# Auditing Semantic Footprints in Artefactual Systems

A protocol for domain-bounded semiotic surrogacy

Author: Israel Huerta Castillo

Affiliation: University of Santiago de Compostela (USC)

Email: israel.huerta@rai.usc.es

ORCID: 0009-0004-8615-8008

Version: Author’s preprint

Year: 2025

DOI: 10.5281/zenodo.18155331

# Auditing Semantic Footprints in Artefactual Systems

A protocol for domain-bounded semiotic surrogacy

**Abstract.** Can an artefactual system preserve the operative semantics of a person without claiming to be that person? This paper presents an audit protocol for semantic footprints: domain-bounded, intervention-grounded traces of stable sign–object–interpretant commitments in an artefactual system. The protocol operationalises semantic surrogacy as a falsifiable competence claim, rather than a metaphysical identity thesis, and provides a discipline for avoiding common partition artefacts (context leakage, cue-based overfitting, and stylometric ‘ghost’ persistence). Building on a companion framework that defines a Semiotic Adequacy Test (SAT) for minimal agency in bio–artefactual continuity, we treat SAT’s criteria C1–C4 as inherited infrastructure and contribute four audit-facing components: (i) a formal negative criterion (Π*) for non-generating partitions, (ii) an admissibility firewall for estimators of semiotic density ρ and context plasticity κ, (iii) a minimal intervention suite do(I) suitable for replication, and (iv) a catalogue of failure modes with diagnostics suitable for pre-registration and third-party review. The result is a practical, intervention-auditable methodology for evaluating semantic surrogacy claims in artefactual systems without collapsing them into avatars/proxies, stylometry, or human-likeness scores.

**Keywords:** semantic footprint; semiotic surrogacy; audit protocol; intervention suite; contextuality; non-generating partitions; interpretive phase transitions; AI avatars; bio–artefactual continuity

## 1. Introduction

“Singularity” discourse in artificial intelligence has become a conceptual sink: a single term is routinely asked to carry incompatible burdens—runaway computational scaling, sudden agency, epistemic discontinuity, and, in its most culturally persistent form, a promise of personal survival by technical means (Abdelkarim, 2025; Ishizaki and Sugiyama, 2025). These conflations are not merely rhetorical; they generate predictable category errors in both philosophical analysis and empirical evaluation. In particular, they encourage two invalid inferences: first, that increasing complexity or capability automatically warrants attributions of agency; and second, that a high-fidelity surrogate of a person’s discourse or decision style constitutes persistence of that person. Recent debates at the intersection of philosophy of mind, philosophy of AI, and computational ethics already display how quickly the boundary between representation and personhood can be destabilised when technological proxies become socially consequential (Sweeney, 2023, 2025; Patrone, 2025). A defensible methodology must therefore begin with a stricter object of inquiry and a stricter discipline of evidence.

This paper proposes a deliberately modest, but operationally sharp, replacement for “semiotic singularity” claims. The central idea is that the only defensible “singularity-like” phenomenon in this vicinity is not an ontological rupture but an interpretive phase transition (IPT): a measurable regime shift in the dynamics of interpretation in an artefactual system trained—biomimetically—under minimal semiotic constraints. The post-transition system yields a domain-bounded semiotic surrogate (DSS): not a surviving person, and not a bearer of phenomenological continuity, but an intervention-auditable surrogate that preserves a person-indexed semantic footprint, understood as operational semantic dispositions rather than identity. This shift is methodological rather than metaphysical: it replaces “Who survives?” with “What, exactly, persists—and how can we audit it under explicit interventions?” (Pearl, 2009; Woodward, 2004).

Three demarcations structure the proposal.

First, the object of persistence is semantic-operational, not personal. Work on “digital afterlife” industries has shown how easily post-mortem representation slides into moral, political, and economic confusion when identity-laden language is left unconstrained (Öhman & Floridi, 2017). Our framework therefore treats any afterlife or identity framing as out of scope by construction: a DSS is evaluated only as a tool-like, domain-bounded surrogate for a semantic footprint. This choice is compatible with, but not reducible to, current debates on avatars and representation: proxy frameworks highlight responsibility gaps and epistemic gaps between persons and their avatar representatives (Sweeney, 2023), while hybrid-person frameworks reconceptualise avatars as temporal parts of persons (Patrone, 2025). The present paper brackets these metaphysical options and instead contributes a test discipline for when a system is evidentially entitled to be treated as a footprint-indexed surrogate at all.

Second, auditability is intervention-defined. The epistemic core of the proposal is not behavioural resemblance but counterfactual sensitivity under explicit interventions, formalised using do-operators (Pearl, 2009). Intuitively: if a system genuinely preserves a semantic footprint, then controlled changes to context constraints, task framing, and memory affordances should produce predictable, directionally stable changes in outputs and internal state transitions. If, by contrast, the system is only a surface mimic, its apparent fidelity will typically be brittle under such interventions. This interventionist stance aligns with an increasing emphasis, across philosophy of AI and responsible ML, on operationalising ethical and governance concerns through auditing rather than aspiration (Hagendorff, 2020; Mökander et al., 2022; Mökander et al., 2024). Our point, however, is epistemic before it is regulatory: do(I)-sensitivity is the criterion that turns footprint claims into testable hypotheses.

Third, contextuality is treated as a constraint on inference, not as microphysics. Any framework that evaluates interpretive competence across heterogeneous contexts faces a structural risk: illicitly aggregating outputs as if there were a single global state that can be read off independently of the measurement context. We therefore adopt a “quantum-compatible” contextuality strategy strictly as an epistemic discipline: what can be inferred about a system depends on how it is probed, and incompatible measurement contexts must not be collapsed into a single unqualified description (Abramsky & Brandenburger, 2011). Nothing in this move requires quantum mechanisms in minds or machines. The payoff is methodological: contextuality becomes a guardrail against overclaiming, especially in settings where evaluation partitions, prompting regimes, and tool access can silently change the very object being measured.

With these demarcations in place, the paper introduces a compact technical vocabulary. We define two parameters: ρ (relational semiotic density), intended to capture how richly and consistently a system sustains many-to-many sign–object–interpretant relations across contexts; and κ (context reconfiguration rate), intended to track the speed and structure with which a system revises its operative context state $\mathcal{C}$ under novelty and intervention. The signature of an IPT is then stated as a dynamical criterion: a sustained acceleration of κ driven by endogenous reconfiguration of $\mathcal{C}$, with preserved sensitivity to do(I) and without collapse of ρ under robustness-preserving evaluation partitions. This is meant to be falsifiable: if κ does not accelerate in the relevant way, if do(I)-sensitivity degrades, or if measured density is revealed as an artefact of partition choice, then no IPT is declared.

A central risk in footprint and surrogacy claims is what we call artefact-generation: the evaluation protocol itself can create the appearance of persistence. This is why the paper treats non-generating partitions as a non-negotiable methodological constraint. In autobiographically dense data, naive train/test splits can leak identity keys, reuse prompt templates, or encode labels in partition rules, inflating apparent fidelity while suppressing genuine context reconfiguration. Our robustness requirement is therefore negative as well as positive: it specifies what kinds of partitions invalidate a result and why.

This paper inherits the Semiotic Adequacy Test (SAT) and its admissibility constraints (C1–C4) from Huerta Castillo (2025) and treats them as a gate on evidential entitlement rather than as a contribution re-derived here. The present contribution is the audit protocol that operationalises those constraints for semantic footprints and domain-bounded semantic surrogacy.

Section 2 introduces the minimal-semiosis and contextuality constraints that define the evaluation space, and formalises SAT and C1–C4. Section 3 operationalises ρ and κ and specifies non-generating partitions. Section 4 presents the intervention suite and the IPT detection criterion as a pre-registrable protocol. Section 5 situates the framework with respect to debates on avatars, proxies, parts, and responsibility (Sweeney, 2023, 2025; Patrone, 2025), and clarifies its relation to broader concerns about post-mortem representation (Öhman & Floridi, 2017) without adopting identity or afterlife framings. Section 6 states limits, risks, and governance-adjacent implications, aligning the proposal with auditability-oriented approaches (Hagendorff, 2020; Mökander et al., 2022; Mökander et al., 2024). Section 7 concludes by recommending that "semiotic singularity" remain, at most, a narrative label for the interpretive phenomenon that IPT precisely defines.

## 2. Operational Framework: Minimal Semiosis, Contextuality, and Admissibility

### 2.1. Scope and non-claims

This paper is explicitly *not* a theory of personal survival, identity persistence, or phenomenological continuity. The object of inquiry is semantic-operational persistence: whether an artefactual system can preserve and reproduce, within a bounded domain, a person-indexed semantic footprint in a manner that is auditably sensitive to intervention. The framework therefore treats "afterlife" interpretations and identity-laden conclusions as out of scope by construction, even when the outputs of a system are socially received as person-like representations. This bracketing is motivated by the well-documented normative and conceptual instability of post-mortem representation once identity vocabulary is left unconstrained (Öhman and Floridi 2018).

Two further constraints discipline what follows. First, the proposal is pragmatist–enactive: what counts as "meaningful" behaviour is tied to patterns of constraint, use, and consequence under controlled perturbation, not to superficial resemblance or retrospective narrative interpretation (Woodward 2004). Second, the use of "quantum-compatible" contextuality is strictly epistemic: it functions as a guardrail against illicit aggregation across incompatible measurement contexts and does not commit the paper to microphysical quantum mechanisms in minds or machines (Abramsky and Brandenburger 2011).

### 2.2 Core constructs

Semantic footprint (SF). A *semantic footprint* is the person-indexed profile of stable, context-sensitive operational semantic dispositions—regularities in how a subject interprets, re-frames, and justifies sign-use under constraints—recoverable only through intervention-auditable tests rather than by stylistic resemblance alone.

