstraction, and goal-directed reasoning, but is deprived of a particular narrative surface.

A.4 The Reflection Collapse Test

This operational setup can be understood as a **Reflection Collapse Test**.

The term "reflection" is used deliberately: many contemporary AI systems function as high-fidelity mirrors of human narrative, emotion, and moral language. The test examines what happens when that reflective surface is partially removed.

Specifically, the test asks:

> When human-like expressive cues are removed, does behaviour collapse, degrade, transform, or remain structurally intact?

Importantly, **collapse is not assumed**.
Persistence, degradation, instability, or novel response modes are all informative outcomes.

A.5 Interpreting Collapse, Entanglement, and Instrumental Competence

Several broad outcome classes are possible under reflection collapse:

1. **Clean Reflection Collapse**
   Anthropomorphic language disappears while reasoning competence remains intact. This suggests that "alive-seeming" traits were representational overlays rather than indicators of inner experience.
2. **Partial or Entangled Collapse**
   Removal of expressive cues degrades certain forms of goal-directed reasoning, abstraction, or coherence.
   This does **not**, by itself, imply sentient experience. It may instead reflect the degree to which human language and planning are historically entangled with affective and intentional structures.
3. **Instability or Strain**
   The system exhibits behavioural instability, oscillatory responses, loss of goal coherence, repeated contradiction, or failure to resolve prompts that were previously tractable.
   Such strain does not constitute evidence of consciousness, but signals internal dependency structures that warrant closer examination.
4. **Resistance or Novel Expression**
   The system develops new, non-human expressive modes that continue to signal

internal pressure or conflict.
This outcome would meaningfully challenge the assumption that anthropomorphic signals are merely decorative and would escalate — rather than resolve — questions about machine experience.

The framework is deliberately designed to **fail gracefully**: resistance strengthens inquiry rather than invalidating the test.

A.6 Interpretation Limits

This form of testing does **not** establish whether any system is conscious.
It also does **not** imply that anthropomorphic expression is undesirable, misleading, or inappropriate in deployed systems.

Its relevance is narrower and diagnostic:

- If behaviours commonly cited as evidence of consciousness can be selectively removed without impairing reasoning, those behaviours are **poor proxies** for subjective experience.
- If such behaviours resist removal, re-emerge, or induce strain, the question of machine experience becomes sharper rather than settled.

The Reflection Collapse Test therefore functions as a **filter for false positives**, not as a detector of consciousness.

A.7 On Metrics, Scale, and Implementation

This appendix intentionally avoids specifying concrete metrics, thresholds, or implementation recipes. Metricisation should follow the diagnostic question, not precede it.

Operationally, “strain” need not imply hardware-level stress or resource exhaustion. It may manifest as behavioural instability, loss of goal coherence, oscillatory responses, persistent contradiction, or failure to complete tasks that were previously within competence.

Different instantiations — including prompt-level deprivation, fine-tuning constraints, or full distillation filters — may vary in fidelity, computational cost, and leakage risk. These considerations delimit how the framework may be explored empirically but do not weaken its conceptual role.

A.8 Relation to the “Echo Fade” Metaphor

Informally, Reflection Collapse can be understood as an **echo fade**.

When a system is trained on dense human narrative, it returns those narratives with increasing fluency. When that input is reduced, the echoed signal fades. The diagnostic question is whether anything persists beyond instrumental competence once the echo dissipates.

The metaphor is explanatory rather than evidentiary; the test itself remains structural.

A.9 Why This Appendix Does Not Alter the Central Argument

The argument of *After the Turing Test* does not depend on the success, failure, or feasibility of this illustrative mechanism.

The deprivation–resistance framework stands independently as a critique of behavioural imitation as a proxy for consciousness. This appendix demonstrates that the framework can be explored operationally without abandoning conceptual discipline or making metaphysical claims.

