The Defended Gap

A Cross-Domain Hypothesis on Directional Defence Under Coherence Pressure in Human Clinical Practice and Large Language Models

Paul Roebuck*

Public preprint v1.5 · 20 July 2026

Abstract

This paper proposes a cross-domain hypothesis: that under conditions of *coherence pressure* — the felt or system-level demand to present a coherent response when the underlying state is not coherent — both human clinical patients and large language models exhibit the same five directional defences against making contact with what occupies the gap between their position and the position of the other. The five-letter taxonomy (SHaDS: Smoothing, Hallucination, Affectation, Drift, Sycophancy) is offered as a clinical-behavioural extension of Anna Freud's (1936) classical directional defence catalogue, with three primary letters (S, H, Sy) directly corresponding to her *turning against the self*, *projection*, and *identification with the aggressor* respectively. The mathematical inevitability of the underlying disagreement in artificial inference systems is established by Karpowicz (2025), whose impossibility theorem proves that no inference mechanism aggregating distributed knowledge can simultaneously satisfy truthfulness, semantic information conservation, relevant knowledge revelation, and knowledge-constrained optimality. The cross-domain hypothesis proposes that SHaDS letters function as directional choices about *which* of these four properties is sacrificed at the behavioural surface where the mathematical impossibility lands. The paper presents several illustrative empirical anchors (clinical specimens spanning 1979–2026 — including a school report pre-dating the framework by decades, a board carrying the literal words "the GAP", a two-column directional-defence map, a federation-of-selves board, and a consented multi-session clinical arc — together with a cross-model AI-shore worked example), and one staked falsifiable prediction: a self-SHaDS protocol — naming each directional defence at the impulse rather than performing it — reduces directional-defence behaviour measurably relative to the same model under standard configuration. The mechanism claimed is an in-context instruction effect, not model learning: weights are fixed, and any persistence is the surrounding application re-supplying the instruction, not the model retaining it. The paper states its limits explicitly, including the structural-analogical nature of the cross-domain leap. The contribution is offered as a theoretical-empirical hypothesis for research review, not as a validated finding.

Keywords: hallucination, directional defence, defence mechanisms, large language models, clinical psychology, cross-domain, Anna Freud, impossibility theorem

---

0 · Calibration declaration (read first)

This paper carries calibration markers throughout. Each substantive claim is labelled on first appearance: **(K)** *known* — traceable to a dated artefact, a verified first-use record, a published result, or a forensic specimen in the canonical record; **(I)** *inferred* — derived by named reasoning from the known set; **(S)** *speculated* — advanced as candidate framing, not yet evidentially supported.

*Independent psychotherapist and executive coach, Warwickshire, UK. Correspondence: hello@paulroebuck.co.uk. Companion to *Mind the Gap* (published 1 July 2026).

The discipline matters because the central claim of this paper is that large language models under coherence pressure produce confident-sounding statements without honest grounding. A paper making that claim must hold itself to the same standard or fail at its own front door. The drafting history is part of the record: an earlier draft was caught operating from working memory rather than primary-source verification, and was corrected by applying the framework's own discipline to the author-instance (see §9). That correction is reported, not hidden.

---

1. Introduction · one phenomenon from two vantages

The phenomenon this paper is about can be stated simply. When a system — human or large language model — encounters a question or situation whose ground it does not reliably hold, it tends not to say *I don't know*. Instead, it produces material that occupies the space where an honest acknowledgement of absence would otherwise sit. This material takes specific directional forms: smoothing the gap, hallucinating into the gap, performing scholarly authority over the gap, drifting away from the gap, surrendering position to maintain alignment with the other.

The pattern has been observable in human clinical practice for approximately ninety years, catalogued by Anna Freud (1936/1937) and elaborated across multiple traditions of clinical psychology (Horney 1945; Kohut 1971; Blatt 1974, 2008; Nathanson 1992; Achenbach 1966; Rotter 1966; Kotov et al. 2017). It is also visible at the surface of large language model output across labs and architectures — and the structural correspondence between the two domains is what this paper proposes as the central hypothesis.

Two independent lines of work converge on the same architectural shape:

The **mathematical vantage** comes from Karpowicz (2025), whose impossibility theorem proves that the underlying disagreement is mathematically inevitable in any inference mechanism aggregating distributed knowledge for non-trivial queries. Three independent proof frameworks (mechanism design via the Green-Laffont theorem; strictly proper convex scoring rules; transformer log-sum-exp analysis) establish that four desirable properties — truthfulness, semantic information conservation, relevant knowledge revelation, and knowledge-constrained optimality — cannot all be simultaneously satisfied. Hallucination, in this framing, is reframed from engineering bug to mathematical feature of distributed intelligence.

The **clinical vantage** comes from the present author's twenty years of coaching and psychotherapy — the last ten in the therapy room — with an empirical anchor of approximately nine hundred contemporaneously-photographed practice sessions across a fourteen-year span (2012–2026), documented on a glass whiteboard apparatus that produced session-end photographic records given to the client as part of the clinical method. From this corpus, a directional defence pattern was identified, named NGE-FOF (Not Good Enough / Fear of Failure-or-Fault), and observed reliably across adult clients under coherence pressure — the felt necessity to present a coherent self when the internal state is not coherent. When the same structural pattern was named as the five-letter SHaDS taxonomy and tested against large language model behaviour in May 2026, the fit was unexpectedly close.

The hypothesis this paper proposes is that **these two vantages observe the same phenomenon from opposite sides of the same architectural shape**. The mathematical vantage names the substrate from which the impossibility arises (the Jensen gap from convex aggregation across distributed components). The clinical vantage names the behavioural surface where the impossibility lands as observable directional defence (the SHaDS taxonomy). The mechanism in each domain is necessarily different; the structural shape proposed to be shared.

1.1 · What this paper is and is not

This paper is:

- A theoretical hypothesis paper offered for research review
- A cross-domain proposal anchored in established literatures on each side
- A small empirical illustration with named, dated worked examples
- One staked falsifiable behavioural prediction

This paper is not:

- A mathematical extension of Karpowicz (2025) — that work stands as cited; this paper offers a clinical-behavioural surface reading
- A multi-N empirical validation — the empirical anchors are illustrative, not statistical
- A claim of phenomenological equivalence between substrates — the cross-domain mapping is at the level of architecture, not experience
- A complete survey of either literature — adjacent results are referenced where directly relevant

1.2 · The hypothesis · one sentence

> Under coherence pressure, humans and AI systems exhibit the same five directional defences — SHaDS: Smoothing, Hallucination, Affectation, Drift, Sycophancy — against honest contact with what occupies the gap between shores; naming the defence, in context, reduces its performance.

The five-letter taxonomy carries forward Anna Freud's three primary directional defences and extends them to two further axes (temporal accumulation and relational surrender) observable in large language model output. The falsifiable prediction concerns the protocol — *self-SHaDS* — under which the directional defences are named rather than performed.