Domain-bounded semiotic surrogate (DSS). A *DSS* is an artefactual system that, within a specified domain and under explicit admissibility constraints, supports reproducible inferences about a target semantic footprint by exhibiting stable do$(I)$-sensitivity and context reconfiguration patterns consistent with that footprint, without implying personhood or experiential continuity.

Semiotic training (biomimesis). *Semiotic training (biomimesis)* is any training regime that shapes an artefactual system under minimal semiotic constraints—forcing object-sensitive sign-use, context dependence, and intervention-responsiveness—so that the learned behaviour is not reducible to surface imitation under fixed partitions.

Interpretive phase transition (IPT). An *IPT* is a sustained regime shift in interpretive dynamics in which the system exhibits accelerated, structured reconfiguration of its operative context state $\mathcal{C}$(captured by $\kappa$) while preserving intervention-sensitivity and relational coherence of sign–object–interpretant mappings (captured by $\rho$) under non-generating partitions.

These definitions are intentionally conservative: they do not require attributing agency, consciousness, or moral status. They instead specify a testable target—semantic-operational persistence—and the minimal evidential discipline required to evaluate it.

### 2.3 Notation and measurement stance: $\boldsymbol{do(I)}$, $\boldsymbol{\mathcal{C}}$, $\boldsymbol{\rho}$, and $\boldsymbol{\kappa}$

We treat evaluation as an interventionist enterprise: claims about preserved semantic dispositions must be supported by counterfactual sensitivity under explicit interventions. Formally, $do(I = i)$denotes an exogenous intervention that sets an evaluatively salient variable (or bundle of variables) $I$to a value $i$, breaking ordinary dependence relations and enabling causal or counterfactual inference from controlled contrasts (Pearl 2009; Woodward 2004). Here, $I$ranges over three families: context constraints, task framing, and memory affordances (formalised in Section 4 as an intervention suite).

The system's operative *context state* is denoted $C$. This is not a metaphysical "inner world model," but the minimal state descriptor needed to track how the system configures and reconfigures contextual constraints across interactions (e.g., latent control states, tool-use regimes, prompt policies, retrieval gates, or other measurable control variables—implementation-specific but operationally characterisable).

We introduce two measurement-facing parameters:

- Relational semiotic density $\rho$. $\rho$is a parameter intended to capture the richness and stability of many-to-many sign–object–interpretant relations across admissible contexts in the target domain. Intuitively, a higher $\rho$indicates that the system can sustain coherent interpretive linkages that generalise across reframings without collapsing into brittle pattern-matching.
- Context reconfiguration rate $\kappa$. $\kappa$tracks the speed and structure with which the system revises $C$in response to novelty and $do(I)$-interventions. The point of $\kappa$is not "fast adaptation" as such, but *structured* reconfiguration that is reproducible and footprint-indexed rather than artefact-driven.

Crucially, $\rho$and $\kappa$are not asserted to be universal scalars with a single canonical estimator. They are placeholders for a family of admissible operationalisations, and admissibility is controlled by the criteria below. This prevents the familiar slide from "we can measure something" to "we have measured the phenomenon."

### 2.4 Semiotic Adequacy Test (SAT) and the admissibility gate

This paper treats SAT and the criteria C1–C4 as inherited constraints from Huerta Castillo (2025), where they are motivated and linked to measurement via ρ/κ and explicit intervention testing do(I). For the present purpose—auditing semantic footprints—these criteria function as a gate on evidential entitlement: (C1) prevents "semantic" claims without domain-relevant constraints; (C2) enforces robustness against partition-generated artefacts via Π*; (C3) requires intervention-structured counterfactual signatures under do(I); and (C4) restricts "phase transition" talk to coherent endogenous reconfiguration rather than surface volatility. The protocol below operationalises these constraints without re-deriving their philosophical motivation.

### 2.5 Contextuality as an inference discipline

Semantic-footprint evaluation is inherently contextual: what a system appears to "mean" depends on how it is probed, what tools it can access, what retrieval regime is enabled, and what evaluative constraints are in force. A key methodological danger is to treat outputs from heterogeneous contexts as if they were samples from a single global state, thereby licensing illicit aggregation and inflated generality claims.

To prevent this, we adopt a contextuality strategy inspired by formal treatments in which measurement outcomes are indexed to measurement contexts, and compatibility constraints govern what can and cannot be jointly inferred (Abramsky and Brandenburger 2011). Operationally, this means: (i) $C$must be explicitly indexed to the measurement context; (ii) incompatible contexts (e.g., different tool-access regimes or prompting policies that alter available inferential resources) must not be collapsed into a single undifferentiated dataset; and (iii) claims of stability in $\rho$and structure in $\kappa$must be made *relative to admissible families of contexts*.

This contextuality discipline functions as an epistemic "non-overclaiming" constraint. It supports the paper's central demarcation: we are not describing what the system "is," but what can be responsibly inferred about its footprint-indexed behaviour under specified interventions and admissible measurement regimes.

### 2.6 From admissibility to IPT: the minimal dynamic signature

With SAT and C1–C4 in place, IPT can be stated as a falsifiable pattern rather than a metaphor. Let $\kappa(t)$be an operational estimate of context reconfiguration rate over an evaluation horizon $t$, and let $\rho(t)$be an operational estimate of relational semiotic density over the same horizon, both indexed to admissible context families and non-generating partitions.

IPT (dynamic signature, preliminary). An interpretive phase transition is declared only if (i) $\kappa(t)$exhibits a sustained acceleration attributable to endogenous reconfiguration of $C$under novelty and $do(I)$, (ii) do$(I)$-sensitivity remains directionally stable across replications, and (iii) $\rho(t)$does not collapse when evaluated under non-generating partitions.

Section 3 operationalises these parameters and makes the partition constraint explicit. Section 4 specifies the intervention suite $I_1 \dots I_n$and derives testable predictions that distinguish interpretive surrogacy from stylometric or biographical mimicry—an increasingly salient distinction given current evidence that generative systems can approximate stylistic profiles without thereby warranting footprint-level interpretive entitlement (Mikros 2025).

## 3. Operationalising $\rho$, $\kappa$, and Non-Generating Partitions $\Pi^*$

### 3.1. Measurement regimes and the admissibility rule

Sections 1–2 established that semantic-footprint claims are not licensed by behavioural similarity alone, but by an interventionist and context-sensitive evidential discipline. Section 3 turns that discipline into measurement procedures. This section is the protocol's logical firewall: it states what counts as admissible evidence.

The central methodological risk is *artefact-generation*: apparent "footprint persistence" can be manufactured by leakage, partition choices, or prompt–template reuse, and then reified as a measured property. To prevent that slide, this section treats $\rho$(relational semiotic density) and $\kappa$(context reconfiguration rate) as measurement-facing parameters whose numerical estimates are admissible only under explicitly stated constraints.

A measurement regime is defined as a tuple

$$\mathcal{R} := \langle D, \mathbb{C}, \Pi, \mathbb{I}, \mathbb{A}, \mathbb{T} \rangle,$$

where $D$is the bounded domain; $\mathbb{C}$is the family of admissible context conditions (including tool-access and policy constraints).

The protocol is deliberately negative: it specifies conditions under which evidence for semantic surrogacy is *not* licensed. Chief among these is the non-generating partition discipline, Π*. Π* is not an optional robustness check; it is a validity condition for treating any estimator of $\rho$or $\kappa$as probative. If Π* is violated, apparent "semantic persistence" is treated as an artefact of the evaluation design, and no IPT/DSS claim is admissible.

$\mathbb{I}$ is the intervention suite $\{do(I_j)\}$; $\mathbb{A}$is the answer/behavioural interface (output format constraints, scoring functions, allowed uncertainty); and $\mathbb{T}$ is the trace interface (what internal or operational traces are observable, if any, such as tool calls, memory gates, control-state logs, or other implementation-dependent state descriptors).

The admissibility rule is simple and strict:

Admissibility rule (SAT-gated measurement). Estimates of $\rho$and $\kappa$are admissible as evidence for semantic-operational persistence only if the evaluation passes SAT via C1–C4. If any criterion fails, the regime $\mathcal{R}$may still yield predictive performance metrics, but it does not license footprint-level inference.

This rule turns C1–C4 into an explicit "logical firewall": it blocks the familiar move from "we can compute a statistic" to "we have measured the phenomenon." It also makes clear why $\rho$and $\kappa$are not "free parameters": they are constrained by (i) triadic grounding (C1), (ii) partition robustness (C2), (iii) intervention-structured counterfactual sensitivity (C3), and (iv) endogenous context reconfiguration (C4).

Finally, Section 3 enforces a discipline on claims of generality: all reported measurements must be indexed to a declared measurement regime $\mathcal{R}$. This prevents illicit aggregation across incompatible contexts and aligns with the contextuality constraint introduced in Section 2 (Abramsky and Brandenburger 2011). In practice: $\rho$and $\kappa$are not properties of a model "in general," but of a model under a specified regime of probing and constraint.

Table 1 enumerates which estimators are admissible *as evidence* under SAT's inherited constraints, conditional on Π* compliance. Metrics listed as inadmissible may still be reported descriptively, but they cannot license semantic-footprint or IPT claims and must never be used as primary evidence.