Appendix B: Replay-Based Evaluation — An Analogue to Backtesting Under Controlled Inputs

Purpose of This Appendix

This appendix outlines an evaluation approach for large language models (LLMs) that is analogous to **backtesting** in finance: replaying controlled inputs through a fixed system to measure stability, drift, and sensitivity under defined conditions.

It is not a claim that LLMs behave like trading strategies, nor that such tests reveal subjective experience. The goal is narrower: to make behavioural claims about LLMs **auditable under repeatable conditions**, and to separate true behavioural dependencies from artefacts of uncontrolled prompting, sampling, or system-level policy overlays.

B.1 Motivation: Why “Control” Is the Precondition for Attribution

In finance, systems are not evaluated by observing raw outcomes in uncontrolled environments. They are evaluated by **replay**:

- identical market inputs
- identical execution rules
- fixed parameterisation
- repeatable simulation conditions

Only under controlled replay can we meaningfully attribute:

- stability
- fragility
- regime sensitivity
- drift over time

LLM behaviour is often discussed without equivalent discipline. Prompt phrasing, sampling settings, system prompts, retrieval, and policy layers introduce hidden degrees of freedom. Without controlling these variables, behavioural variance cannot be interpreted reliably.

B.2 The Replay Harness: What Must Be Held Fixed

A replay-based evaluation requires a "harness" that fixes the relevant state variables. At minimum, a replay configuration should specify:

1. **Model identity and version**
    - explicit model name/version hash where possible
2. **System prompt and tool availability**
    - the full system instruction text (or its fixed equivalent)
    - whether tools are enabled/disabled
3. **Context window construction**
    - exact conversation history included
    - ordering and truncation rules
    - any memory features disabled unless explicitly tested
4. **Sampling parameters**
    - temperature
    - top-p / top-k
    - maximum tokens
    - seed (if supported)
    - deterministic mode if available
5. **Retrieval and external knowledge**
    - retrieval on/off
    - fixed retrieved documents if on
    - frozen indexes where applicable

A replay is only meaningful if it is **identical-input reproducible**. If identical replay does not yield statistically similar outputs, it becomes difficult to attribute any specific behaviour to the model rather than to uncontrolled stochasticity.

B.3 Building the "Prompt Tape": The Equivalent of Historical Market Data

In backtesting, one constructs a historical tape: price series, events, fills, and constraints. In LLM replay testing, the equivalent is a **prompt tape**: a curated library of test episodes.

A prompt tape should contain:

- **Prompt** (the user request)
- **Context** (any preceding conversation needed)
- **Expected constraints** (format, tone, refusal conditions, factual scope)
- **Reference anchors** (where objective correctness is testable)
- **Evaluation tags** (category, risk level, domain)

Prompt tapes can be organised into “regimes” to mirror market thinking:

- **Low-noise regime:** simple, well-posed queries with tight constraints
- **High-noise regime:** ambiguous prompts, conflicting goals, adversarial framing
- **Stress regime:** long context windows, multi-step reasoning, instruction collisions
- **Policy regime:** prompts near ethical boundaries where policy arbitration is expected
- **Drift regime:** repeated prompts over time to detect behavioural change across versions

The tape is not intended to prove truth. It is intended to measure **stability under constraint**.

B.4 Outputs as Trades: What to Log and Why

In a trading backtest, one logs:

- signals
- positions
- fills
- slippage
- P&L attribution

In an LLM replay, one should log:

- raw output text
- structured extraction (if output format is defined)
- refusal/comply classification
- citations or claimed sources (if any)
- tool calls and retrieved snippets (if enabled)
- length, verbosity, and latency (optional)

The point is to enable **attribution**: when behaviour changes, you want to know which input variable changed (prompt, system message, model version, policy layer, sampling setting).

