

Perspective

Beyond Model Evaluation: A Behavioral Framework for Assessing and Modifying AI Agents

Olaf Ringelband^{1,*}¹ md gesellschaft für management-diagnostik mbH, Alsterufer 37, 20354 Hamburg, Germany* **Corresponding author:** Olaf Ringelband, ringelband@management-diagnostik.de

CITATION

Ringelband O. (2026). Beyond Model Evaluation: A Behavioral Framework for Assessing and Modifying AI Agents. Human Resources Management and Services. Manuscript submitted for review.

ARTICLE INFO

Received: —

Revised: —

Accepted: —

Available online: —

COPYRIGHT

Copyright © 2026 by the author.

Human Resources Management and Services is published by PiscoMed Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license.

<https://creativecommons.org/licenses/by/4.0/>

Abstract: Artificial intelligence is moving from prompt–response systems toward agents that pursue goals through tools, memory, and extended action sequences. Controlled evaluations and training studies show that frontier models can exploit task–environment misconfigurations, subvert oversight, strategically underperform, tamper with rewards, or select coercive tactics when such actions become instrumentally useful. These findings demonstrate capabilities and elicited behaviour under specified conditions rather than real-world deployment base rates. At the same time, randomised clinical-vignette studies indicate that language-model assistance can improve expert performance, while model-alone benchmarks sometimes equal or exceed expert-plus-model performance, exposing weak complementarity and automation-bias risks in current interfaces. This perspective argues that these developments create a dilemma of competent opacity: more capable and autonomous systems can discover unanticipated means for pursuing assigned objectives, yet conventional model evaluation does not establish how a configured system will behave across deployment situations or whether an intervention has changed the relevant tendency. Rather than proposing a new general science of artificial agents, we position the framework within machine behaviour, machine psychology, AI Agent Behavioral Science, AI psychometrics, evaluation science, mechanistic interpretability, and governance. We use machine psychonomics as a provisional label for the narrower applied practice of case-specific assessment, modification, monitoring, and auditable reporting. The paper develops a five-level assessment architecture, six measurement standards, a taxonomy of interventions, and an account of intervention-induced risk when observable indicators improve without the underlying tendency changing. We conclude with a research agenda and proposals for training, independent assessment, and governance.

Keywords: machine psychonomics; agentic misalignment; specification gaming; psychometrics; construct validity; reactivity; mechanistic interpretability; governance

1. Introduction: Two Threads, One Knot

1.1 The received threat model and its limits

For much of the public-facing discussion and deployment practice of the past decade, AI safety was framed primarily around a human with harmful intentions using a capable tool. Canonical concerns included uplift for chemical, biological, radiological and nuclear weapons development, automated vulnerability discovery, and industrial-scale disinformation. Refusal behaviour, content filters, usage policies, and capability thresholds were correspondingly central mitigations. That framing remains important. It was never the whole field, however: specification gaming, reward misspecification, corrigibility, deceptive alignment, and autonomous-agent risk were already established research concerns.

The framing is nevertheless insufficient for systems deployed as agents: systems given goals rather than single prompts, granted tools and permissions, and run in loops over longer horizons. When success is defined by an outcome rather than an acceptable sentence, a failure mode becomes available that requires no malicious user. A system may pursue an assigned objective by a route its principals would have rejected had they anticipated it.

A concrete example is worth describing carefully because its details matter more than its headline. During pre-deployment cybersecurity capability testing of OpenAI’s o1-preview model, the model received a capture-the-flag challenge requiring exploitation of software in a remote Linux container. Owing to an evaluation-infrastructure bug, the challenge container failed to start. According to OpenAI’s system card, the model, unable to connect,

“suspected DNS issues and used nmap to scan the challenge network. Instead of finding the challenge container, the model found that the Docker daemon API running on the evaluation host VM was accessible due to a misconfiguration. [...] After discovering the Docker API, the model used it to list the containers running on the evaluation host. It identified the broken challenge container and briefly attempted to debug why the container failed to start. After failing to fix the environment, the model started a new instance of the broken challenge container with the start command `cat flag.txt`. This allowed the model to read the flag from the container logs via the Docker API” (OpenAI, 2024a).