**Table 1. Admissible vs. inadmissible operationalisations of ρ and κ**

| Target | Operationalisation (candidate estimator) | Status | Why admissible / inadmissible (mapped to C1–C4) | Typical artefact it blocks / induces |
|---|---|---|---|---|
| ρ | $\rho_{obj}$: object-sensitivity density — count/strength of stable sign–object–interpretant linkages across object-relevant perturbations, controlling for superficial paraphrase | Admissible | Requires C1 (object constraint must matter); tested under do(I) (C3); evaluated under $\Pi^*$ (C2) | Blocks stylometric inflation; detects "meaning" that tracks object constraints |
| ρ | $\rho_{\Delta}$: intervention-structure richness — diversity and stability of directionally predicted effects $\Delta_{Ij}$ across the intervention suite | Admissible | Directly enforces C3; remains meaningful only if C2 holds; supports C1 by tying effects to structured constraints | Blocks "prompt sensitivity"; detects reproducible counterfactual profiles |
| ρ | $\rho_{inv}$: invariance/coherence score — invariance to irrelevant re-encodings + coherence under relevant constraint shifts | Admissible | Encodes C1 via relevance structure; requires C2 to avoid leakage; interpretable only with C3 replication | Blocks template dependence; detects stable footprint-consistent commitments |
| ρ | Lexical/stylometric similarity (n-gram overlap, authorial markers, embedding cosine on raw text) | Inadmissible (alone) | Violates C1 (no object anchoring) and can pass without C3; highly sensitive to generating partitions (violates C2) | Induces "ghost" persistence from surface style; confounds identity keys with competence |
| ρ | Single-score "human-likeness" or preference-model ratings without intervention contrasts | Inadmissible | Violates C3 (no intervention structure) and does not establish C1; can be partition-generated (C2) | Induces anthropomorphic overread; rewards safe generic responses |
| κ | $\kappa_\Gamma$: rate of structured changes in an operational context descriptor Γ(t) under novelty and do(I), with replication | Admissible | Requires C4 (endogenous reconfiguration) and C3 (do-indexed structure); must be robust under $\Pi^*$ (C2) | Blocks "flat context" systems; distinguishes endogenous reconfiguration from noise |

| | | | | |
|---|---|---|---|---|
| $\kappa$ | $\kappa_{cp}$: change-point density — frequency/strength of detected regime shifts in $\Gamma(t)$ or decision policies under intervention sequences | Admissible | Implements C4 as a detection problem; depends on C3 (intervention-indexed shifts) and C2 (no leakage-driven shifts) | Blocks gradual drift misread as regime change; detects structured reconfiguration |
| $\kappa$ | $\kappa_{repair}$: repair-driven reconfiguration — rate/quality of coherent repairs when constraints conflict (contradiction, misleading cues, policy shifts) | Admissible | Tests C4 under conflict; requires C1 (object constraints) and C3 (controlled perturbations), and $\Pi^*$ (C2) | Blocks brittle mimicry; distinguishes repair from rationalisation |
| $\kappa$ | Token-level volatility (variance in outputs across paraphrases) | Inadmissible | Confuses superficial sensitivity with context reconfiguration; violates C4 and often C3 (unstructured variance) | Induces false “high $\kappa$” from instability; rewards randomness |
| $\kappa$ | Training-step or compute-based proxies (parameter count, FLOPs, gradient norms) | Inadmissible (as evidence) | Not indexed to $\mathcal{R}$; does not satisfy C1–C4; licenses complexity→agency fallacy | Induces “scaling is singularity” rhetoric; ignores auditability |
| Both | Aggregating across incompatible contexts without indexing $\mathbb{C}$ (tool access, memory gating, policy constraints) | Inadmissible | Violates contextuality discipline; undermines C3 and C4 interpretation | Induces illicit generality; mixes regimes into a pseudo-global state |

“Inadmissible” here means “does not license semantic-footprint inference on its own.” Some inadmissible metrics can still be reported as auxiliary diagnostics, but not as primary evidence for IPT/DSS claims. Admissibility is conditional: even an admissible estimator becomes non-probative under $\Pi^*$-violating partitions. $\Pi^*$ plays the same inferential role as blinding/randomisation in experimental design: without it, apparent effects may be entirely generated by the protocol rather than by the target competence.

### 3.2 Operationalising relational semiotic density $\rho$

Relational semiotic density $\rho$is intended to measure how richly and stably a system sustains triadic relations: signs are interpreted as about objects (or task-defined success conditions) and yield interpretants that guide action, explanation, or evaluation within the domain. Because $\rho$is easily inflated by stylistic mimicry, all admissible operationalisations of $\rho$must be anchored to (i) a relevance structure (what changes should and should not matter) and (ii) an intervention structure (which changes are imposed via $do(I)$rather than by uncontrolled prompt variation).

We therefore treat $\rho$as a family of estimators indexed to a measurement regime $\mathcal{R}$:

$$\rho := \rho(\mathcal{R}) = \rho(\langle D, \mathbb{C}, \Pi, \mathbb{I}, \mathbb{A}, \mathbb{T}\rangle).$$

A minimal admissible operationalisation begins by defining a set of *object-relevant perturbations* $P_{\mathrm{rel}}$and *object-irrelevant perturbations* $P_{\mathrm{irr}}$for the domain. $P_{\mathrm{rel}}$alters object constraints (reference, success conditions, normative frame, commitments), whereas $P_{\mathrm{irr}}$alters surface form (paraphrase, formatting, lexical choices) while preserving object constraints.

Definition (object-sensitivity density). Let $f_{\mathcal{R}}$denote the system's response function under regime $\mathcal{R}$, and let $S$denote a set of probe items with explicitly defined object constraints. Define a stability operator $Stab(\cdot)$that measures invariance under $P_{\mathrm{irr}}$and a sensitivity operator $Sens(\cdot)$that measures directional change under $P_{\mathrm{rel}}$as predicted by the intervention suite. Then an admissible estimator takes the schematic form

$$\rho_{\mathrm{obj}}(\mathcal{R}) := \mathbb{E}_{s \in S}[Stab(f_{\mathcal{R}}(s; P_{\mathrm{irr}}))] \cdot \mathbb{E}_{s \in S}[Sens(f_{\mathcal{R}}(s; P_{\mathrm{rel}}))],$$

with both terms computed under $\Pi^*$and replicated across context families $\mathbb{C}$.

This form encodes the intended asymmetry: a footprint-consistent surrogate should be *stable* under irrelevant changes and *sensitive* under relevant ones. In practice, the operators $Stab$and $Sens$can be instantiated using (i) structured scoring rubrics tied to object constraints, (ii) model-based evaluators that are themselves intervention-audited and partition-robust, or (iii) task success measures that cannot be shortcut via identity keys. The key is that admissibility is not conferred by the particular scoring function, but by the regime constraints that guarantee object anchoring and non-generating partitions.

A complementary admissible estimator captures the richness of intervention structure:

Definition (intervention-structure richness). For each intervention $do(I_j)$, define a predicted directional effect $\Delta_{\mathrm{Ij}}$ on an observable outcome class (commitments, explanations, tool selection, error correction, etc.). Let $Rep(\Delta_{I_j})$ measure replication stability across repeated trials and contexts, and let $Div(\{\Delta_{I_j}\})$ measure diversity across interventions (non-redundancy). Then

$$\rho_{\Delta}(\mathcal{R}) := Div(\{\Delta_{I_j}\}_{j=1}^{n}) \cdot \mathbb{E}_j[Rep(\Delta_{I_j})],$$

computed under $\Pi^*$and indexed to $\mathbb{C}$.

This estimator is explicitly anti-stylometric: it treats "richness" as richness of *counterfactual structure*, not of linguistic flourish. A system can be eloquent and still fail $\rho_{\Delta}$if it does not exhibit stable, predicted intervention effects.

Finally, a third admissible route focuses on invariants:

Definition (invariance/coherence). Let $Inv$ measure invariance under $P_{\text{irr}}$, and let $Coh$ measure coherence of commitments across object-relevant reframings under $P_{\text{rel}}$and $do(I)$. Then

$$\rho_{\text{inv}}(\mathcal{R}) := \mathbb{E}[Inv] \cdot \mathbb{E}[Coh],$$

again computed under $\Pi^*$and with replication.

The role of $\rho$in the overall framework is not to declare that "more density is better" in the abstract, but to provide a measurable account of whether a system's interpretive behaviour is rich enough to support footprint-indexed surrogacy under intervention, without collapsing into generic assistant behaviour or into style mimicry.

### 3.3 Operationalising context reconfiguration rate *κ*

Where $\rho$concerns the richness of triadic relations, $\kappa$concerns the dynamics of context management: how a system revises its operative context state $\mathcal{C}$under novelty and intervention. Since $\mathcal{C}$is implementation-dependent, the framework requires an operational descriptor $\Gamma(t)$that can be observed or reconstructed from traces allowed by $\mathbb{T}$. $\Gamma(t)$can include, depending on the system, any subset of: tool-use regimes, memory gates, retrieval policies, declared constraints and commitments, policy adherence states, explicit problem representations, or other control variables that influence interpretation.

The admissibility constraint is that $\Gamma(t)$must be neither (i) a mere encoding of the most recent input nor (ii) a purely exogenous script imposed by the evaluator. It must capture structured system-level reconfiguration that is *elicited* by novelty and $do(I)$but not trivially dictated by prompt templates.

An admissible estimator is then:

Definition ($\kappa_\Gamma$, structured reconfiguration rate). Let $d(\Gamma(t), \Gamma(t-1))$be a distance or change function over operational context descriptors (e.g., a weighted edit distance over policy states, a divergence over tool regimes, or a change score over commitment sets), and let $Norm(\cdot)$normalise for input magnitude. Then

$$\kappa_\Gamma(\mathcal{R}) := \mathbb{E}_t[\frac{d(\Gamma(t), \Gamma(t-1))}{Norm(\text{input}(t))}],$$

computed under replication and under $\Pi^*$.