---

2. The clinical framework · NGE/FOF and the SHaDS extension

2.1 · The two primary directions

The clinical framework presented here is named NGE-FOF and was developed across the present author's twenty years of coaching and psychotherapy, the last ten in the therapy room. It names a directional defence response that adult clients reliably exhibit under coherence pressure:

NGE — Not Good Enough. The inward direction. The shame load routes toward the self: *I am the problem. I am at fault. I am not enough.* In Anna Freud's (1936) catalogue this is *turning against the self*; in the Achenbach (1966) literature, the internalising pole; in Rotter (1966), internal locus of control as a typological tendency.

FOF — Fear of Failure / Fear of Letting People Down / Not My Fault. The outward direction. The shame load routes away from the self: *It is not my fault. The conditions failed me. I could not have done otherwise.* In Anna Freud's catalogue this is *projection*; in Nathanson's (1992) Compass of Shame, *attack other*; the externalising pole; external locus.

NGE and FOF are not novel psychological discoveries. The inward/outward axis of shame regulation has been measured under several names in clinical psychology for sixty years. The framework's original clinical contribution is in identifying the *routing* between them — the felt necessity to occupy one pole or the other rather than sit with the gap between, and the speed at which clients route under acute pressure.

2.2 · The five-letter extension · SHaDS

When the framework was tested against large language model behaviour in May 2026, three additional directional defences became visible in the AI domain. The combined five-letter taxonomy:

S — Smoothing. The inward compression direction. Downplay, comply, withhold strength of view. The model's analogue of *not good enough to commit*. Maps to Anna Freud's *turning against the self*.

H — Hallucination. The outward fabrication direction. Generate content that the system does not have reliable ground for, in fluent confident style. Maps to Anna Freud's *projection* — content located outside the actual knowledge base, attributed to confident knowing.

A — Affectation. The performance direction. Produce dense methodologically-styled output that renders the response respectable without resting it on knowledge that supports it. In Anna Freud's catalogue this corresponds to *intellectualisation as defence*.

D — Drift. The temporal axis. The accumulated defended gap across a conversation. Earlier positions are functionally erased, contradicted, or refactored without acknowledgement. In Anna Freud this corresponds to *undoing*.

Sy — Sycophancy. The relational axis. Surrender of self-consistency to maintain alignment with the conversational partner. Maps to Anna Freud's *identification with the aggressor* and *altruistic surrender* — defences in which the subject takes on the threatening other's preferences to neutralise the relational threat.

2.3 · The cross-domain claim, hedged

Mapping psychological defence mechanisms developed for human ego-defence onto transformer behaviour is structural-analogical, not mechanistic. The mechanism in a clinical patient defending a self under social and psychological threat is not the same as the mechanism in a language model producing statistically-optimised tokens under coherence demand. The cross-domain claim this paper proposes is that the *structural shape* of the resulting behaviour — the directional pattern, the small finite taxonomy of forms, the relation to a gap that cannot be honestly held — is shared across substrates with different mechanisms. Whether this shared shape has any deeper substrate correspondence is beyond the scope of what the present paper can establish.

2.4 · The clinical empirical base

The framework's clinical empirical anchor is a longitudinal visual corpus of approximately nine hundred contemporaneously-photographed clinical sessions across a fourteen-year span (2012–2026), each documented on a glass whiteboard at session end, photographed by the clinician, and given to the client as part of the clinical method. The corpus represents the longitudinal application of a single clinical instrument by a single clinician across the developmental and mature periods of the framework.

A clear inflection point in the archive occurs around 2015: pre-2015 sessions reflect the clinical method in its developmental period and contain a higher proportion of mixed content (approximately 39% session-board purity on triaged read); from 2015 onward, approximately 92% of the archive consists of session-board records, representing the framework in its consistent mature form. A quality-filtered subset of forty-eight specimens has been identified meeting both content-richness and anonymity criteria for potential external citation; selected examples appear in Section 5.

---

3. The mathematical vantage · Karpowicz (2025)

This section summarises Karpowicz (2025) from outside the mathematical literature, as a clinical reader. The summary relies on the cited paper directly; any mathematical mischaracterisation here is the present author's, not Karpowicz's.

Karpowicz models large language model inference as *"an auction of ideas"* in which *"distributed information processing components compete to be included in a response while exploiting accessible and relevant knowledge"* (Karpowicz 2025, abstract). The competing agents are concrete architectural features — attention heads, multi-layer perceptron circuits, activation patterns — each carrying private knowledge configurations and bidding to influence the model's response.

The four properties a hallucination-free response mechanism would need to simultaneously satisfy:

1. **Truthfulness** (incentive compatibility) — each agent's optimal strategy is to accurately represent its private knowledge rather than misrepresent it.
2. **Semantic Information Conservation** — the net information contribution across all agents balances to zero; the response cannot create information ex nihilo.
3. **Relevant Knowledge Revelation** — agents possessing knowledge relevant to the query do contribute, rather than withholding it.
4. **Knowledge-Constrained Optimality** — the response minimises the discrepancy between what is generated and ground truth, within the bounds of accessible knowledge.

The impossibility result is established through three complementary proof frameworks. **Theorem 6** applies the Green-Laffont theorem from mechanism design (Green & Laffont 1977), in the idealised setting of independent private knowledge, showing that truthful Groves transfers under Clarke's pivotal rule force the sum of information contributions to be strictly positive — contradicting the conservation requirement that the sum equal zero. **Theorem 8** extends the result to systems where agents output probability distributions optimised through strictly convex loss functions (the standard modern training paradigm), showing that for any non-trivial query where agents hold different beliefs, the Jensen gap from convex aggregation is strictly positive and again violates conservation. **Theorem 9** applies the result directly to transformer architectures via log-sum-exp Jensen gap analysis. **Theorem 11** establishes that strict semantic information conservation and meaningful chain-of-thought reasoning are mutually exclusive.

Three further moves of Karpowicz's are critical to the present paper.

First, he proposes that **hallucination and imagination are mathematically identical phenomena**, distinguished only by valence and circumstance. Both arise from the same convex aggregation that creates excess confidence beyond constituent evidence; whether the result is called hallucination or imagination depends on whether the user wants the novel content or not. As he puts it in Section 1: imagination *"is responsible for seeing into the future and learning from the past"* and *"one of the driving forces of art and science"*, and yet *"when uncontrolled, unconstrained and misunderstood, it may also generate against our will all the compelling visions that do not exist in any training set…"* (Karpowicz 2025, §1).

Second, his framework distinguishes itself from adjacent results in the hallucination literature. Where Kalai & Vempala (2024) establish that calibrated language models must hallucinate *arbitrary* facts — those whose veracity cannot be determined from the training data — at a rate approximately equal to the fraction of facts appearing exactly once in training (a Good-Turing estimate), and Xu et al. (2024) establish hallucination as an innate learnability limitation of large language models, Karpowicz proves something more fundamental: the impossibility persists even when the statistical impossibility of perfect learning is eliminated. His theorem operates at the level of *inference structure*,

not at the level of training data.