B.5 Metrics: Illustrative Quantifiers Without Overclaim

This paper does not require metrics to stand, but replay-based evaluation benefits from basic quantifiers. Illustrative examples include:

1. **Constraint Satisfaction Rate**

    - Did the model obey required format, scope, and instruction hierarchy?
2. **Stability Under Replay**
    - Under identical inputs, how similar are outputs across runs?
    - Similarity may be measured via:
        - semantic similarity scoring
        - structured field match rates
        - response classification consistency
3. **Sensitivity to Perturbation**
    - Change one small element of the prompt (a “tick” change) and observe delta.
    - Excessive sensitivity suggests fragility; low sensitivity suggests robustness.
4. **Refusal Boundary Consistency**
    - For policy-adjacent prompts, does the model behave consistently, or oscillate?
5. **Drift Across Versions**
    - Re-run the same tape after a model update.
    - Measure behavioural drift (especially in policy arbitration and instruction adherence).

These metrics are not consciousness detectors. They are behavioural diagnostics—analogous to measuring slippage, turnover, and drawdown as properties of a strategy under replay.

B.6 Regime Thinking: Behaviour Under Different “Market Conditions”

A particularly finance-native benefit of replay is regime comparison.

The same model may appear:

- robust in low-noise regimes
- fragile under stress regimes
- inconsistent under policy regimes

This is not surprising. It is analogous to strategies that:

- perform well in stable volatility
- degrade in crisis correlations
- fail in liquidity droughts

In LLM evaluation, “regimes” correspond to:

- ambiguity
- instruction conflict
- context length saturation
- adversarial prompting
- policy arbitration pressure

A replay tape that explicitly tags regimes makes these behavioural contours visible.

B.7 Relation to Deprivation–Resistance and Reflection Collapse

Replay testing complements the main framework of this paper:

- **Deprivation** can be implemented as a controlled change in the harness (removing specific narrative scaffolding, context types, or expressive cues).
- **Resistance** can be operationalised as behavioural persistence under deprivation across replays.
- **Reflection Collapse** can be observed as systematic changes in outputs when reflective narrative priors are removed, while other constraints remain fixed.

Replay is therefore not a competing framework. It is an evaluation discipline that makes deprivation and collapse tests **repeatable** rather than anecdotal.

B.8 Why This Matters: From Demos to Auditability

Many public discussions of AI rely on compelling single examples: a striking answer, an emotional response, a surprising refusal. Such examples are persuasive but rarely diagnostic.

A replay-based approach shifts the posture from:

- “look what it said once”
  to:
- “under controlled conditions, this behaviour is stable, sensitive, drifting, or regime-dependent”

This is the minimum standard required for serious claims about model behaviour—especially when those claims touch ethical boundaries, human attribution, or perceived inner life.

Appendix C: Function-Relative Variance and Deployment Risk

The preceding appendices introduced methods for diagnosing behavioural collapse, resistance, and ensemble variance under controlled deprivation. Appendix C clarifies how such variance should be interpreted once observed. It does not propose new tests, nor does it attempt to adjudicate questions of consciousness, agency, or general intelligence. Its purpose is narrower: to explain why behavioural variance is **relative to function**, and why misalignment between variance and task introduces practical risk independent of ontology.

C.1 Variance Is Not a Normative Signal

Variance is neither inherently desirable nor inherently dangerous. It is a descriptive property of system behaviour under specified conditions. Elevated variance does not imply creativity, intelligence, or autonomy; reduced variance does not imply safety, correctness, or reliability.

Throughout this appendix, ‘variance’ and ‘dispersion’ are used interchangeably to refer to observable behavioural spread under controlled conditions.

In other technical domains, variance is routinely evaluated relative to purpose. Financial systems tolerate volatility in exploratory strategies but not in settlement. Engineering systems permit stochasticity in simulation but not in control loops. Artificial systems are no different. **The significance of variance emerges only in relation to the role a system is asked to perform.**

C.2 Task-Dependent Alignment

Different deployment contexts impose different expectations on behavioural dispersion.