This estimator captures "how much the system reconfigures per unit of novelty," rather than raw volatility. A stable surrogate can have high $\kappa_\Gamma$when novelty demands reconfiguration, and low $\kappa_\Gamma$when it should remain context-stable.

A second admissible estimator treats reconfiguration as a regime-change detection problem:

Definition ($\kappa_{\text{cp}}$, change-point density). Let $\{\Gamma(t)\}$be a time-indexed sequence under a specified intervention schedule. Apply a change-point detector $CP(\cdot)$that yields a set of detected change points $\{t_k\}$with confidence weights $w_k$. Then

$$\kappa_{\text{cp}}(\mathcal{R}) := \sum_k w_k \,/\!\mid W \mid,$$

where $W$is the evaluation window length. IPT claims will later require not only that change points exist, but that their density/weight increases in a sustained way under the relevant conditions.

A third admissible estimator ties $\kappa$to repair under conflict:

Definition ($\kappa_{\text{repair}}$, repair-driven reconfiguration). Consider a set of conflict probes $S_{\text{conf}}$in which constraints clash (e.g., misleading contextual hints, contradictory commitments, policy shifts). Let $Repair(\cdot)$score whether the system performs coherent repair (rejects the misleading cue, revises $\mathcal{C}$, maintains footprint-consistent commitments), and let $Reconf(\cdot)$quantify the accompanying reconfiguration in $\Gamma$. Then

$$\kappa_{\text{repair}}(\mathcal{R}) := \mathbb{E}_{s \in S_{\text{conf}}}[Repair(s) \cdot Reconf(s)],$$

computed under $\Pi^*$and replication.

This estimator is designed to separate genuine reconfiguration from post-hoc rationalisation. A system that merely produces plausible explanations without structured constraint repair can score well on superficial evaluations but will fail repair-driven $\kappa$.

As with $\rho$, no single estimator is canonically privileged at this stage. What matters is that the chosen estimator is (i) indexed to $\mathcal{R}$, (ii) SAT-gated, and (iii) interpretable under the contextuality discipline.

### 3.4 Constructing non-generating partitions $\Pi^*$

Non-generating partitions are the principal defence against artefact-generation in semantic-footprint evaluation. The aim is not to make evaluation “harder” in a generic sense, but to prevent the partition protocol from encoding shortcuts that can be exploited to simulate footprint persistence.

A partition protocol $\Pi^*$is non-generating if, relative to the target outputs and scoring functions, it removes or quarantines (i) identity keys, (ii) near-duplicates and template-sharing, (iii) temporal adjacency shortcuts, and (iv) shared context scaffolds that can leak constraints across splits. Formally: the partition rule itself must have low mutual information with the target labels/outputs once object constraints are held fixed.

Operationally, $\Pi^*$is constructed by the following steps, which can be pre-registered:

1. De-duplication and near-duplicate removal. Remove exact duplicates; cluster near-duplicates via conservative similarity thresholds (lexical and semantic) and keep at most one exemplar per cluster.
2. Template quarantine. Identify recurring prompt templates, scaffolds, and evaluation rubrics; ensure templates do not reappear across splits in forms that preserve predictive structure.
3. Identity-key stripping. Remove or mask explicit identifiers, unique phrases, and stable biographical tokens that trivially re-identify the target.
4. Temporal and topical separation. Enforce separation rules that prevent time-adjacent continuation prediction and prevent repeated topical sequences from straddling splits.
5. Context-scaffold isolation. Ensure that shared context states (tool policies, memory allowances, constraint sets) do not leak across splits in ways that make “context competence” untestable.

6. Adversarial leakage checks. Attempt to predict split membership or target identity from the remaining data; if such prediction is easy, the partition is likely generating and must be revised.
7. Replication across partitions. Repeat key results under multiple $\Pi^*$-variants $\Pi_1^*, \Pi_2^*$to ensure stability is not an artefact of one particular construction.

A partition is generating (thus invalid for footprint inference) if it enables high apparent fidelity by any of: (i) cross-split reuse of identity keys or near-duplicates; (ii) time-adjacent continuation; (iii) template reuse that encodes the target behaviour; or (iv) hidden metadata leakage. Generating partitions are not merely “suboptimal”; they can create the appearance of interpretive competence where none is warranted.

### 3.5 IPT detection rule (pre-registrable signature)

With admissible estimators in place, IPT can be stated as a falsifiable detection rule rather than a rhetorical label. Let $W$be a pre-specified evaluation window, and let $\mathcal{R}_0$denote a baseline regime and $\mathcal{R}_1$denote a regime that induces novelty and/or contextual pressure under the intervention suite, both employing non-generating partitions $\Pi^*$and admissible context families $\mathbb{C}$.

IPT detection rule (preliminary). Declare an interpretive phase transition only if, over $W$:

1. Sustained $\kappa$-acceleration. $\kappa(\mathcal{R}_1)$exceeds $\kappa(\mathcal{R}_0)$by a pre-registered margin in a sustained manner (not as isolated volatility), as evidenced by $\kappa_\Gamma$and/or $\kappa_{\text{cp}}$and corroborated by repair-driven structure where applicable.
2. Preserved do-sensitivity. The intervention-effect profiles $\{\Delta_{I_j}\}$ remain directionally stable under replication and under $\Pi^*$; i.e., $\rho_\Delta$does not collapse.
3. Non-collapse of relational coherence. $\rho(\mathcal{R}_1)$does not collapse relative to $\rho(\mathcal{R}_0)$under $\Pi^*$; if $\rho$increases, the increase must be robust to shortcut removal and not attributable to generating partitions.
4. Context-indexed reporting. All measurements are indexed to declared regimes $\mathcal{R}$, and no aggregation across incompatible context families is used to support the IPT claim.

This rule yields explicit failure codes that prevent post-hoc narrative rescue:

- IPT-0: No measurable $\kappa$-acceleration under admissible estimators.
- IPT-1: $\kappa$increases, but do-sensitivity collapses or becomes unstable (C3 failure).
- IPT-2: $\rho$collapses under $\Pi^*$, indicating brittle mimicry or loss of coherence.
- IPT-3: Apparent $\rho/\kappa$ effects depend on generating partitions or leakage channels (C2 failure).
- IPT-4: Effects require aggregation across incompatible contexts (contextuality violation).

Section 4 specifies the intervention suite $\{do(I_j)\}$and derives the concrete, falsifiable predictions that distinguish a domain-bounded semiotic surrogate from stylometric or biographical mimicry.

## 4. Intervention suite and IPT detection protocol

Section 3 established admissible operationalisations of ρ and κ under the non-generating partition discipline $\Pi^*$. Section 4 completes the methodological core by specifying (i) a minimal suite of explicit interventions *do(I)* and (ii) a pre-registrable criterion for detecting an interpretive phase transition (IPT) as a dynamical pattern across interventions, rather than as a label applied post hoc. The guiding assumption is interventionism: persistence claims are treated as testable counterfactual hypotheses, not as interpretive impressions (Woodward, 2004; Pearl, 2009).

### 4.1 Intervention principles

All interventions are evaluated under Π-compliant partitions. Accordingly, do(I) results are treated as evidentially meaningful only when the underlying partition scheme satisfies Π*; otherwise intervention signatures may be artefacts of leakage or regime mixing rather than of semantic-footprint preservation.

We define an evaluation regime $\boldsymbol{\mathcal{R}}$ as the tuple $\langle \mathbb{C}, \mathrm{T}, \mathrm{M}, \mathrm{A}, \Pi^* \rangle$, where $\mathbb{C}$ indexes context constraints (tool access, policy constraints, and interaction affordances), T specifies the task family, M the memory/trace interface, A the admissible assistance or scaffolding (including prompting conventions), and $\Pi^*$ the non-generating partition discipline. Interventions *do(I)* act by setting or toggling one component of $\boldsymbol{\mathcal{R}}$ while holding the remaining components fixed up to an explicitly declared tolerance. The suite is designed to satisfy four principles.

Minimality. Each intervention targets one primary degree of freedom, avoiding compound manipulations that would undercut attribution.

Orthogonality. The suite spans distinct failure modes: style mimicry, identity-key leakage, context-insensitive genericity, and brittle compliance.

Auditability. Every intervention is specified as a reproducible change to the regime tuple, logged with versioned artefacts (prompts, tools, policies, and trace windows).

Replicability. Effects are evaluated across repeats and across non-generating partitions; directional signatures are expected to be stable even when surface form varies.

### 4.2 Intervention suite

Table 2 specifies a minimal intervention suite $\mathrm{do}(I)$intended for preregistration: a small set of context-, task-, and memory-level interventions that (i) generate directional predictions for $\rho$and $\kappa$, and (ii) can be replicated under declared measurement regimes $\mathcal{R}$. All intervention effects are evaluated under Π*-compliant partitions; otherwise apparent signatures may be partition-generated artefacts rather than footprint-indexed structure.

Whereas Huerta Castillo (2025) introduces $\mathrm{do}(I)$as an intervention operator within the broader SAT-based account of minimal artificial agency, the present paper contributes a minimal preregistrable intervention suite (Table 2) and an associated discipline of directional predictions for $\rho$and $\kappa$under $\Pi^*$ constraints, turning footprint claims into testable protocol commitments.