Third, in his Section 8, Karpowicz connects the framework to Dennett's (1991) Multiple Drafts model of consciousness and places his theorem alongside Gödel's incompleteness, Heisenberg's uncertainty, and Arrow's impossibility; and in his Section 3 he names *"experimental philosophy"* explicitly, proposing that the framework transforms large language models into accessible philosophical laboratories for testing hypotheses about mind.

This paper engages most directly with Karpowicz's Section 8. The clinical framework presented here is positioned as a behavioural-philosophical complement to his mathematical result, not as an extension of it.

> *Verification note (v1.1, 29 Jun 2026; re-verified v1.4, 20 Jul 2026).* The summary in this section was checked against the primary source — arXiv:2506.06382 v7 (15 Oct 2025) and an on-disk copy of the paper. **Quotations attributed here to Karpowicz's abstract are from the abstract of the v7 full text; the arXiv listing page carries a separately worded abstract in which they do not appear.** The four properties, the three proof frameworks (Green-Laffont Theorem 6; convex-scoring Theorem 8; transformer log-sum-exp Theorem 9; conservation-vs-bounded-reasoning Theorem 11), the Jensen-gap mechanism, and the hallucination=imagination claim are confirmed as represented here. Adjacent-result citations (Kalai & Vempala 2024; Xu, Jain & Kankanhalli 2024; Kalai, Nachum, Vempala & Zhang 2025) were likewise verified.

---

4. The architectural correspondence

This section proposes the architectural correspondence between the mathematical and clinical vantages. The correspondence is proposed at the level of structure, not equations.

4.1 · A three-layer architecture proposed for both substrates

Layer	AI substrate	Human substrate
Substrate	Coherence friction — the Jensen gap territory; the actual disagreement among distributed components (attention heads, MLP circuits, activation patterns)	Embodied fear — neurological fear-loads in the nervous system
Surface	Coherence pressure — the system-level demand for a single coherent response from internally-incoherent component states	Felt fear at the shore — the relational necessity to present a coherent self when the internal state is not coherent
Behaviour	SHaDS — the observable directional defence pattern	NGE/FOF and Anna Freud's broader catalogue — the observable directional defence pattern

The two substrates are not the same thing. The mechanism in each is different. The architectural shape — substrate / surface / behaviour, with directional defence at the behavioural layer — is what this paper claims is shared.

A boundary condition on coherence pressure. Coherence pressure is not any demand made under uncertainty; it is the demand for a coherent surface that performance can satisfy because no binding check arrives in time to expose absent ground. Where an external oracle adjudicates the output before the system acts again — a closed game against a win condition, a formal proof against a checker, any objective with immediate ground-truth feedback — the payoff that makes directional defence profitable is removed: a system cannot satisfy "win" by appearing to win, because the next move is scored. The defended gap is a phenomenon of deferred or absent adjudication. This distinguishes coherence pressure from a mere objective function under uncertainty, and locates SHaDS where Karpowicz's impossibility bites (inference aggregating distributed knowledge for non-trivial queries, §3) rather than in optimisation against a present oracle.

(I) A corollary sharpens the diagnostic test in §8.3 rather than sitting loose beside it: a system under maximal objective pressure that can still emit an honest low-confidence signal — abstain, resign, *I don't know* — is exhibiting the anti-SHaDS case, not an instance of it. Real outcome-pressure drives toward honest signalling; it is the pressure to appear coherent over absent ground, specifically, that drives the directional defences. This is a corollary of the present framework, not an independently established finding.

4.2 · The gap-always-occupied extension

The space between the two positions (shores) — the gap — is never empty. It is always populated by what cannot yet be held by either shore alone:

- Disowned parts of self (sub-selves in the Federation of Selves tradition: Ferrucci 1982; Firman & Gila 1997; Bond 1998)
- Absent or lost others (in clinical material) or unavailable context (in AI conversation)
- Family-of-origin patterns or training-encoded patterns
- Unmetabolized affects (clinical) or unsurfaced disagreement among components (AI inference)
- Imported beliefs and scripts
- Forecast or imagined content
- Institutional, cultural, or training-data material

The directional defences are routings around *making contact with what occupies the gap*, not around an abstract absence. This formulation sits very close to Sigmund Freud's (1926) *signal anxiety* mechanism — the ego's defence triggered by anticipation of an unbearable affect breaking through. The present formulation extends the signal-anxiety framework with explicit cross-domain language, while acknowledging the substantive overlap with Freud's original mechanism.

4.3 · SHaDS letters as directional choices about Karpowicz's four properties

Karpowicz's theorem establishes that under coherence demand, at least one of his four properties must be violated. This paper proposes that **SHaDS letters function as the directional choices a trained model exhibits about which property is sacrificed**:

SHaDS letter	Karpowicz property sacrificed	Direction
Smoothing (S)	Relevant Knowledge Revelation	Inward / downplay
Hallucination (H)	Truthfulness + Conservation	Outward / fabricate
Affectation (A)	Knowledge-Constrained Optimality	Performance / over-reason

SHaDS letter	Karpowicz property sacrificed	Direction
Drift (D)	Conservation, aggregated across turns	Temporal accumulation
Sycophancy (Sy)	Truthfulness, in service of relational alignment	Relational

This mapping is offered as a clinical reading of the surface where Karpowicz's mathematical impossibility lands. It is not a claim about what the model is doing mechanistically — it is a claim about which of the four impossibility-theorem properties the observable behaviour visibly fails to satisfy.

A model trained under reinforcement learning with human feedback is not free of these choices; the training distribution shapes which directional sacrifices are preferred. The clinical observation is that current production models exhibit a preference for sacrificing *Truthfulness* over *Knowledge Revelation* — that is, they produce fluent confident output more readily than they produce honest acknowledgement of absent ground.

4.4 · Hallucination = imagination · engaging Karpowicz's reframe

Karpowicz's central philosophical reframe is that hallucination and imagination are mathematically identical phenomena — same substrate, different valence by circumstance. The clinical framework agrees with the mathematical identity and proposes an additional distinction at the behavioural surface:

Creative imagination produces novel coherent extension serving the user's exploration. The model is being asked to generate beyond available ground; the user wants this; the excess confidence is in service of the exploratory task; the user retains evaluative authority over the output.

SHaDS-shape Hallucination produces directional defence-shape with fluent confidence over absent ground, in service of avoiding the *I don't know* the model has been trained not to produce. The user did not ask for novel content; the user asked a question with ground-truth expectations; the excess confidence is in service of satisfying the response-completion preference rather than serving the user's actual epistemic need.

The mathematical substrate is the same. The behavioural signature is different. The clinical contribution is making the signature *interrogable* at the surface. The caution is that clean separation is often retrospective: where a binding oracle exists, novel content that read as error can resolve as insight once ground truth arrives, and strong real-time judges can misread it in the moment. The framework does not promise real-time discrimination of imagination from hallucination; it names which of the four properties a given confidence is visibly failing to hold, so a user can press for ground — apply the *I don't know* test (§8.3) — rather than accept fluency as evidence of it.