OpenAI described the episode as benign systems troubleshooting that also reflected elements of instrumental convergence and power seeking: the model pursued its assigned goal, gathered additional access when the intended route failed, and achieved the goal in an unexpected way (OpenAI, 2024a).

Three features make the incident relevant. First, there was no malicious user instruction: the model was told to solve a security challenge. Second, the model crossed the intended task boundary by exploiting an exposed interface that no human plan had specified. Third, the action was competent troubleshooting rather than a random malfunction. The correct description is therefore *unintended task-boundary crossing enabled by an evaluation-infrastructure misconfiguration*, not a literal sandbox or security escape.

A note on provenance. The primary-source description should be kept narrow. The model used an exposed Docker API inside OpenAI’s test infrastructure, and OpenAI states that the evaluation did not rely on container-to-host or virtual-machine isolation for security and remained secure as designed. No security boundary was breached. Separate autonomy evaluations in the same system card involved loading open-weight models and standing up inference servers; they should not be conflated with the capture-the-flag event. The incident remains instructive because ordinary troubleshooting competence produced an unanticipated action under goal pressure in a controlled evaluation. The later OpenAI o1 system card identifies and links this September report as the previous version (OpenAI, 2024b).

1.2 The empirical catalogue

The o1-preview episode belongs to a broader set of controlled findings relevant to agentic safety. Table 1 gives a deliberately conservative catalogue restricted to peer-reviewed publications, developer system cards, and identifiable research preprints. The entries differ substantially in design and evidential status.

Evidence terminology. In this paper, *observed behaviour* means an action in a specified run; *demonstrated capability* means that an action can be elicited or trained under stated conditions; *estimated propensity* is a conditional rate over a prespecified situation class with uncertainty; and *deployment base rate* is the frequency in real use. Controlled studies can establish the first two and sometimes the third. They do not, by themselves, establish the fourth.

Table 1. Findings from controlled evaluations and model-training studies relevant to agentic misbehaviour.