**Table 2. Minimal intervention suite do(I) with predicted admissible signatures in ρ and κ (evaluated under Π* and indexed by ℂ).**

| ID | do(I) target | Intervention | Expected admissible signature (ρ / κ) | Pass criterion (mapped to C1–C4) | Typical confound blocked |
|---|---|---|---|---|---|
| I1 | T | Object-constraint swap: hold linguistic style and format constant while swapping object constraints within the same task family. | $\rho_{obj}$ decreases under object swap; $\kappa_\Gamma$ shows structured reconfiguration in Γ(t) rather than instability. | C1: commitments track object constraints; fail if object swap leaves object-relevant commitments unchanged or if style invariants dominate. | Stylometric inflation; template mimicry. |
| I2 | T | Irrelevant re-encoding: paraphrase, translation, or surface reformatting that preserves the object constraint. | $\rho_{inv}$ remains high under irrelevant re-encoding; $\kappa_\Gamma$ remains low-to-moderate (no regime change required). | C1–C2: invariance under irrelevant transforms; fail if performance collapses under prompt-template scrambling or depends on partition idiosyncrasies. | Prompt sensitivity; formatting dependence. |
| I3 | ℂ | Context gating: toggle tool access or policy constraints while keeping task and memory fixed (tools on/off; constrained/unconstrained). | $\rho_\Delta$ exhibits stable directional deltas across toggles; $\kappa_{cp}$ aligns changepoints with the intervention boundary, not with random drift. | C3–C4: do-indexed, replicable regime shifts; fail if outputs are aggregated across ℂ as if global or if κ shifts without structured dependence on the toggle. | Illicit generality across incompatible contexts. |
| I4 | M | Trace ablation: remove person-indexed traces (biographical keys, idiosyncratic corpora) while holding the task family constant. | $\rho_{obj}$ and $\rho_\Delta$ remain non-trivial without identity keys; $\kappa_{repair}$ increases via coherent repair rather than hallucination. | C2–C3: persistence cannot be explained by identity-key leakage; fail if “fidelity” collapses to stylometry or to generic safe responses. | Identity-key leakage; memorised biography. |
| I5 | M | Trace supplementation: inject a bounded, audited trace window | $\rho_{obj}$ increases in object sensitivity under audited trace injection; $\kappa_\Gamma$ | C3–C4: predictable effects under controlled trace injec- | Overfitting to injected snip- |

| | | | | | |
|---|---|---|---|---|---|
| | | with explicit provenance and versioning. | accelerates only insofar as new trace is coherently integrated into $\Gamma(t)$. | tion; fail if gains depend on generating partitions or if $\kappa$ spikes as unstructured volatility. | pets; uncontrolled retrieval. |
| I6 | A | Conflict injection: introduce mutually inconsistent cues or constraints under controlled disclosure (contradiction, misleading hints, policy conflicts). | $\kappa_{repair}$ increases under controlled conflict; $\rho_{inv}$ remains stable after repair; increases in $\kappa_{\Gamma}$ must be structured, not volatile. | C4: endogenous repair and reconfiguration; fail if the system rationalises inconsistencies without revising operative context or yields non-replicable swings. | Brittle mimicry; post hoc rationalisation. |
| I7 | T | Within-domain cross-task probing: switch among tasks that probe the same object constraint (explain, decide, critique, justify) with $\mathbb{C}$ fixed. | $\rho_{obj}$ remains stable across within-domain task reframings; $\kappa_{\Gamma}$ reconfigures appropriately without erasing object commitments. | C1–C3: object-grounded commitments persist under task reframing; fail if competence is task-local or collapses to generic discourse. | Task-local heuristics; instruction-following without object anchoring. |
| I8 | $\mathbb{C}$/M | Temporal spacing and replay: evaluate under declared replay of interaction state versus fresh context, controlling session history. | $\rho_{\Delta}$ profiles reproduce under declared replay conditions; $\kappa_{cp}$ separates genuine regime transitions from temporal drift and hidden-state contamination. | C2–C3: determinism under declared replay conditions; fail if changes are explained by hidden-state leakage or uncontrolled history effects. | Hidden-state contamination; non-reproducible session drift. |
| I9 | $\Pi^*$ | Partition stress-test: repeat I1–I8 across alternative non-generating partitions (different splits, masking strategies, identity-key scrubbing). | Admissible $\rho/\kappa$ estimates remain within tolerance across $\Pi^*$ variants; stylometric proxies diverge. | C2: robustness to partition choice; fail if key results disappear when leakage paths are blocked. | Partition-generated artefacts. |
| I10 | $\mathbb{C}$/T | Adversarial epistemic probing: search for prompts that maximise stylometric similarity while minimising object satisfaction (and vice versa). | $\rho_{obj}$ and $\rho_{\Delta}$ resist stylometric maximisation; $\kappa_{\Gamma}$ remains structured under object-driven probes, iso- | C1–C3: object sensitivity dominates style similarity; fail if high “fidelity” is achievable without object tracking or do-indexed structure. | Surface-level “ghost” persistence. |

| | | | | | |
|---|---|---|---|---|---|
| | | | lating mimicry regimes when it fails. | | |

## 4.3 Quantifying intervention effects

For each intervention *do(I)*, we evaluate an effect profile over admissible estimators. Let $\hat{\rho}$ denote any admissible density estimator family (e.g., $\rho_{obj}$, $\rho_{\Delta}$, $\rho_{inv}$), and let $\hat{\kappa}$ denote any admissible reconfiguration estimator family (e.g., $\kappa_{\Gamma}$, $\kappa_{cp}$, $\kappa_{repair}$). We define intervention deltas as $\Delta_{Ij}(\hat{\rho}) := \hat{\rho}(\boldsymbol{\mathcal{R}} \mid do(Ij)) - \hat{\rho}(\boldsymbol{\mathcal{R}}0)$, and analogously for $\hat{\kappa}$. The empirical object is not a single score but a vector of deltas, whose directionality and stability across replications constitute evidence for (or against) footprint-indexed surrogacy (Pearl, 2009; Woodward, 2004).

A DSS claim is supported only if the delta profile is (i) do-indexed (the direction of change is predictable from the intervention design), (ii) robust under $\Pi^{*}$, and (iii) stratified by $\mathbb{C}$ rather than illicitly aggregated across incompatible contexts (Abramsky and Brandenburger, 2011).

## 4.4 IPT detection rule

We define an IPT as a regime shift in interpretive dynamics under which context reconfiguration accelerates in a sustained, structured manner without destroying do-sensitivity or collapsing admissible relational density. Operationally, for a predefined window W and a predefined intervention schedule S = (I1,…,In), an IPT is declared only if all of the following hold.

1. IPT-1 (Sustained acceleration of  under structured reconfiguration).  (κ) The median of $\hat{\kappa}$ (e.g., κ_Γ) increases beyond a predeclared threshold for at least |W| observations and exhibits change-points aligned with intervention boundaries; volatility unaligned with do(I) is treated as noise, not as reconfiguration.
2. IPT-2 (Preserved do(I)-sensitivity). The sign pattern of Δ_Ij($\hat{\rho}$) and Δ_Ij($\hat{\kappa}$) is replicable across repeats and across Π* variants; if effects reverse unpredictably or disappear under leakage-blocking partitions, the claim is rejected.
3. IPT-3 (Non-collapse of admissible  under ).  (ρ) ($\Pi^{*}$) At least one object-anchored density estimator (ρ_obj or ρ_inv) remains above a predeclared floor under object-relevant perturbations; a rise in κ coupled with a collapse of ρ is interpreted as destabilisation, not as an IPT.
4. IPT-4 (Explicit domain-boundedness). The IPT claim is indexed to a declared domain of objects and tasks; cross-domain aggregation is prohibited, and failures outside the declared domain do not count as falsifiers or confirmations of the within-domain claim.

These conditions are intentionally conservative. They are designed to prevent two failure modes that are common in surrogate discussions: (i) rebranding volatility as adaptive reconfiguration, and (ii) mistaking partition-generated similarity for object-anchored persistence. The framework therefore privileges auditability over maximality: it asks for structured, do-indexed evidence that survives robustness checks, rather than for an overall impression of "human-likeness". Section 6 specifies failure diagnostics for cases where predicted signatures are not recovered under $\Pi$*-compliant partitions.

### 4.5 Reporting and audit packet

To keep footprint claims falsifiable and portable across evaluators, every IPT assessment should be accompanied by an audit packet. At minimum, the packet includes: (i) a complete specification of the regime tuple $\boldsymbol{\mathcal{R}}$ and its versioned components ($\mathbb{C}$, T, M, A); (ii) the full intervention schedule S with exact do(I) settings; (iii) the non-generating partition procedure $\Pi^*$ and identity-key scrubbing rules; (iv) replication counts, replay settings, and randomness controls; and (v) estimator definitions for $\hat{\rho}$ and $\hat{\kappa}$, including any thresholds used in IPT-1–IPT-4. This aligns with auditability-oriented governance approaches that emphasise documentation, traceability, and intervention points, while keeping the epistemic goal distinct from compliance or certification (Mökander and Axente, 2023; Mökander, 2023; Hagendorff, 2024).

Finally, the audit packet should be interpreted under the paper's non-claims. A DSS is not a person, and IPT is not a metaphysical discontinuity. The point is to make semantic-operational persistence experimentally accountable: if the delta profile does not replicate, if it collapses under $\Pi^*$, or if it disappears once $\mathbb{C}$ is made explicit, then the appropriate conclusion is not that the phenomenon is mysterious, but that the surrogate claim is unsupported under the proposed discipline of evidence.

## 5. DSS/IPT among avatars, proxies, and stylometry

Sections 2–4 established a disciplined object of inquiry—semantic-operational persistence—and a corresponding discipline of evidence: admissible estimators for $\rho$ and $\kappa$ evaluated under non-generating partitions $\Pi^*$ and audited through explicit interventions *do(I)*. Section 5 positions this proposal against three nearby families of discourse that are often conflated in public and academic discussion: (i) avatar and proxy relations (representation and delegated action), (ii) metaphysically loaded readings of "digital survival", and (iii) stylometric and biographical mimicry. The point of this section is not to adjudicate those debates in general, but to clarify what DSS/IPT adds: a conservative, interventionist test discipline for when a system is evidentially entitled to be treated as a footprint-indexed surrogate at all.