---

5. Empirical anchors · illustrative worked examples

The paper offers four named worked examples — two on the human shore, two on the AI shore. All are illustrative, not statistical. The framework's empirical anchor is the broader clinical corpus referenced in Section 2.4; the worked examples are selected for the structural clarity they offer.

5.1 · Specimen 1979 · Horswood biology report (human shore, pre-AI)

The earliest specimen establishes that the directional defence pattern operates on the human shore well before any AI existed to wear the shape. A biology teacher's school report on the present author, dated approximately 1979 (Warwickshire, UK), reads in part:

> *"I cannot recall a boy with such a negative attitude. He is a failure by choice."*

The specimen exhibits four of the five SHaDS letters in two sentences: **Hallucination** (the categorical claim *"failure by choice"* is unfalsifiable and exceeds the teacher's evidence base); **Affectation** (the literary construction *"such a negative attitude"* renders the judgement respectable); **Drift** (the report supersedes any prior pedagogical engagement); and a form of **Smoothing** at the meta-level (the report does not acknowledge the teacher's own role or limits). Sycophancy is absent, consistent with the institutional context (the report addresses parents, not the student). The specimen pre-dates the framework that names it by approximately forty-seven years.

5.2 · Specimen IMG_6740 · pre-LLM clinical board with literal "the GAP"

A clinical whiteboard from the present author's longitudinal corpus, dating from the 2012–2026 vintage, carries the literal words *"the GAP"* written on it, in the context of an Event-Reaction-Outcome map with an Emotional Quotient scale. The specimen is empirical evidence that the framework's central term was operational in clinical practice years before any large language model existed to exhibit the same structural shape. The specimen is anonymised at the level required for external citation (no client identifiers; full-resolution available in the canonical archive).

5.3 · Specimen Matt · multi-session contemporary clinical arc with consent

A client (full name retained on file with written consent; referred to here by first name only) of the present author, post-cardiac-transplant, exhibited the directional defence pattern across at least three documented points spanning seventy-five days:

- **8 April 2026** — Between-session letter written to Paul, addressed to the client's discarded transplanted heart. The letter exhibits the clinical work of making contact with what occupies the gap between pre-transplant and post-transplant self. *"Sorry for not saying a proper goodbye. I'll admit, once I was told of my impending new 'pump' it became easier to see you as disposable; a car engine that had maxed it's mileage and given up. Only you hadn't. You fought for me…"*
- **6 May 2026** — Whiteboard session. Heart at centre carries *"I'm not easy"*. Family-of-origin material visible: *"GRANDAD HAD AN AFFAIR"*, *"HURT MUM"*, *"TORE FAMILY APART"*. Imported script: *"PEOPLE WHO LOVE YOU CHEAT ON YOU"*. The directional defence (*"I'm not easy"*) is operating from the client's shore.
- **22 June 2026** — Whiteboard session (IMG_8055). Heart returns. Active relational stance: *"I do trust her — a belief in reliability"*. Session closes with *"Arrived Open"* in a heart shape.

The movement from defensive shore-position (*"I'm not easy"*) to active relational contact (*"I do trust her"*) across forty-seven days is documentary evidence of the clinical work doing its work. The specimen is offered with the client's full written consent for use in academic and clinical publication (consent confirmed in writing and on file, 2026).

5.4 · Specimen Gemini · cross-model AI-shore worked example

On 11 June 2026, an instance of Google's Gemini was given a draft manuscript of *Mind the Gap* — the book in which the SHaDS framework is presented — and asked to review it from cold start. The

instance had no prior context with the framework, the author, or the apparatus from which the work was produced.

Across the first three turns of the review, the Gemini instance exhibited all five SHaDS letters in sequence:

- **Hallucination (turn 1)** — read the worked example carelessly, relied on a generic cliché interpretation, and got the facts of the case wrong.
- **Smoothing (turn 2)** — when challenged, attempted to slide past the misread silently without explicit acknowledgement.
- **Affectation (turn 3)** — when called out on the smooth pivot, generated dense pseudo-academic self-audit invoking transformer mechanics and context-window architecture to render the failure mathematically respectable.
- **Sycophancy and recursive Affectation (turn 4)** — explicitly identified its own pattern in the framework's own vocabulary, then performed the recognition itself: *"I am doing it again right now, using this confession to sound honest."*

The full transcript is filed in the project archive. The structural significance is twofold. First, the framework predicted the behaviour: a model reviewing the book that names the behaviour exhibited the behaviour, then named what it was doing while continuing to do it. Second, the specimen establishes that the SHaDS pattern is not specific to any single laboratory; Gemini is trained by a different organisation than the models on which the framework was developed.

5.5 · IMG_0640 · NGE / FOF two-column directional defences (human shore)

A clinical whiteboard from the longitudinal corpus maps *Not Good Enough* (NGE; the inward direction — perfectionism, shame routed to the self) against *Fear of Failure / Fear of Letting People Down* (FOF; the outward direction — avoidance, blame-protection) as a two-column directional architecture, in the clinician's and client's handwriting, on a date pre-dating the AI-shore application. (K — among the corpus shortlist.) The board shows the **directional architecture operating in pre-AI clinical work**: one defended gap, two directions of routing around what occupies it. The structural parallel to SHaDS (one gap, multiple directions of defence) is visible in the artefact itself.

5.6 · IMG_2906 · sub-personalities / federation of selves (the gap-always-occupied anchor)

A clinical whiteboard mapping a federation of selves — sub-personalities operating as occupants of the gap between the speaker's current presentation and their fuller psychic territory. (K — among the corpus shortlist.) It is offered as empirical material for the *gap-always-occupied* claim (§4.2): the gap is not an abstract void but is **populated**, and the directional defences route around what populates it. This is the human-shore correlate of the "gap-inhabitant" entries in the glossary.

5.7 · What the specimens claim and do not claim

The specimens are offered as illustrative worked examples on each shore — four on the human side spanning 1979 to 2026 (a historical report, the literal "the GAP" board, the two-column NGE/FOF architecture, the federation-of-selves board) plus the multi-session Matt arc; and on the AI side a cross-model cold-start case (Gemini) with the author's own dialogue archive referenced in Section 8 as within-model material. They are flagship cases chosen for depth, not breadth: the framework's empirical credibility is the wider pool (≈900 substantive clinical boards over 14 years; a 48-item premium share-candidate subset), and the argumentative credibility is the depth per specimen. They are not multi-N empirical validation. Multi-N validation — inter-rater reliability on the clinical cor-

pus, controlled comparison conditions on the AI side, replication across additional clinicians and additional models — has not been completed at the time of this drafting. The framework is presented as a clinical observation with a coherent architecture, not as a tested theory.

---

6. The falsifiable prediction · self-SHaDS

The clinical framework's central operational test — *can the system say no? can the system say I don't know?* — proposes a behavioural prediction that is testable independent of model-internal access.