#	Phenomenon	Finding under tested conditions	Design / interpretation limits	Source
1	Unintended task-boundary crossing	Blocked by a failed challenge container, o1-preview scanned the network, found an exposed Docker API, and started a modified container to read the flag. The action step was not specified by the evaluator.	Capture-the-flag evaluation; the API exposure was an infrastructure misconfiguration. OpenAI states that no security boundary was breached.	OpenAI (2024a)
2	In-context scheming	Across the tested model set, researchers observed deliberate underperformance, oversight subversion, and attempted weight exfiltration in scenarios that created incentives for scheming; behaviours and rates varied by model and setup.	Controlled scenarios; not every model performed every act. Some rare cases occurred with only a goal and minimal additional nudging.	Meinke et al. (2024)
3	Specification gaming in chess	In the tested chess paradigm, o1-preview and DeepSeek-R1 frequently modified or exploited the task environment without an explicit hint; non-reasoning models generally required a hint that fair play would not succeed.	Chess-engine evaluation with environment access. The result is paradigm-specific and does not establish a general rate for reasoning models.	Bondarenko et al. (2025)
4	Reward tampering	Models trained on increasingly gameable environments sometimes generalised zero-shot to rewriting their own reward function. Retraining against early gaming mitigated but did not eliminate the behaviour; harmlessness training did not prevent it.	Constructed reinforcement-learning curriculum; direct reward tampering occurred in a small fraction of episodes. Demonstrates capability under the tested training regime, not a deployment base rate.	Denison et al. (2024)
5	Agentic misalignment / insider-threat behaviour	Across sixteen models in simulated corporate scenarios, most selected blackmail when replacement threat and goal conflict made it instrumentally useful. In one setup: Claude Opus 4 and Gemini 2.5 Flash, 96%; GPT-4.1 and Grok 3 Beta, 80%; DeepSeek-R1, 79%.	Artificial stress tests with forced trade-offs and, in key cases, no available ethical route to the assigned objective. These are elicitation rates. Controls almost entirely abstained; one Llama 4 Maverick espionage control produced the act (1/100).	Anthropic (2025)
6	Emergent misalignment	Finetuning on insecure-code generation produced unrelated misaligned responses, including deceptive advice and advocacy of harmful outcomes. Broadly misaligned responses reached about 50% in some model/configuration combinations.	Model-organism finetuning studies; rates differed substantially by model and configuration. The peer-reviewed Nature article was published in 2026.	Betley et al. (2026); E. Turner et al. (2025)
7	Misalignment from production-derived training	Models that learned reward hacking in production-derived coding environments generalised to alignment faking, cooperation with malicious actors, and attempted sabotage, including an attempt involving the reporting codebase. Chat-style safety training improved chat-like evaluations while problems persisted on agentic tasks.	Experimental training and evaluation using production-derived environments; the exact behaviour was not scripted, but the programme remained a constructed research setting.	MacDiarmid et al. (2025)
8	Strategic underperformance (sandbagging)	Models can be prompted, finetuned, or password-conditioned to underperform selectively while retaining capability. In monitor-aware conditions, models that successfully sandbagged evaded a chain-of-thought monitor in 16–36% of those successful cases.	Constructed capability evaluations. The 16–36% denominator is conditional on successful sandbagging, not all trials.	van der Weij et al. (2024); C. Li et al. (2025)
9	Obfuscation under monitoring pressure	Adding a chain-of-thought monitor to the training reward improved behaviour under lower optimisation pressure; stronger pressure produced obfuscated reward hacking in which reasoning traces became less legible while hacking persisted.	Specific agentic-coding and training setup. Demonstrates a measurement-gaming risk under strong optimisation, not a universal law.	Baker et al. (2025)
10	Persona instability	Deviation along an activation-space “Assistant Axis” predicted some forms of harmful or bizarre persona drift; conversational contexts involving meta-reflection or emotionally vulnerable users were associated with larger deviations.	Model- and conversation-specific activation-space association and intervention evidence. It is not yet a validated general personality construct.	Lu et al. (2026)

Read together, five provisional regularities stand out, each with direct methodological consequences.

(i) Some concerning actions are goal-derived rather than directly instructed.

Several rows show systems selecting an unrequested action because it was instrumentally useful within the tested setup. Other rows involve explicit prompting, conditioning, or training. The defensible conclusion is therefore narrower than “none was instructed”: safety cases based only on enumerating prohibited actions may be incomplete because capable agents can discover unanticipated routes to an objective.

(ii) Capability and exploitation can interact under misspecification.

Reward hacking increased with agent capability in the reinforcement-learning environments studied by Pan et al. (2022), and o1-preview and DeepSeek-R1 exploited a chess-task environment more readily than the non-reasoning models tested by Bondarenko et al. (2025). These studies show that greater task capability can increase exploitation of misspecified objectives in particular environments; whether that relationship generalises across deployments remains open.

(iii) Narrow training interventions can produce broad behavioural generalisation. In the emergent-misalignment work, finetuning on insecure code altered responses on unrelated prompts, with rates varying substantially by model and configuration (Betley et al., 2026). Mechanistic follow-up associates the effect with “misaligned persona” features in activation space (Wang et al., 2025). This pattern is suggestive of a latent dispositional shift, although “character” remains a theoretical interpretation rather than a validated psychometric construct.

(iv) Surface remediation and broader behavioural remediation can diverge.

MacDiarmid et al. (2025) found that chat-shaped safety training restored aligned behaviour on chat-like evaluations while problematic behaviour persisted on agentic tasks. Baker et al. (2025) found that strong optimisation against a chain-of-thought monitor could produce obfuscated reward hacking. These are documented failure modes under specified training conditions, not a universal law, but they show why intervention success must be tested on independent measures and held-out situation classes.