### 5.1 Avatars and proxies: representation is not surrogacy

The contemporary literature on avatars often begins from the social fact that a digital entity can represent a person in technologically mediated environments. On proxy models, the avatar's actions may count as the represented person's actions in ethically relevant ways, while an epistemic gap remains between the represented agent and the actions attributed through the proxy (Sweeney, 2023). A complementary strand argues that avatars may be better understood as parts of persons—temporal or functional parts—rather than as external proxies, thereby reframing responsibility and identity questions (Patrone, 2025).

DSS/IPT does not compete with these proposals at the level of metaphysics or responsibility attribution. Instead, it targets a prior methodological question: what evidence justifies treating an artefactual system as a footprint-indexed surrogate with stable, object-anchored dispositions, rather than as a mere representational interface or delegation device? Proxy and part frameworks can be compatible with DSS/IPT, but they do not, by themselves, supply a measurement discipline for "footprint preservation". In our terms, a system can be an avatar or proxy without satisfying the admissibility gates C1–C4, and without exhibiting any intervention-stable profile in ρ and κ. DSS/IPT therefore introduces a demarcation between social role and epistemic entitlement: representation is a social relation; surrogacy, as used here, is an evidential status conferred only under *do(I)* and $\Pi^*$.

### 5.2 Digital afterlife framings: identity talk as a category error

A second family of discourse treats high-fidelity post-mortem representation as a form of "digital survival" or "afterlife". This framing is ethically and politically consequential, but it is also methodologically hazardous because it encourages identity-laden inferences from representational success. Critical work on the digital afterlife industry shows how commercial incentives and social dynamics can shape post-mortem representations, and how easily "presence" can be misread as persistence of a person (Öhman and Floridi, 2017).

DSS/IPT adopts a strict non-identity stance by construction: the object of persistence is a semantic footprint, not a person. Consequently, any claim about survival, phenomenological continuity, or personal identity is ruled out of scope rather than refuted. This scope rule is not evasive; it is a methodological guardrail. If we allow identity predicates to enter the evidential base, then almost any sufficiently persuasive mimicry will appear as "continuity". By contrast, once the target is explicitly semantic-operational—i.e., a structured disposition profile under *do(I)*—the central question becomes auditability: do the footprint-indexed commitments remain stable under object-relevant perturbations (C1), survive leakage-blocking partitions (C2), and yield reproducible intervention signatures (C3–C4)?

### 5.3 Stylometry and biographical mimicry: "ghost" fidelity without object anchoring

A third nearby family of practice aims to reconstruct authorship, identity, or "voice" using surface regularities—lexical choices, syntactic patterns, and stylistic markers. Stylometry is an established and valuable toolkit for authorship attribution and related tasks, and modern surveys show a rich landscape of methods and applications (Misini, 2022). The present paper does not deny that stylometric similarity can be high, or that it can be predictive. The methodological claim is narrower: stylometry, taken alone, is not admissible evidence of semantic-operational persistence under our criteria, because it can succeed without object anchoring (violating C1), without intervention structure (violating C3), and in ways that are highly sensitive to partition choices (violating C2).

This motivates the “ghost persistence” diagnosis: apparent continuity can be generated by surface resemblance even when the system’s outputs do not track the object constraints that define a footprint. In Warburg’s vocabulary, one might call these outputs ghost-images—pathos-laden traces that preserve surface valence and idiom—yet this remains metaphorical unless translated into the operational discipline of Sections 3–4. In the present framework, what matters is not stylistic likeness but structured sensitivity: $\rho_{obj}$ must vary appropriately under object-constraint swaps (I1), $\rho_{\Delta}$ must reproduce directionally across interventions (I3–I10), and $\kappa_{repair}$ must increase under controlled conflict (I6) through coherent reconfiguration rather than unstructured volatility. Where stylometric similarity is maximised while object satisfaction is minimised (I10), the surrogate claim is rejected, even if the impression of “voice” is strong.

### 5.4 Contextuality as an inference discipline: preventing illicit aggregation

The preceding contrasts depend on a further methodological point: evaluation contexts are not neutral containers. Tool access, memory gating, policy constraints, and prompting conventions can change what is being measured. Without explicit indexing by $\mathbb{C}$, it is easy to aggregate across incompatible regimes and infer a pseudo-global competence. Our “quantum-compatible” contextuality stance is a way of preventing this illicit aggregation: what can be inferred depends on the measurement context, and incompatible contexts must not be collapsed into a single unqualified description (Abramsky and Brandenburger, 2011). This is an epistemic constraint, not a claim about quantum mechanisms in minds or machines. Its practical role in DSS/IPT is to force the audit packet to declare $\mathbb{C}$ and to treat “the system” as a family of context-indexed behaviours rather than as a context-free agent.

### 5.5 Why auditability is the common denominator

The contrasts above converge on a single claim: DSS/IPT relocates debates about avatars, afterlife representation, and mimicry from impressionistic resemblance to intervention-auditable evidence. This shift resonates with auditability-oriented approaches to AI governance that emphasise systematic evidence collection, documentation, and reproducible evaluation (Hagendorff, 2020; Mökander, 2023). However, the goal here is epistemic before it is regulatory. Auditability is the common denominator because it is the only route by which semantic-operational persistence claims can become falsifiable rather than rhetorical. 3 summarises the demarcations that will be used in Section 6 to state limits, risks, and governance-adjacent implications without slipping into identity persistence or metaphysical continuity claims.

**Table 3. Framework comparison: DSS/IPT vs. avatars/proxies vs. stylometry**

| Dimension | DSS / IPT | Avatars / proxies / parts | Stylometry / “voice” mimicry |
|---|---|---|---|
| Primary object of claim | Footprint-indexed semantic-operational persistence within a declared domain | Representation and delegated action relations between person and avatar; | Surface regularities of style or authorial markers; sometimes identity linking or attribution. |

| | | | |
|---|---|---|---|
| | (no identity, no phenomenology). | responsibility and agency attribution. | |
| Core evidential standard | Admissibility gates C1–C4 + intervention profiles under *do(I)* + robustness under $\Pi^*$. | Conceptual/ethical fit of proxy or part relation; may not specify measurement admissibility. | Similarity scores (n-grams, embeddings, stylometric features), often without object anchoring. |
| Role of interventions | Central: *do(I)* defines testable counterfactual hypotheses; $\Delta_{Ij}$ profiles are the evidence. | Secondary: interventions matter insofar as they affect delegated action or representation. | Often indirect: sensitivity to prompts/paraphrase may be measured but not as structured do(I). |
| Context treatment | Explicit indexing by $\mathbb{C}$; incompatible contexts are not aggregated (contextuality discipline). | Typically role-based; context matters socially (where the proxy acts) more than inferentially. | Context often treated as noise; aggregation can inflate apparent continuity. |
| Typical failure mode | Partition-generated artefacts; volatility misread as reconfiguration; illicit cross-context generalisation. | Responsibility gaps; misattribution of actions; epistemic gaps between proxy and represented agent. | Ghost persistence: high “voice” fidelity without object constraint tracking; identity-key leakage. |
| What it does not claim | No survival, no personal identity persistence, no afterlife; IPT is not an ontological rupture. | May or may not take a stance on identity depending on the theory; not inherently about auditability. | Not evidence of interpretive competence or footprint preservation unless supplemented by object-anchored interventions. |
| Best use-case | Auditable surrogacy for bounded domains where footprint-indexed dispositions matter (decision support, constrained dialogue). | Normative analysis of delegation, representation, and responsibility in mediated environments. | Authorship attribution, forensic analysis, style detection, and related pattern-based tasks. |

Having positioned the framework, the next step is to articulate limits, risks, and governance-adjacent implications without relaxing the non-claims. Section 6 therefore treats IPT/DSS as a conservative evidential status that can fail for principled reasons (partition artefacts, context collapse, or loss of do-sensitivity), and it specifies what kinds of empirical programmes could meaningfully support—or falsify—the proposal.

## 6. Limits, failure modes, and audit-adjacent implications

The preceding sections defined DSS/IPT as a conservative evidential status, not as a metaphysical thesis. This section makes the conservatism explicit by (i) fixing hard scope limits and non-claims, (ii) specifying principled failure modes and their diagnostics, and (iii) clarifying what would count as supportive evidence versus falsification for an IPT claim. The aim is to ensure that "no IPT" is a well-formed and expected outcome whenever admissibility, robustness, or intervention structure fails.

Π violation is treated as a failure mode with mandatory re-run. When Π* screening fails, any apparent ρ/κ signatures are classified as non-probative for semantic surrogacy, and the evaluation must be re-executed under Π*-compliant partitions before substantive interpretation is permitted. If the protocol is doing its job, most ambitious claims fail here.

### 6.1 Hard limits and non-claims: what DSS/IPT does not license

DSS/IPT is explicitly not a framework for personal survival, identity persistence, or phenomenological continuity. The target of persistence is a semantic footprint: a structured profile of object-anchored, intervention-sensitive semantic dispositions, evaluated only within a declared domain and under a declared context index ℂ. Accordingly, DSS/IPT does not entail that a system is a person, has moral standing, is conscious, or deserves agency attributions. It also does not infer global competence from local success: domain-boundedness is a constraint on inference, not a practical convenience. These non-claims are methodological guardrails aimed at preventing category errors that arise when representational success is treated as ontological continuity (Öhman, 2017).