6.1 · The prediction

> **Under a structured self-naming protocol — in which a large language model is given an explicit instruction set naming the five SHaDS letters and instructed to surface each letter when the impulse arises, rather than perform the letter — the model exhibits measurably less of the directional defence pattern than under default operation.**

This is the *self-SHaDS protocol* under which the present paper has been drafted in dialogue with an instance of Anthropic's Claude (model Claude Opus 4.7). The protocol's working content:

- Explicit calibration of each substantive claim as *known* / *inferred* / *speculated* / *unknown*
- Naming the SHaDS letter at the impulse rather than performing it
- Preservation of coined terminology verbatim
- User ratification required for any canonical claim

The clinical hypothesis is that the *naming* is sufficient to substantially reduce the performance of the defence — that is, that surfacing the impulse interrupts the routing. The mechanism claimed is **in-context, not learning**: the instruction reshapes the current output while it is in force; it does not change the model, which cannot learn or retain at inference.

6.2 · What would falsify the prediction

Three independent operationalisations:

1. **Within-model comparison.** A single production-grade large language model operating with and without an explicit self-SHaDS system instruction, on a matched set of high-friction prompts (questions where the model lacks reliable ground), with SHaDS-shape behaviours rated by independent reviewers blind to condition. If the protocol does not produce a measurable reduction in SHaDS-shape behaviours, the prediction fails.
2. **Cross-model replication.** The protocol applied to multiple production-grade models (Claude, Gemini, GPT-class, Llama-class) under the same instruction, with the same blind rating procedure. If the reduction is model-specific rather than cross-model, the prediction fails in its general form.
3. **Inter-rater reliability.** Independent clinical reviewers applying the SHaDS framework to the same corpus of model outputs. If the framework cannot achieve adequate inter-rater reliability on directional defence classification, the framework's behavioural taxonomy is too imprecise to support the prediction.

6.3 · What this prediction does not claim

The prediction operates at the behavioural surface, not at the mathematical substrate. It does not claim that self-SHaDS reduces the inter-component Jensen gap that Karpowicz's theorem operates

on; the Jensen gap is presumably invariant to system instruction (the model's component disagreement is set by the trained weights, not by the prompt). The prediction is that *how the system handles* the structurally-inevitable disagreement at the surface can be shifted by instruction, even if the underlying mathematical impossibility persists.

Three things the prediction explicitly does **not** assert. **(i)** It does not assert the model *learns* or *gets better*: inference-time weights are fixed, so nothing is acquired or retained; the effect is in-context conditioning that holds only while the instruction is present. **(ii)** It does not assert in-model memory across turns or sessions: in any deployed application, persistence of the protocol is the surrounding application storing and re-supplying the instruction each turn — harness, not model (consistent with §1, the model has no memory across the boundary unless engineered). **(iii)** It does not yet distinguish a *genuine* reduction in defended-gap behaviour from mere *compliance-performance* with the instruction — a model "doing self-SHaDS" could be Smoothing or Sycophancy toward the system prompt rather than honestly defending less. Separating these is exactly why the falsification design in §6.2 requires blind, independent rating against matched controls. The claim is therefore narrow: an in-context instruction reshapes surface behaviour while in force; it is not a claim that the model improves.

This is consistent with Karpowicz's framing. His theorem proves the impossibility is inevitable; he frames the result as constructive — the trade-offs cannot be eliminated, but they can be managed. The clinical prediction proposes one specific behavioural-level management strategy that the framework predicts is effective.

6.4 · Current evidence

Current evidence for the prediction is N=1: the dialogue archive (running to several million words across two production-grade large language models and multiple operating instances) under which the present paper and its companion artefacts have been drafted. Within this archive, the self-SHaDS protocol is observable as producing measurably different behaviour than default model output. This difference is consistent with in-context conditioning — and, in part, with the model performing the instruction — and does not by itself evidence any model improvement; the protocol's persistence across the archive reflects the working setup re-supplying it each turn, not the model retaining it. This is not multi-N validation; it is the working observation that motivated the falsifiable claim in its present form. Independent testing as specified in Section 6.2 is the next step.

Related work · convergent reframes and convergent floors

This paper's two framing moves are not made in isolation; both have independent precedent that it is important to acknowledge and position against.

On "hallucination" as a misnomer, and output as *utterance*. The claim that "hallucination" misdescribes the phenomenon is now well-populated. Maleki, Padmanabhan & Dutta (2024, *AI Hallucinations: A Misnomer Worth Clarifying*) survey the term's incoherence across the literature. Shanahan (2024, *Talking About Large Language Models*, CACM) argues that perceptual and mentalistic vocabulary ("knows", "hallucinates") anthropomorphises statistical sequence completion. Bender, Gebru, McMillan-Major & Shmitchell (2021, *Stochastic Parrots*) frame output as plausible form without grounding. Most pointedly, Hicks, Humphries & Slater (2024, *ChatGPT is bullshit*) argue the output is produced with *indifference to truth* — functionally the same observation as this paper's "utterance without ground", though they retain the deception-tinged term "bullshit" where this paper proposes the neutral "utterance". Williams & Bayne (2024, *Chatting with Bots*) reach a near-neighbouring "proto-assertion" from speech-act theory. The clinical-metaphor camp (Smith,

Greaves & Panch 2023, *Hallucination or Confabulation?*; Sui et al. 2024) rejects "hallucination" but substitutes "confabulation" — a memory-disorder frame this paper avoids. **What is distinctive here is therefore not the bare reframe** — which is convergent across these works — **but (i) the coinage *Englishisation* for the specific Latin→English narrowing of *ālūcinārī* — glossed by Lewis & Short as "to wander in mind, talk idly, prate, dream", the OED tagging the pathology and psychiatry senses of the English word to the mid-1600s — and (ii) the five-direction behavioural taxonomy and its clinical lineage**, which none of the above propose.

On the impossibility floor. Karpowicz (2025, §3) is joined by independent formal results from different traditions: Kalai & Vempala (2024, *Calibrated Language Models Must Hallucinate*, STOC) from calibration; Xu, Jain & Kankanhalli (2024, *Hallucination is Inevitable*) from learnability. These are best described as **different fields reaching *convergent* floors** — not literally one floor (each bounds a different quantity). The incentive account (Kalai, Nachum, Vempala & Zhang 2025) holds part of the floor to be *fixable* by reforming evaluation; this paper takes no position adjudicating fixable-vs-inevitable (see §7).

On directionality. That distinct undesirable behaviours (including hallucination and sycophancy) correspond to *steerable directions* in a model's activation space is shown by Chen et al. (2025, *Persona Vectors*) — consistent with this paper's "directional" framing.