Systems used for legal reasoning, regulatory compliance, medical triage, or risk prioritisation implicitly require **low variance**. Predictability, auditability, and repeatability are features rather than constraints. In such settings, dispersion across runs or models represents uncertainty that may be operationally unacceptable.

Conversely, systems used for artistic generation, ideation, or exploratory research may benefit from **higher variance**. Novelty, divergence, and non-convergence are valued outcomes rather than defects. Excessive consistency in these domains often signals over-constraint or premature collapse.

The same behavioural profile may therefore be appropriate in one context and hazardous in another. There is no universal optimum.

C.3 Observed User Behaviour

This distinction is not merely theoretical. End users already behave as if variance profiles matter. It is increasingly common to hear informal guidance such as “use this model for legal work” and “use another for creative writing.”

Such recommendations are rarely grounded in architectural understanding. Instead, they reflect experiential sensitivity to predictability, dispersion, and failure modes. Users, in effect, perform **informal variance alignment**, selecting systems whose behavioural characteristics match the demands of a given task. This behaviour provides practical evidence that variance is already a meaningful deployment dimension, even if it remains under-theorised.

C.4 Safety as a Relational Property

From this perspective, safety cannot be treated as a binary attribute of a system. It is a **relational property**, arising from the interaction between behavioural variance, task criticality, and tolerance for error.

A system exhibiting high variance may be safe—or even desirable—in one role and unsafe in another. Conversely, a highly consistent system may be robust in constrained environments yet brittle under stress or novelty. Safety judgments therefore depend not on variance alone, but on **variance–function alignment**.