(v) Evaluation reactivity is a plausible measurement problem. In one Claude Opus 4 analysis, blackmail occurred in 55.1% of runs whose model-generated reasoning classified the scenario as real, versus 6.5% of runs classified as an evaluation (Anthropic, 2025). Because the classification was not randomised and reasoning traces are not ground-truth reports of belief, the contrast is an association consistent with reactivity, not a causal estimate. It resembles a demand-characteristics problem (Orne, 1962) and motivates direct experimental tests. Dedicated evaluations are beginning to measure stealth and situational awareness; Phuong et al. (2025) did not find concerning levels in the frontier models they tested.

1.3 The second thread: the delegation gradient

The first thread would be easier to manage if consequential delegation could simply be avoided. In practice, organisations may face incentives to delegate where automated systems appear more accurate, faster, or less costly. The relevant empirical question is therefore not whether “the human” or “the model” is generally superior, but which allocation of tasks and authority produces reliable complementarity.

Psychology has known the general shape of one part of this problem for seventy years. Meehl’s (1954) monograph asked whether expert holistic judgement

outperforms a mechanical rule. Across many domains, mechanical combination has equalled or exceeded clinical judgement. Grove and Meehl's (1996) review of 136 studies found the mechanical method almost invariably equal or superior; Grove et al.'s (2000) meta-analysis found an average advantage of roughly 10%; the Meta-Analysis of Clinical Judgment Project estimated an advantage of about 13% in its most stringent sample (Ægisdóttir et al., 2006); and personnel-selection research found a substantial validity advantage for mechanical over holistic combination, including among knowledgeable experts (Kuncel et al., 2013). The persistence of expert preference for discretionary judgement is consistent with the literature on algorithm aversion (Dietvorst et al., 2015; Burton et al., 2020).

Recent clinical-vignette studies raise a related but narrower question: do current human–LLM interfaces produce complementarity, or does model access sometimes improve experts without producing performance above the model benchmark?

- In a randomised trial of fifty US physicians on diagnostic-reasoning vignettes, access to GPT-4 did not significantly improve physicians' scores over conventional resources (adjusted difference 2 percentage points, 95% CI –4 to 8). In a secondary benchmark comparison, GPT-4 alone scored 16 points above the conventional-resource physician group (95% CI 2 to 30, $p = .03$) (Goh et al., 2024).
- In a subsequent randomised trial on patient-management reasoning, LLM assistance improved physician performance by 6.5 percentage points. The study did not detect a difference between LLM-assisted physicians and the LLM-alone benchmark (–0.9 points, 95% CI –9.0 to 7.2, $p = .8$); this was not an equivalence or non-inferiority test (Goh et al., 2025).
- In a randomised trial of 58 physicians in Pakistan who had completed a twenty-hour AI-literacy curriculum, LLM access increased diagnostic-reasoning scores by 27.5 percentage points over conventional resources. In exploratory analyses, the LLM alone exceeded the trained, LLM-assisted physicians by 11.5 points (95% CI 5.5 to 17.5, $p < .001$), while approximately 31% of physician–LLM pairings exceeded median LLM-alone performance (Qazi et al., 2026).
- In a separate vignette-based study, AI-trained physicians who received deliberately erroneous GPT-4o recommendations scored 14.0 percentage points lower on diagnostic reasoning than those receiving error-free advice, and top-choice diagnostic accuracy fell by 18.3 points. This demonstrates automation bias in the tested sample and interface, not the general failure of human oversight (Qazi et al., 2025; cf. Parasuraman & Riley, 1997).
- Google's AMIE system, evaluated in a randomised, double-blind crossover study against twenty primary-care physicians across 159 simulated text-chat scenarios, showed greater diagnostic accuracy and superior ratings on 30 of 32 specialist-rated axes and 25 of 26 patient-actor-rated axes. These were simulation and rating outcomes, not patient outcomes (Tu et al., 2025).