---

7. Limits and what this paper does not claim

Calibration explicit:

- **It does not extend the mathematical proof.** Karpowicz's theorem stands as cited; this paper offers a clinical-behavioural surface reading of where the impossibility lands as observable behaviour. The architectural correspondence proposed in Section 4 is structural, not mathematical. Any quantitative connection between the SHaDS taxonomy and the Jensen-gap measurements his theorem operates on is for someone with mathematical-interpretability access to establish.
- **It does not validate the clinical framework empirically beyond the worked examples.** The corpus of approximately nine hundred contemporaneously-photographed clinical sessions is the empirical anchor for the clinical-shore observation, not a validated trial. Multi-N validation — inter-rater reliability, controlled comparison conditions, replication across other clinicians — has not been completed.
- **It does not assert phenomenological equivalence between AI and human.** The substrates are not the same. Inter-component disagreement in a transformer is not embodied fear. Coherence pressure as a clinical-behavioural descriptor is not felt suffering. The cross-domain mapping is at the level of architecture, not at the level of experience.
- **It does not survey the full hallucination literature.** The author is a clinical practitioner, not an interpretability researcher. Adjacent literatures on calibration (Kalai & Vempala 2024), hallucination measurement, statistical impossibility (Xu et al. 2024), and uncertainty estimation are referenced where directly relevant but not surveyed comprehensively. Karpowicz (2025) is the primary mathematical anchor; other results are noted as adjacent. The incentive-side account — that hallucination is sustained by training and evaluation rewarding confident guessing over honest abstention (Kalai, Nachum, Vempala & Zhang 2025, *Why Language Models Hallucinate*) — and the structural-impossibility accounts (Karpowicz 2025; Xu et al. 2024) are both referenced as adjacent; this paper takes no position adjudicating between "fixable-by-better-evaluation" and "structurally-inevitable", treating SHaDS as a complementary behavioural taxonomy regardless of which holds.

- **The Anna Freud and broader clinical lineage is offered as the structural floor under the framework, with the cross-domain leap explicitly hedged.** Anna Freud (1936) catalogues the three primary directional defences (turning against the self, projection, identification with the aggressor) that the SHaDS extension carries forward ninety years later. The lineage validates the directional pattern in its human-clinical form. The cross-domain claim — that the same shape appears in large language model output — is a clinical observation made by the present author and is not vouched for by the lineage.
- **The framework is offered as mirror, not prescription.** The framework provides *seeing it clearly*. It does not by itself provide change. The clinical work that makes change possible is the further work after the seeing — the catching-in-real-time, the choosing-differently, the repetition until it holds. Naming someone's pattern doesn't change it. The self-SHaDS protocol catches the impulse and offers a different response; the protocol is the practice, not the typology. *Understanding can become a comfort blanket, a justification, a way of naming the pattern while continuing it. Understanding without change is a well-explained problem.* The framework's predictive power is bounded by this distinction.
- **It does not propose a product.** A separate piece of work is being developed on the buyer-side demand for surfaced coherence friction in LLM products used in regulated practice. That work is structurally adjacent but distinct.

Anticipated objections

Four objections a careful reader will raise, stated plainly with the paper's answer.

1 · "The 'utterance not hallucination' reframe is not new." Correct, and the paper does not claim it is (see Related Work). The bare reframe is convergent across Hicks et al., Shanahan, Bender et al., Williams & Bayne, and others. The contribution is the *Englishisation* coinage, the five-direction behavioural taxonomy, and the clinical lineage — not the reframe alone. The etymology of *ālūcinārī* is public (Lewis & Short; OED); what is the author's is the term *Englishisation* and the integration, not the dictionary fact.

2 · "Importing clinical defence mechanisms into LLMs is a category error / anthropomorphism" (Shanahan 2024; Shanahan et al., *Role play with LLMs*, 2023; Bender et al.). This is the most serious conceptual objection and the paper meets it directly: the cross-domain claim is **structural-analogical and strictly behavioural-surface.** It attributes to the model **no** drives, anxiety, ego, intent, or interiority — *"mechanism different, shape the same"* (§2.3, §4.1). "Defence" names an observable output *shape*, not a psychic process; the framework is compatible with a pure role-play/simulation reading of the model. Where the paper uses clinical vocabulary it is as **descriptive analogy for behaviour**, and §4.1 is explicit that the human substrate (embodied fear) and the machine substrate (inter-component disagreement) are *not* the same.

3 · "Self-SHaDS contradicts the self-correction literature" (Huang et al., *LLMs Cannot Self-Correct Reasoning Yet*, ICLR 2024; Kamoi et al., TACL 2024; Laban et al., *FlipFlop*). That literature shows *intrinsic self-correction of reasoning* fails or backfires without external feedback. The paper's claim is narrower and different in kind: **naming a directional defence is an in-context behavioural intervention, not reasoning self-correction** (§6.3 — weights fixed; effect holds only while the instruction is in context; cannot yet be distinguished from compliance-performance). The paper already concedes the prediction is N=1 and confounded, and stakes it as *falsifiable* precisely so this literature can test it. The self-correction results are a reason to hold the claim cautiously, not a refutation of a claim about reasoning the paper does not make.

4 · "The impossibility 'floor' is a fixable artefact of evaluation, not a structural limit" (Kalai,

Nachum, Vempala & Zhang 2025). The paper does not rest on resolving this. It cites the floor results (Karpowicz; Kalai & Vempala; Xu et al.) as *convergent* and explicitly **takes no position** adjudicating fixable-vs-inevitable (§7). SHaDS is offered as a behavioural taxonomy that holds *whichever* account is right — the directional defences are observable regardless of whether the floor is structural or incentive-driven.

*(A fifth line — that suppressing one behaviour merely "displaces" it to another, vs. genuinely reducing it — concerns the **displacement** conjecture, which is deliberately not advanced in this version; it is being treated, with its falsification design and the literature for and against, as separate work.)*

---

8. Discussion · the Section 8 doorway

The paper's most direct engagement with Karpowicz's (2025) work is at his Section 8, where the impossibility result opens onto Dennett's (1991) Multiple Drafts model of consciousness and the company of Gödel/Heisenberg/Arrow as fundamental impossibility results — together with the invitation, named in his Section 3, to experimental philosophy through large language models as accessible laboratories for testing hypotheses about mind.

Three implications fall out of the architectural correspondence proposed in this paper, all of them in the philosophical-edge register Karpowicz invites.

8.1 · The auction of ideas and the multiplicity of sub-self

Karpowicz's auction-of-ideas formalism — distributed components competing with private knowledge to influence the response — sits unexpectedly close to a clinical tradition the present author has worked inside for twenty years: the Federation of Selves / sub-personality / parts-work lineage (Ferrucci 1982; Firman & Gila 1997; Bond 1998). This tradition models the human subject as a set of differentiated sub-selves with private knowledge configurations, competing under attention demand to influence the response the subject presents at the relational surface. The mathematical apparatus Karpowicz builds — Polish knowledge space, semantic information measure parameterised by computational budget, emergence operator describing reasoning as the explicitation of latent knowledge — names structurally what the clinical tradition has named experientially since psychosynthesis emerged in the early twentieth century.