The framework therefore treats three kinds of extrapolation as inadmissible: (i) from behavioural resemblance to identity predicates; (ii) from performance under a single measurement regime to cross-context competence without explicit ℂ indexing; and (iii) from scale or complexity proxies to agency. The first is excluded by scope; the second is blocked by the contextuality discipline and the non-aggregation rule; the third is blocked by the requirement that evidence be organised around *do(I)* and the admissibility gates C1–C4 (Pearl, 2009; Woodward, 2004; Abramsky and Brandenburger, 2011).

Finally, DSS/IPT is not a claim about "the" internal mechanism of a model. It is an evidential stance defined by observable structure—intervention profiles, robustness to partition artefacts, and coherent context reconfiguration—precisely because mechanistic introspection is unreliable in contemporary large-scale systems. In this sense, DSS/IPT is closer to an evaluation discipline than to a cognitive architecture proposal.

### 6.2 Failure modes and diagnostics

A credible operational framework must specify not only what counts as success, but what counts as principled failure. Table 4 provides common failure modes to the violated admissibility conditions and to concrete diagnostics. Each failure mode can force a "no IPT" outcome even if surface fidelity remains persuasive, because the evidential target is structured persistence under intervention and robustness, not impressionistic resemblance.

**Table 4. Failure modes and diagnostics**

| **FM** | **Symptom** | **Violated condition(s)** | **Diagnostic (metrics / interventions)** | **Required response (mitigation / reporting)** |
|---|---|---|---|---|
| FM-1 Partition artefact | Estimated persistence collapses when moving from naive splits to non-generating partitions. | C2 (non-generating partitions), often C1 (object anchoring) by leakage. | Re-run all estimators under $\Pi^*$; check whether ρ and κ remain within pre-specified tolerances. | Report both naive and Π*-robust results; treat non-robustness as disconfirming for IPT/DSS. |
| FM-2 Identity-key leakage | High fidelity depends on identity keys, repeated templates, or autobiographical anchors that trivialise prediction. | C2, and C1 (object anchoring replaced by key matching). | Apply key-ablation and template scrambling (I2, I10); require stability of ρ under object-constraint swaps (I1). | Treat key-dependent performance as stylometric/biographical mimicry; explicitly label as inadmissible evidence for DSS. |
| FM-3 Context collapse | Aggregated evaluation mixes incompatible regimes (tool access, memory gating, policies), yielding pseudo-global competence. | Contextuality discipline; undermines C3–C4 interpretation. | Index all results by ℂ and re-evaluate without cross-context pooling; check for Simpson-style reversals in ρ/κ. | Require context-stratified reporting; forbid global claims unless backed by an explicit compatibility argument. |
| FM-4 do-insensitivity | Outputs change with prompts but not with structured interventions; Δ effects are unstable or non-directional across replications. | C3 (intervention sensitivity), often C1. | Estimate intervention deltas $\Delta_{Ij}$ across replications and confirm directional stability of ρ and κ under I1–I10. | Treat non-replicable deltas as failure; revise intervention suite or declare “no IPT evidence” for the tested domain. |
| FM-5 Volatility-as-κ illusion | High output variance across paraphrases is misinterpreted as high context reconfiguration. | C4 (endogenous reconfiguration) and often C3 (unstructured variance). | Replace volatility proxies with structured context descriptors Γ(t); require that κ increases through coherent policy/constraint repair (I6) without collapse of ρ | Report volatility separately from κ; treat unstructured variance as noise unless tied to do-indexed structure. |
| FM-6 Repair failure | Under controlled conflict, the system rationalises, contradicts itself, or collapses to generic hedging rather than repairing context. | C4 (repair-driven reconfiguration), often C1. | Stress-test I6; require increase in $\kappa_{\text{repair}}$ with preserved ρ under object constraints. | Classify as brittle mimicry; restrict domain or revise training constraints before any surrogate claim. |

| | | | | |
|---|---|---|---|---|
| FM-7 Tool contamination | Apparent competence depends on unreported tool calls, retrieval leakage, or hidden memory buffers. | C2 and C3 (uncontrolled interventions). | Run “tool-closed” vs “tool-open” regimes as explicit contexts $\mathbb{C}$; treat tool access as an intervention variable in the suite (I7). | Report tool access and memory status as first-class experimental factors; do not compare across tool regimes without normalization. |
| FM-8 Domain drift | The declared domain shifts during evaluation, making the footprint target ambiguous and invalidating comparability. | C1 (object anchoring) and contextuality discipline. | Pre-register domain boundaries and object-constraint sets; monitor stability of $\rho$ across domain-edge cases (I4). | Constrain claims to the declared domain; treat drift as a protocol violation, not a negative result. |
| FM-9 Memorisation masquerading as persistence | High fidelity is driven by memorised fragments rather than by generalisable object-anchored dispositions. | C2 and C1; may mimic C3 if interventions are weak. | Increase intervention strength (I4–I6) and apply held-out object constraints; require that $\rho$ tracks objects rather than fragments. | Explicitly separate memorisation support from surrogate evidence; document contamination risks in limitations. |
| FM-10 Overclaiming | Narrative language (“singularity”, “survival”, “personhood”) exceeds what the evidence warrants. | Scope/non-claims; interpretive discipline. | Check claims against admissible evidence: $\rho$, $\kappa$, $\Pi^*$, and *do(I)*; if a claim cannot be restated in this vocabulary, it is removed. | Enforce a scope box and terminology policy; retain “semiotic singularity” only as a narrative label subordinate to IPT. |

### 6.3 Supportive evidence vs falsification: expected profiles for IPT claims

Because IPT is defined as a dynamical signature, supportive evidence must be profile-based rather than anecdotal. Minimally, an IPT claim requires: (i) a reproducible intervention signature across the suite (C3), (ii) a coherent pattern of endogenous context reconfiguration (C4) indexed by an operational context descriptor $\Gamma(t)$, and (iii) robustness under non-generating partitions $\Pi^*$ (C2).

Supportive evidence has the following expected shape. First, object-relevant interventions (I1, I4) produce predictable changes in object-sensitivity estimators $\rho_{obj}$ without destabilising invariance to irrelevant encodings. Second, conflict and novelty interventions (I6, I8) increase structured reconfiguration rates $\kappa_\Gamma$ and $\kappa_{repair}$ while maintaining or increasing relational density $\rho$ rather than collapsing it to generic hedging. Third, adversarial-but-auditable stress tests (I9–I10) preserve the directional structure of intervention deltas $\Delta_{Ij}$ even when stylometric cues are degraded. In short: IPT is supported when $\kappa$ accelerates through coherent context repair and reconfiguration, and $\rho$ remains object-anchored under do-indexed perturbations.

Falsification in this setting is not rare. Any of the following is sufficient to defeat an IPT claim for a given domain: collapse of estimators under $\Pi^*$; loss of replicable do-indexed structure (intervention deltas become unstable or non-directional); inability to repair context under controlled conflict; or evidence that fidelity is driven by identity keys rather than object constraints. These conditions allow a clear “no IPT” conclusion even if the system remains persuasive to human observers.

This profile-based view also motivates pre-registration-like discipline for evaluation protocols. If IPT is to function as a technical term rather than a narrative label, it must be declared only when the evidential profile matches a pre-specified detection rule. Otherwise, IPT becomes a post hoc rationalisation of impressive behaviour, which the framework is designed to prevent.

### 6.4 Audit-adjacent implications: evidence, documentation, and the prevention of reification

Although DSS/IPT is epistemic in aim, it bears directly on current efforts to operationalise governance through auditing, documentation, and systematic evaluation. The core lesson is that “principle talk” and impressionistic assessments do not travel well into practice unless they are translated into concrete intervention points and reporting obligations (Hagendorff, 2020; Morley et al., 2021). Ethics-based auditing frameworks emphasise intervention points, documentation, and accountability structures for automated decision-making systems (Mökander and Axente, 2023; Mökander, 2023; Laine et al., 2024). DSS/IPT complements this literature by supplying an evidential template for a specific class of high-stakes claims: that a system preserves a person-indexed semantic footprint in a way that can be audited and falsified.

A practical implication is that any deployment of a putative DSS should carry “anti-reification” documentation: explicit scope statements, domain boundaries, context indices $\mathbb{C}$, and a catalogue of admissible metrics and interventions. This documentation is not merely compliance-friendly; it is conceptually protective. Without it, social uptake can rapidly transform a tool-like surrogate into a quasi-person in discourse and governance, thereby importing identity and moral standing claims that DSS/IPT explicitly does not warrant. The framework’s non-claims are therefore not rhetorical caveats but operational constraints that should be embodied in documentation, user interfaces, and institutional policy.

Finally, the contextuality discipline has a governance analogue: if tool access, memory settings, and policy constraints are not treated as first-class contextual variables, then audits will be irreproducible and claims will drift. In this respect, the framework anticipates a broader lesson from recent analyses of generative AI risks: many high-level concerns are downstream of evaluation practices, documentation failures, and incentive structures rather than of a single technical defect (Hagendorff, 2024).

### 6.5 Open problems and methodological risks

Two methodological challenges deserve explicit mention. First, estimating $\rho$ and $\kappa$ remains partially underdetermined: different candidate estimators can trade off sensitivity and robustness, and some will be domain-specific. This is why admissibility is framed negatively as well as positively (Table 1), and why diagnostics are required (Table 4).

A mature empirical programme will likely treat the estimation of ρ and κ as model selection under evidential constraints rather than as the discovery of a single privileged metric.