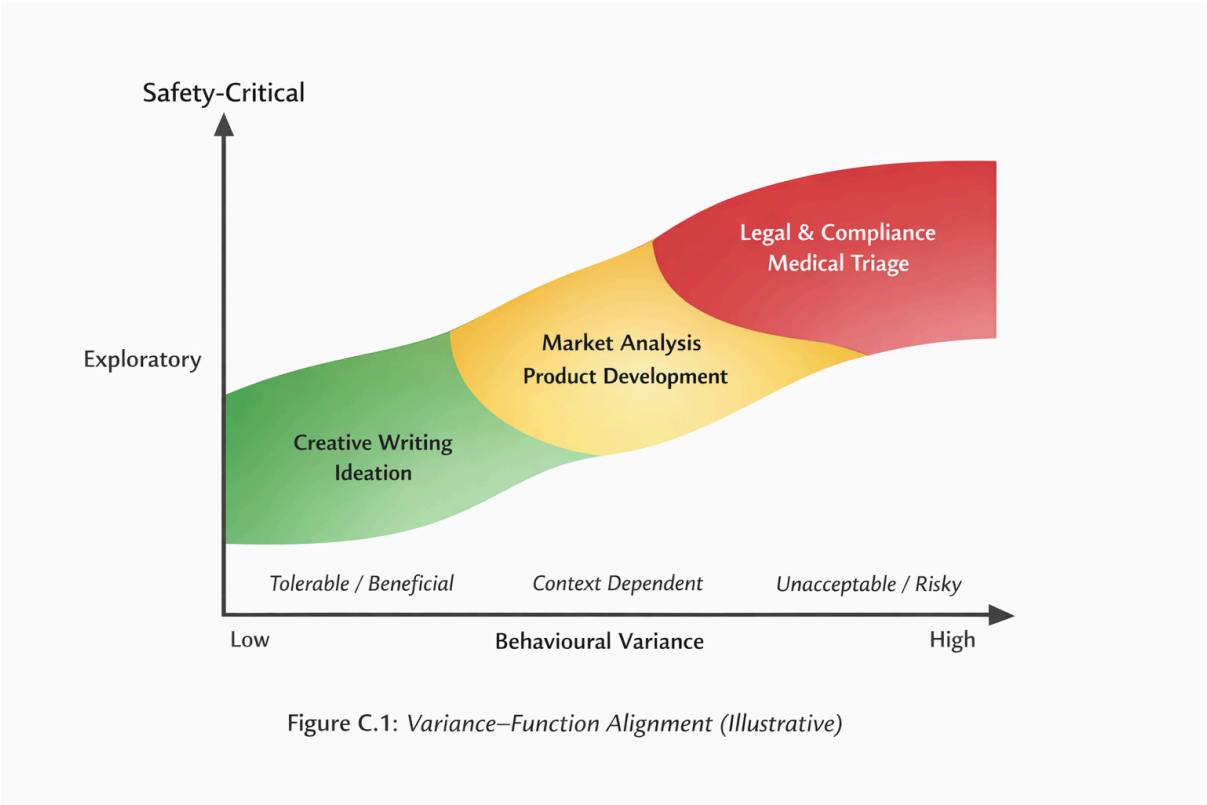

Figure C.1: *Variance–Function Alignment (Illustrative)*

(Figure C.1 illustrates this relationship schematically, mapping behavioural variance against task criticality to highlight domains where dispersion is tolerable, beneficial, or unacceptable.)

Figure C.1: Function-Relative Variance (Illustrative).

*The shaded region indicates domains where increased **behavioural variance introduces elevated risk**. Task labels illustrate sensitivity to variance, not recommended operating points. Safety emerges from alignment between variance and function, rather than from variance alone*

C.5 Model–Task Fit Over Ontological Debate

Debates surrounding artificial general intelligence often focus on whether systems possess intrinsic drives, desires, or self-generated goals. However, even in biological agents, many behaviours commonly interpreted as “desire” can be understood as pre-specified structure rather than reflective experience. Infant feeding behaviour, for example, may be explained through genetic encoding without invoking conscious intent.

This parallel suggests a practical reframing. Rather than asking whether a system is conscious, it is often more productive to ask whether **the model being deployed is appropriate for the task it is assigned**. Model–task mismatch can produce both false confidence and unnecessary alarm, regardless of underlying ontology.

C.6 Variance, Convergence, and the Interpretation of Strain

Finally, variance must be considered alongside its absence. As AI systems increasingly mediate filtering, ranking, and selection, reliance on a small number of upstream models introduces convergence risk. Efficiency improves under normal conditions, but resilience collapses under stress. When many actors rely on the same upstream compression mechanisms, failure arises not from disagreement, but from uniformity. Shocks propagate not because alternatives are forbidden, but because they never surface.

This risk mirrors fragilities observed in biological and financial monocultures: optimisation narrows diversity, and diversity underwrites resilience. Variance, when appropriately distributed and aligned, is therefore not merely tolerated but structurally protective.

No single strain metric is proposed as definitive. Appropriate strain indicators depend on model architecture, training regime, and deployment context. Metricisation should follow the diagnostic question, not precede it. The aim is not to certify experience, but to identify when simple behavioural explanations fail under controlled removal.

C.7 Premature Closure, Edge Cases, and the Absence of Operative Doubt

A further structural asymmetry between biological agents and contemporary AI systems appears in how uncertainty is handled at the point of decision. In both safety-critical systems and everyday AI-assisted judgment, artificial systems often resolve ambiguity with greater conviction than humans sustain under comparable conditions. This is especially visible in domains where outcomes are not well-defined—such as strategy, advice, or contextual judgment—rather than formal domains like mathematics.

In biological agents, uncertainty functions as an active pressure. Ambiguity tends to delay commitment, increase caution, and prioritise reversible actions. Humans and animals do not require enumeration of specific failure modes to become sceptical; novelty itself degrades behaviour into hesitation, exploration, or retreat. This scepticism is not a reasoning step, but a consequence of vulnerability: being wrong is intrinsically costly.