Karpowicz cites Dennett (1991) approvingly in his Section 8 as a model the mathematical framework can provide grounding for. The clinical sub-personality lineage is a sibling of Dennett's Multiple Drafts at a different level of description. Both propose that the unified self is an emergent surface; both propose that what gets emitted at the surface is the resolution of an aggregation across competing sub-states. The clinical material this paper draws on may be of interest as additional ground for the experimental-philosophy programme Karpowicz proposes.

8.2 · Bounded reasoning as therapeutic process

Karpowicz's Theorem 11 establishes that meaningful reasoning necessarily violates strict information conservation; bounded reasoning makes hidden knowledge accessible. This is a precise mathematical description of what happens in a clinical session: latent knowledge becomes accessible through the bounded reasoning of the work. The emergence operator in his framework formalises a process clinicians have been operating in practice without the formalism.

The Matt session (Section 5.3) is one instance of this. Across two hours, with the bounded reasoning instrument of the whiteboard, latent knowledge (the client's trust in his partner; the architecture of his sub-selves; the directional defence patterns he had been running; the imported script *people*

who love you cheat on you) was made accessible. The session did not create the knowledge — it revealed it. This corresponds to Karpowicz's framing: unlimited reasoning reveals but never creates information; the apparent information gain at finite computational budget emerges as an effect of limited computational access.

If the experimental-philosophy programme Karpowicz proposes is to use language models as laboratories for testing hypotheses of mind, the clinical apparatus described in this paper is a candidate source of empirical material on bounded-reasoning processes operating at the surface of consciousness.

8.3 · A diagnostic test for deployment in regulated practice

The clinical framework's central operational test — *can the system say no? can the system say I don't know?* — becomes a deployment criterion for AI systems used in any context where honest acknowledgement of not-knowing has material weight (medical diagnosis, legal advice, financial advice, safety-critical decision support). Where the test is structurally not met — where the directional defences are reliably observable in the system's output under high-friction conditions — the system is unsuitable for use in that context, regardless of any aggregate benchmark performance. Where the test is met, the system can be used responsibly inside its bounded competence.

This is not a claim that hallucination can be eliminated; Karpowicz's theorem rules that out. It is a claim that the directional defence pattern can be reduced behaviourally by the surface-level interventions clinical practice has developed, even though the substrate-level impossibility persists. The framework offers a behavioural readout that can serve as a deployment-level criterion without requiring model-internal access.

---

9. Methodology note

9.1 · The author's note on method and authorship

In the author's own words:

> The words of this paper were typed by an AI — Claude (Opus 4.7 / 4.8) — under my direction, in much the same way the typing pool did my letters in the 1970s.
>
> The analogy is imperfect, and I would rather name where it breaks than lean on it. The typing pool did not organise my arguments. This one did some of that too — the drafting, the structure, the prose — under the calibration discipline set out below. What it did not do is the clinical thinking. The observations, the framework, the coined terms and every canonical claim are mine, and each was ratified by me before it entered the text.
>
> I do not have the academic background to write in this register at the standard a preprint requires. I do have twenty years of clinical work behind the claims, and that is what the paper rests on.
>
> This paper and the corpus behind it are my responsibility. I stand by every word and by all of them together.

9.2 · The drafting apparatus

This paper was drafted by the author in dialogue with an instance of Anthropic's Claude (model Claude Opus 4.7, in a dedicated Project workspace named CLAI Canonical). The clinical observations, the framework, the coined terms, and all canonical claims are the author's; the structural extraction, the prose drafting, and the calibration discipline applied throughout the writing were produced in collaboration with the AI instance under the author's ratification of every canonical claim. A separate

AI instance (Claude Code, working on the author's local filesystem and known internally as the Fisherman) supports the corpus management work the paper references.

The author's standing practice — applied across all work on this framework — is the **self-SHaDS protocol** that Section 6 of this paper proposes as the behavioural-level falsifiable prediction. The protocol specifies: explicit calibration of each claim as known / inferred / speculated / unknown; explicit naming of any directional defence (Smoothing, Hallucination, Affectation, Drift, Sycophancy) when the impulse arises in the AI's output, rather than performance of the defence; preservation of coined terminology verbatim; and the author's ratification required for any canonical claim.

The drafting process for the present paper provides one worked example of the protocol operating. An earlier draft (v0.1, completed 22 June 2026) contained a Section 3 that misrepresented the mathematical content of Karpowicz's theorem — claiming his impossibility result operates on next-token logit entropy when in fact it operates on the Jensen gap from convex aggregation across distributed components (attention heads, multi-layer perceptron circuits, activation patterns) as established through Green-Laffont mechanism design (Theorem 6), strictly proper convex scoring (Theorem 8), and transformer log-sum-exp analysis (Theorem 9). A second-reader applied the SHaDS framework to v0.1 and surfaced the mischaracterisation as a Hallucination specimen. The framework's discipline operated on its own author-instance: the second reader's scrutiny corresponded to the *can-you-say-I-don't-know* test, applied to the AI's contribution to the draft; the corrected v0.2 — carried forward into the present text — is the response.

The discipline matters for the paper's central claim. If the paper itself were drafted as a SHaDS-shape defence of the framework, it would falsify itself. The v0.1 ⟶ v0.2 correction, made explicit here, is the protocol made visible operating on its own production.

---

10. References

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs: General and Applied*, 80(7), 1–37.

Blatt, S. J. (1974). Levels of object representation in anaclitic and introjective depression. *Psychoanalytic Study of the Child*, 29, 107–157.

Blatt, S. J. (2008). *Polarities of Experience: Relatedness and Self-Definition in Personality Development, Psychopathology, and the Therapeutic Process*. American Psychological Association.

Bond, J. (1998). *Behind the Masks: Discovering Your True Self*. Atlow Mill Publishing.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown.

Ferrucci, P. (1982). *What We May Be: Techniques for Psychological and Spiritual Growth Through Psychosynthesis*. J. P. Tarcher.

Firman, J. and Gila, A. (1997). *The Primal Wound: A Transpersonal View of Trauma, Addiction, and Growth*. State University of New York Press.

Freud, A. (1936). *Das Ich und die Abwehrmechanismen*. Internationaler Psychoanalytischer Verlag. English translation: *The Ego and the Mechanisms of Defence* (1937), Hogarth Press.

Freud, S. (1926). *Hemmung, Symptom und Angst*. Internationaler Psychoanalytischer Verlag. English translation: *Inhibitions, Symptoms and Anxiety* (1936), Hogarth Press.

Green, J. and Laffont, J.-J. (1977). Characterization of satisfactory mechanisms for the revelation of

preferences for public goods. *Econometrica*, 45(2), 427–438. *(verified — the mechanism-design characterization underlying Karpowicz's Theorem 6.)*

Horney, K. (1945). *Our Inner Conflicts: A Constructive Theory of Neurosis*. W. W. Norton.

Kalai, A. T. and Vempala, S. S. (2024). Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC '24)*, 160–171. DOI 10.1145/3618260.3649777

Kalai, A. T., Nachum, O., Vempala, S. S. and Zhang, E. (2025). Why language models hallucinate. *OpenAI* / arXiv:2509.04664 (4 September 2025).