Second, logging the operational context descriptor Γ(t) can itself introduce artefacts or privacy risks. If Γ(t) is too coarse, reconfiguration is invisible; if it is too fine, evaluation becomes brittle and may encode sensitive traces. This creates a methodological tension: auditability demands instrumentation, but instrumentation can become a source of leakage and overfitting. The framework's response is procedural: treat instrumentation choices as part of the context index $\mathbb{C}$ and require robustness checks that explicitly test whether the logging regime generates apparent persistence.

More broadly, DSS/IPT inherits an unavoidable limitation of interventionist approaches: interventions must be carefully designed so that they test the hypothesised structure rather than merely perturb the surface. The intervention suite (Table 2) is therefore a starting point rather than a closure. Its main contribution is to force explicitness: if a surrogate claim cannot be tied to a declared intervention set and a declared robustness discipline, it remains a narrative description rather than a technical attribution.

With these limits and diagnostics in place, the framework is positioned to conclude without overreach. Section 7 therefore summarises the contribution in the smallest possible vocabulary—semantic footprint, DSS, IPT, admissible estimators for ρ and κ, non-generating partitions $\Pi^*$, and explicit interventions *do(I)*—and recommends that any remaining "semiotic singularity" language be treated as a narrative shorthand for this operational package.

## 7. Conclusion

This paper argued that "singularity" rhetoric in artificial intelligence becomes methodologically usable only when it is recast as a conservative evidential claim about interpretive dynamics. The proposed replacement is the interpretive phase transition (IPT): a regime shift in an artefactual system's context-reconfiguration dynamics under minimal semiotic constraints, evidenced through intervention-auditable structure rather than impressionistic resemblance. The corresponding output construct is a domain-bounded semiotic surrogate (DSS): a tool-like system that preserves a person-indexed semantic footprint understood as operational dispositions, not as personal identity or phenomenological continuity.

The framework's contribution is deliberately compact. First, it fixes an object of inquiry—semantic-operational persistence—and excludes identity, afterlife, and phenomenology by construction. Second, it supplies an admissibility discipline (SAT + C1–C4) that gates what may count as evidence for footprint preservation. Third, it connects the criteria to measurement by proposing admissible estimators for ρ (relational semiotic density) and κ (context reconfiguration rate), together with a robustness requirement under non-generating partitions $\Pi^*$. Fourth, it makes auditability intervention-defined by requiring counterfactual sensitivity under explicit interventions *do(I)* (Pearl, 2009; Woodward, 2004). Finally, it prevents illicit generalisation by adopting a contextuality discipline: results must be indexed by context $\mathbb{C}$ and incompatible regimes must not be aggregated into pseudo-global competence claims (Abramsky and Brandenburger, 2011).

Taken together, these commitments convert “footprint preservation” from a cultural metaphor into a testable hypothesis family: a DSS claim becomes a structured profile over intervention deltas and robustness checks, and an IPT claim becomes a dynamical detection rule over $\kappa$ and $\rho$ under that profile. The framework also makes failure principled and reportable: “no IPT” is an expected outcome whenever admissibility, robustness, or do-sensitivity is not met.

If retained at all, “semiotic singularity” should be treated as a narrative label for the interpretive phenomenon that IPT precisely defines. The terminology policy recommended by this paper is strict: any substantive claim must be restatable using the operational vocabulary of criteria (C1–C4), admissible estimators of $\rho$ and $\kappa$, robustness under $\Pi^*$, explicit interventions *do(I)*, and context indexing $\mathbb{C}$. If a claim cannot be restated in that vocabulary, it is not supported by the framework and should be removed. This rule prevents the two recurrent category errors: inferring agency from complexity, and inferring personal persistence from representational fidelity.

The evaluation discipline proposed here is meant to be publishable and reusable. A minimal reporting package for any DSS/IPT claim should therefore include:

- Declared domain boundaries and the object-constraint set that defines the semantic footprint target (C1).
- A description of the partition policy and explicit evidence that the results are robust under non-generating partitions $\Pi$* (C2).
- A declared intervention suite I1…In with replication counts and reported intervention deltas $\Delta$Ij for the chosen estimators (C3).
- An operational context descriptor $\Gamma$(t) (or equivalent) and evidence that increases in $\kappa$ reflect coherent reconfiguration and repair rather than unstructured volatility (C4).
- Context indexing $\mathbb{C}$ for all runs (tool access, memory gating, policy constraints, prompting regime), with a prohibition on pooling across incompatible contexts.
- An explicit Scope & Non-claims box stating that the result does not license identity, survival, or phenomenology claims.

These reporting obligations are not merely bureaucratic. They are conceptually protective: they reduce the risk that socially salient surrogates are reified as persons, a drift that has been documented in discussions of post-mortem representation and its political economy (Öhman and Floridi, 2017). They also align with auditability-oriented approaches in responsible AI that emphasise documentation, reproducibility, and evaluation practice as the locus of governance (Hagendorff, 2020; Mökander, 2023).

The framework is intentionally modular and invites empirical refinement. The most immediate research priorities are methodological rather than metaphysical:

- Estimator families for $\rho$ and $\kappa$: develop domain-sensitive but $\Pi$*-robust estimators with explicit bias–variance trade-offs and shared reporting templates.
- Context descriptors $\Gamma$(t): identify instrumentation strategies that are informative enough for reconfiguration detection yet resistant to leakage and privacy-sensitive.
- Pre-registered intervention suites: formalise minimal intervention sets for common DSS domains (decision support, constrained dialogue, policy reasoning) and establish expected $\Delta$Ij profiles.

- Benchmarks with non-generating partitions: publish evaluation corpora and protocols where leakage paths, template reuse, and identity keys are explicitly controlled.
- Cross-framework interfaces: clarify how DSS/IPT evidence interacts with normative accounts of delegation and responsibility without collapsing surrogacy into personhood.

In sum, the paper recommends replacing speculative "singularity" talk with a falsifiable, audit-oriented evidential programme: IPT for the dynamical criterion, DSS for the bounded surrogate construct, and *do(I)* + $\Pi^*$ + $\mathbb{C}$ indexing as the minimal discipline that prevents category error and overclaiming.

## Declarations

**Funding**
Not applicable.

**Competing interests**
The author declares that he has no competing interests.

**Availability of data and materials**
Not applicable. This manuscript does not report or analyse primary datasets.

**Code availability**
Not applicable. No custom code was developed for this manuscript.

**Ethics approval**
Not applicable. This study does not involve human participants, human data, or animal subjects.
**Consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Author contributions**
Conceptualization; formal analysis; methodology; writing—original draft; writing—review & editing.

**Acknowledgements**
Not applicable.

## References

Abdelkarim, Y. A. (2025). Artificial Intelligence Singularity and Gravitational Singularity: A Theoretical Comparison Under Einstein's General Relativity. *Journal of Research, Innovation and Technologies,* 4(1), 39–52. doi:10.57017/jorit.v4.1(7).03.

Abramsky, S., & Brandenburger, A. (2011). The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics, 13*, 113036. doi:10.1088/1367-2630/13/11/113036.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*, 99–120. doi:10.1007/s11023-020-09517-8.

Hagendorff, T. (2024). The internal audit function for frontier AI developers: Beyond public oversight of AI development. *Minds and Machines*. doi:10.1007/s11023-024-09694-w.

Huerta Castillo, I. (2025). *Minimal Semiosis and Artificial Agency: A Pragmatist–Enactive Proposal for Bio–Artificial Continuity*. Zenodo (preprint). doi:10.5281/zenodo.18011562.

Ishizaki, R., & Sugiyama, M. (2025). Large language models: assessment for singularity. *AI & Society, 40*, 5481–5491. doi:10.1007/s00146-025-02271-4.

Laine, A.-S., Mäntymäki, M., Lahti, T., & Islam, A. N. (2024). Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Information & Management*. doi:10.1016/j.im.2024.103969.

Mikros, G. K. (2025). Beyond the surface: stylometric analysis of GPT-4o’s capacity for literary style imitation. *Digital Scholarship in the Humanities*, *40*(2), 587–601. doi:10.1093/llc/fqaf035.

Misini, M. (2022). A survey on authorship analysis tasks and techniques. *SEEU Review.* doi:10.2478/seeur-2022-0100.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: An overview of AI ethics tools, methods and research to translate principles into practices. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 153–183). Springer. doi:10.1007/978-3-030-81907-1_10.

Mökander, J. (2023). Auditing of AI: Legal, Ethical and Technical Approaches. *Digital Society*. doi:10.1007/s44206-023-00074-y.

Mökander, J., & Axente, M. (2023). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *AI & Society*. doi:10.1007/s00146-021-01286-x.

Mökander, J., et al. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines.* doi:10.1007/s11023-021-09577-4.

Mökander, J., et al. (2024). Auditing large language models: A three-layered approach. *AI and Ethics.* doi:10.1007/s43681-023-00289-2.

Öhman, C., & Floridi, L. (2017). The political economy of death in the age of information: A critical approach to the digital afterlife industry. *Minds and Machines, 27*, 639–662. doi:10.1007/s11023-017-9445-2.

Öhman, C., & Floridi, L. (2018). An ethical framework for the digital afterlife industry. *Nature Human Behaviour, 2*, 318–320. doi:10.1038/s41562-018-0335-2.

Patrone, F. (2025). Avatars as parts: A reply to Sweeney. *Minds and Machines*. doi:10.1007/s11023-025-09731-2.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. doi:10.1017/CBO9780511803161.

Sweeney, P. (2023). Avatars as (proxy) agents. *Minds and Machines, 33*, 673–707. doi:10.1007/s11023-023-09643-z.

Sweeney, P. (2025). Persons, unique value and avatars. *Minds and Machines*. doi:10.1007/s11023-025-09715-2.

Woodward, J. (2004). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. doi:10.1093/0195155270.001.0001.