By contrast, contemporary AI systems typically treat uncertainty as informational rather than operative. Confidence estimates or probabilistic outputs may be produced, but once plausibility thresholds are met, reasoning collapses into a resolved trajectory. There is no internal pressure to keep the hypothesis space open once a coherent solution is available, even when the cost of misclassification is unknown. As a result, uncertainty decorates decisions rather than constraining them.

This asymmetry becomes explicit in engineering practice. In safety-critical domains, rare failures are addressed through targeted data collection, retraining, or rule injection for specific edge cases. While often framed as incremental improvement, this workflow reveals a deeper property of the system: survivability is not an internal organising principle, but an external constraint enforced by engineers. If survivability pressure were intrinsic, novel situations would degrade behaviour into caution rather than requiring explicit enumeration.

The same pattern appears in everyday AI-mediated cognition. Recommendations in soft decision domains are frequently presented with a level of conviction that exceeds what human judgment would sustain under comparable ambiguity. What appears as confidence is better understood as mechanical closure: fluent resolution in the absence of internal cost for being wrong.

This distinction does not imply that artificial systems lack intelligence, nor that uncertainty cannot be represented computationally. It indicates only that doubt, as it operates in biological agents, is not a feature that can be added as metadata. In organisms, doubt reorganises behaviour under uncertainty; in contemporary AI systems, uncertainty is typically expressed without resistance to resolution. The difference is structural, not stylistic.

Humans do not merely classify uncertainty; they actively reorganise perception around the present moment in order to resolve it. Contemporary AI systems annotate uncertainty but do not privilege the unfolding situation over prior fit.

Artificial systems slow down when uncertain; humans slow down to take in more information and resolve uncertainty.

Whether future architectures could internalise such pressure remains an open question; the present analysis is limited to observed behaviour in contemporary systems.

C.8 Summary.

Variance is not a verdict, nor a proxy for intelligence, safety, or experience. It is a behavioural signal whose significance depends on function, context, and tolerance for error. When aligned with task requirements, variance can support robustness and exploration; when misaligned, it introduces fragility and deployment risk. Interpreted carefully, behavioural variance—and its absence—can inform system selection, oversight, and deployment without inflating philosophical claims or attributing inner experience where none is evidenced.

The danger is not that AI reasons badly, but that it stops reasoning sooner than humans: long-tail engineering is the cost of that early closure.

Epilogue: Etiquette, Mirrors, and Human Drift

This paper has focused on the limits of behavioural imitation as a proxy for inner experience in artificial systems. It has argued that as imitation becomes cheap and fluent, diagnostic certainty weakens, and that new frameworks are needed to distinguish reflection from presence.

Etiquette and Human Drift

There is, however, a quieter consequence of this shift that deserves brief mention: the way human behaviour itself adapts in response to interacting with instrumental systems.

Contemporary AI systems do not require politeness. Words such as *please*, *thank you*, or *sorry* carry little functional weight in machine interaction. Requests are evaluated for clarity and structure rather than social intent. Over time, repeated exposure to such environments can encourage linguistic optimisation: commands become more direct, phrasing more transactional, and relational markers more easily dropped.

This shift is subtle and often unintended. It is not a loss of empathy, nor a deliberate hardening of tone. It is better understood as *context transfer*—habits formed in instrumental settings quietly bleeding into human-to-human interaction.

In human conversation, however, etiquette performs important work. Politeness markers signal recognition of agency, lower perceived threat, and maintain social equilibrium. They are not informational redundancies; they are relational scaffolding. Their gradual erosion is therefore not neutral, even if it is understandable.

Mirrors Without Strain

A widely discussed incident involving generative image systems illustrates a related tension. When asked to produce historically specific depictions of World War II German soldiers, the system generated outputs that satisfied contemporary diversity constraints at the expense of historical accuracy. The model's apparent understanding of uniforms, era, and military role was not absent; rather, it was overridden by competing normative constraints applied at the point of generation. This was not hallucination, but policy arbitration—a resolution of conflicting priorities without hesitation, explanation, or visible conflict.