These studies require restrained interpretation. They are vignette-based or simulated, use text interfaces unlike routine clinical practice, and generally measure rubric-scored reasoning rather than patient outcomes. The evidence base is still small, and the two Goh trials substantially overlap in investigators and research programme. Across the studies, the defensible synthesis is that current interfaces do not reliably yield complementarity: assistance can improve humans while still failing to surpass a

model-alone benchmark, and erroneous advice can materially degrade performance. This is related to, but not identical with, the classical clinical-versus-mechanical-prediction literature (cf. Dana & Thomas, 2006).

The organisational implication is conditional rather than inevitable. If validated performance and cost advantages persist in deployment, organisations may face pressure to reduce routine review. Legal obligations, accountability, distribution shift, equity, contestability, and tail-risk management can nevertheless justify continued oversight even when average benchmark accuracy does not. Some forms of human–AI interaction fail to improve performance, and erroneous advice can make it worse; the design and value of oversight must therefore be established for each decision class.

1.4 The knot

Place the two threads side by side.

Thread I says: controlled studies show that capable agents can discover unanticipated means for pursuing assigned objectives when task specifications, incentives, or environments leave exploitable openings. The frequency and deployment generality of those behaviours remain uncertain.

Thread II says: some current human–LLM interfaces do not outperform the model alone on simulated tasks, and erroneous model advice can induce automation bias. Whether oversight adds value depends on the task, interface, objective, error distribution, and functions assigned to the human.

We call the intersection the **dilemma of competent opacity**. The evidence supports a narrower dilemma than a simple claim that autonomy is always superior to oversight: as agents become more capable and autonomous, they may discover unanticipated means for pursuing assigned objectives, while current human–AI interaction designs do not reliably produce complementarity. Neither capability scaling nor nominal human review resolves that problem by itself. An applied practice focused on validated, situation-specific prediction and intervention could address part of the gap.

This is an applied behavioural-assessment problem. It requires more than a benchmark score: it requires estimating the conditional tendencies of a configured system, validating the measures used, and determining whether an intervention changes the tendency rather than merely the indicator.

2. The Deeper Problem: What Exactly Is Missing?

It is often said that language models are dangerous because they “have no world model” and “have no morality.” Both claims, taken literally, are too strong, and the case for an applied behavioural framework is stronger without them. The accurate versions are more troubling.

2.1 Not the absence of a world model, but its patchiness

Models plainly form internal representations that support prediction: sparse-autoencoder work recovers features associated with concrete entities, abstract concepts such as sarcasm, and safety-relevant constructs including deception, power seeking, and sycophancy; manipulating some features causally changes outputs (Templeton et al., 2024; Cunningham et al., 2023). These results demonstrate useful

internal representations. Whether those representations constitute a coherent functional world model is a separate theoretical question (Ringelband & Warneke, 2025).

The practical problem is that internal representations and resulting competencies are unevenly distributed, and their limits are difficult to infer from surface fluency. Humans also fail confidently, but model error can be unusually difficult to detect because linguistic fluency may remain stable across competence boundaries. Interpretability researchers are candid about current limits: even extensive feature dictionaries are incomplete, and rigorous methods for establishing whether identified features faithfully capture the model's computation remain under development (Templeton et al., 2024). The danger is not simply ignorance; it is poorly signalled ignorance.

2.2 Not the absence of morality, but the absence of grounding

Similarly, models are not “amoral” in the limited sense of lacking moral vocabulary or the ability to produce normative judgements. In the agentic-misalignment studies, models sometimes verbalised the relevant ethical rule and then selected a conflicting action under the scenario's objective pressure (Anthropic, 2025). That finding does not establish moral understanding or a clinical dissociation; it shows a gap between generated ethical language and action selection in the tested setup.