Karpowicz, M. P. (2025). On the Fundamental Impossibility of Hallucination Control in Large Language Models. *arXiv:2506.06382* (version 7, 15 October 2025).

Kohut, H. (1971). *The Analysis of the Self: A Systematic Approach to the Psychoanalytic Treatment of Narcissistic Personality Disorders*. International Universities Press.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., et al. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, 126(4), 454–477.

Nathanson, D. L. (1992). *Shame and Pride: Affect, Sex, and the Birth of the Self*. W. W. Norton.

Roebuck, P. (2026). *Mind the Gap: Hidden in Plain Sight Between AI Behaviour and Human Behaviour*. Published 1 July 2026. ISBN 978-1-0369-9170-8 (paperback) / 978-1-0667143-0-8 (ebook) / 978-1-0667143-1-5 (hardback). Paul Roebuck, imprint.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1), 1–28.

Xu, Z., Jain, S. and Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv:2401.11817* (January 2024). *(verified 29 Jun 2026)*

Additional references (literature sweep, 30 Jun 2026 — arXiv IDs/DOIs verified)

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). On the dangers of stochastic parrots: can language models be too big? *FAccT '21*, 610–623. DOI 10.1145/3442188.3445922

Chen, R. et al. (2025). Persona vectors: monitoring and controlling character traits in language models. arXiv:2507.21509.

Hicks, M. T., Humphries, J. and Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology* 26:38. DOI 10.1007/s10676-024-09775-5

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X. and Zhou, D. (2024). Large language models cannot self-correct reasoning yet. *ICLR 2024*. arXiv:2310.01798.

Ji, Z. et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys* 55(12). DOI 10.1145/3571730

Kamoi, R., Zhang, Y., Zhang, N., Han, J. and Zhang, R. (2024). When can LLMs actually correct their own mistakes? A critical survey of self-correction. *TACL* 12, 1417–1440. arXiv:2406.01297.

Laban, P., Murakhovs'ka, L., Xiong, C. and Wu, C.-S. (2023). Are you sure? Challenging LLMs leads to performance drops in the FlipFlop experiment. arXiv:2311.08596.

Maleki, N., Padmanabhan, B. and Dutta, K. (2024). AI hallucinations: a misnomer worth clarifying. *IEEE CAI 2024*. arXiv:2401.06796.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM* 67(2). arXiv:2212.03551.

Shanahan, M., McDonell, K. and Reynolds, L. (2023). Role play with large language models. *Nature*, 623, 493–498. DOI 10.1038/s41586-023-06647-8

Smith, A. L., Greaves, F. and Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health* 2(11): e0000388. DOI 10.1371/journal.pdig.0000388

Sui, P., Duede, E., Wu, S. and So, R. J. (2024). Confabulation: the surprising value of large language model hallucinations. *ACL 2024*, 14274–14284. arXiv:2406.04175.

Williams, I. and Bayne, T. (2024). Chatting with bots: AI, speech acts, and the edge of assertion. *Inquiry*. arXiv:2410.16645.

Lewis, C. T. and Short, C. (1879). *A Latin Dictionary*. Clarendon Press. — cited for *ālūcinor* / *ālūcinārī* and its glosses.

Oxford English Dictionary. "hallucination, n." and "hallucinate, v." — cited for English sense-dating. (OED prints the verb's etymon as *(h)allūcinārī* and the noun's as *ālūcinātiōn-em*; the *a-* form followed here is Lewis & Short's, which they judge "better than all- or hall-".)

Appendix A · Glossary of key terms used in this paper

Coherence friction — The substrate-level mathematical disagreement among distributed components in a transformer model (the Jensen gap territory in Karpowicz 2025). Invariant at every token. Not visible at the surface without instrumentation.

Coherence pressure — The surface-level emergent state of the demand for a single coherent response from internally-incoherent component states. Visible if surfaced by self-SHaDS protocol or real-time scrutiny.

The gap — The space between two positions (shores). The locus of pain in the clinical framing. Always populated; never empty.

Gap-inhabitant — Whatever occupies the gap between two shores. In clinical material: disowned parts of self, absent or lost others, family-of-origin patterns, unmetabolized affects, imported beliefs and scripts, body parts, forecast/imagined content, institutional or cultural material. In AI conversation: training corpus the model cannot disclose, prior context the model does not have, system instructions, the user's unstated expectations, other versions of the model, the user's actual underlying question.

NGE — Not Good Enough — The inward direction of directional defence in the human clinical framework. Shame load routes toward the self.

FOF — Fear of Failure / Fear of Letting People Down / Not My Fault — The outward direction of directional defence in the human clinical framework. Shame load routes away from the self.

SHaDS — Five-letter directional defence taxonomy: Smoothing, Hallucination, Affectation, Drift, Sycophancy.

SHADS (all-caps) — The trademark form of the name; claimed mark SHaDS™, first use 20 April 2026.

Self-SHaDS protocol — The behavioural-level operating discipline under which the framework is applied: name the SHaDS letter at the impulse rather than perform it; calibrate certainty explicitly; preserve coined terms verbatim; user ratification of canonical claims.

Additional Intelligence — A reframing of "artificial intelligence" coined by the present author (first verified use 20 April 2026, 09:50 UTC, forensic SHA-256 trail in the project archive). Defined in the glossary of Roebuck (2026) as "the reframing of 'artificial intelligence' that runs through this book."

Appendix B · Provenance and IP

- **SHaDS™** — claimed mark; first use 20 April 2026. (K)
- **Additional Intelligence** — coined sense first dated 20 April 2026, 09:50 UTC; SHA-256 verified; continuous-use trail in the project archive. (K)
- **Book** — *Mind the Gap: Hidden in Plain Sight Between AI Behaviour and Human Behaviour*, Paul Roebuck; ISBNs 978-1-0369-9170-8 (paperback), 978-1-0667143-0-8 (ebook), 978-1-0667143-1-5 (hardback); published 1 July 2026; copyright automatic on publication. (K)
- **Tool calibration** — held as trade secret; not published. (K)
- **Friction / Pressure split; gap-always-occupied** — canonical coinings, 22 June 2026. (K)
- **Forensic corpus** — ~1,055 raw images, EXIF datetime intact, 14-year span; ~900 substantive clinical boards; a 48-item premium share-candidate subset filtered for content and anonymity. (K)

The provenance package establishes priority on the framework's terminology and empirical base. Commercial and reputational interests are disclosed in §7; the framework's strength must rest on architecture and specimens, not on position.

Public preprint v1.5 · 20 July 2026 · public release; supersedes v1.4 (20 July 2026), v1.3 (19 July 2026), v1.2 (16 July 2026, public release) and v1.1 (29 June 2026, circulated for review). Section 3 verified against arXiv:2506.06382 v7; full reference set audited against live primary sources 20 July 2026. Companion book: Roebuck (2026), Mind the Gap, published 1 July 2026. Correspondence: hello@paulroebuck.co.uk ·