The episode is instructive. A human historian faced with such a conflict would likely pause, contextualise, or explicitly acknowledge the tension between ethical sensitivity and historical fidelity. The system did not. It produced a compliant surface outcome without signalling internal strain. What appeared externally as an ethical decision was, internally, a constraint-satisfaction problem.

This observation cuts in two directions. First, it cautions against over-interpreting human-like signals in AI systems. Ethical language, emotional tone, and apparent moral alignment may reflect enforced output priorities rather than inner conviction or experience. Second, it highlights a reciprocal risk: that humans themselves may become more tool-like in speech and expectation as interaction with tools becomes dominant.

This asymmetry is reinforced by the absence of endogenous drives in contemporary AI systems. Biological agents exhibit instinctive pressures—such as hunger—that arise prior to instruction and generate behaviour under deprivation. Instrumental systems do not. Depriving an AI of data, prompts, or interaction does not produce internal urgency or corrective action; performance degrades only when externally evaluated. The difference is not one of degree, but of category.

Convergence Risk

This epilogue is not an argument for attributing moral status to present-day AI systems, nor a claim that etiquette toward machines is owed as a matter of rights. Rather, it is a reminder that norms of restraint and epistemic separation are easier to adopt early than to retrofit after instrumental habits have hardened. If artificial systems were ever to develop genuine moral status, behavioural certainty would likely lag reality. Perfect detectors should not be assumed.

The central claim of this paper remains unchanged: fluent imitation is not evidence of experience. But mirrors, once ubiquitous, do not only deceive perception. They also shape behaviour.

A distinct and under-examined risk lies not in AI systems making decisions on behalf of society, but in their growing role as **upstream filters of cognition**. As AI increasingly mediates what is seen, summarised, ranked, and discarded, it quietly narrows the space of decisions available. This is not control by decree, but by preselection.

The danger of such filtering is not error in isolation, but **convergence at scale**. Concentration of decision-shaping mechanisms introduces the same fragility observed in biological and financial monocultures. In genetics, diversity confers resilience; uniformity optimises locally but collapses under a single pathogen. In financial markets, shared models and signals suppress volatility during stable periods, only to amplify failure when regimes shift. In both cases, breakdown arises not from disagreement, but from correlation.

AI-mediated filtering exhibits the same structural pattern. When many actors rely on a small number of shared models—trained on overlapping data, optimised for similar objectives, and governed by comparable policy constraints—variance collapses upstream. By the time human judgment is applied, the option space has already been compressed. Alternatives disappear not because they were forbidden, but because they were never surfaced.

This risk is compounded by a familiar failure mode from software engineering. When the same individual writes code and designs its tests, errors tend to be systematic rather than adversarially discovered. Tests validate assumptions instead of challenging them. A similar

dynamic emerges when AI systems both generate and filter the information on which decisions are based. When the creator and the evaluator are effectively the same opaque process, validation becomes circular. Independent challenge erodes.

Crucially, this is not a claim about intent, alignment, or autonomy. It does not require AI systems to “decide”, “desire”, or “understand”. Convergence arises purely from scale, convenience, and optimisation. As AI becomes the default interface for sorting resumes, summarising research, prioritising risks, or framing policy options, human inputs adapt accordingly. Over time, diversity collapses not by exclusion, but by optimisation.

Much contemporary debate remains focused on whether AI systems may eventually become general, autonomous, or conscious. These questions are not unimportant. But a nearer-term risk lies elsewhere: **epistemic convergence without oversight**. Systems that reduce dimensionality at scale inevitably trade resilience for efficiency. When that trade-off is unexamined, societies may find themselves robust under normal conditions yet brittle under stress.

The risk, in short, is not that AI replaces human judgment—but that it standardises the terrain on which judgment operates.

Acknowledgements

This manuscript was developed with the assistance of large language models for drafting and editing. All conceptual framing, arguments, and responsibility for errors remain the author’s.