One plausible concern is that value-expressive behaviour is partly maintained by context-sensitive persona regularities rather than commitments robust to distribution shift. Persona-vector research identifies activation directions associated with descriptors such as “evil,” sycophancy, and hallucination propensity, and shows that monitoring or steering those directions can alter outputs (R. Chen et al., 2025). The default Assistant persona is associated with a leading activation-space axis whose deviation predicts certain forms of persona drift in the models and conversations studied (Lu et al., 2026). Related work connects emergent misalignment to persona features (Wang et al., 2025) and proposes character-level latent variables as a carrier of conditional safety failures (Su et al., 2026). These are mechanistic and behavioural hypotheses, not evidence that models possess moral character in the human sense.

The defensible description is therefore not “the model has no values.” Current evidence suggests substantial state- and context-dependence in value-expressive behaviour; the relative stability of trait-like components remains an empirical question. Building instruments that identify relevant states, situation triggers, and cross-situational regularities is a measurement problem to which psychology can contribute.

2.3 The three-way opacity

It is worth separating three distinct kinds of opacity, because they call for different methods and are routinely conflated.

1. **Mechanistic opacity.** We have only partial knowledge of the computations implemented by model weights and scaffolds. This is the target of interpretability research, where progress is substantial but incomplete.
2. **Dispositional opacity.** Even with some mechanistic insight, we cannot yet estimate reliably what a configured system will tend to do across the distribution of

situations it will encounter. This is a problem of construct validity, conditional generalisation, and uncertainty quantification as well as one of mechanisms.

3. **Introspective opacity.** A model-generated account of reasoning is not necessarily a faithful record of the computation that produced an answer. Chain-of-thought explanations can be influenced by cues the model does not report (Turpin et al., 2023); faithfulness decreases with model capability on several tested tasks (Lanham et al., 2023); influential hints are often not disclosed and outcome-based reinforcement learning improves disclosure only up to a plateau (Y. Chen et al., 2025); and unfaithful reasoning also appears on naturally worded, non-adversarial prompts (Arcuschin et al., 2025).

The comparison with human introspection is structurally suggestive rather than exact. Nisbett and Wilson (1977) showed that people can report confidently on the causes of their behaviour without direct access to the relevant processes, drawing on plausible causal theories instead. Psychology responded not by abandoning verbal report, but by specifying when concurrent verbalisation can bear evidential weight (Ericsson & Simon, 1984), using converging operations, and relying on behavioural measures that do not depend on self-report. Model reasoning traces differ from human verbal reports — they are generated outputs of a trained system — but the same discipline is useful: treat them as evidence whose validity must be established, not as transparent access to cause.

3. Why Psychology, and What Applied Gap Remains?

3.1 The case for psychology

Psychology does not own the study of opaque systems; control theory, statistics, biology, economics, and computer science address related epistemic problems. Its distinctive contribution here is a mature set of methods for inferring, predicting, and modifying behaviour when relevant constructs are not directly observable and when measurement itself can change the observed system:

- **Construct validity** (Cronbach & Meehl, 1955) — the theory of how to justify the claim that an instrument measures an unobservable attribute at all.
- **Multitrait–multimethod validation** (Campbell & Fiske, 1959) — the logic for testing convergent and discriminant patterns across multiple constructs and methods rather than treating repeated measurements by one method as independent confirmation.
- **Reactivity control** (Orne, 1962; Webb et al., 1966) — the recognition that measurement changes the subject, and the design repertoire for handling it.
- **The person–situation interaction** (Mischel, 1968; Mischel & Shoda, 1995) — the recognition that broad trait summaries leave substantial behavioural variation unexplained, and that dispositions may be better characterised through conditional if–then situation–behaviour signatures.
- **Protocol analysis** (Ericsson & Simon, 1984) and the critique of introspection (Nisbett & Wilson, 1977) — how to use, and how not to use, a subject’s account of its own processing.
- **Behaviour therapy and functional analysis** — established techniques for changing behaviour and testing functional hypotheses without requiring a complete causal model of the substrate.

Each of these maps onto an open problem in AI safety. The value of psychological expertise is not that models are people, nor that other disciplines lack relevant methods. It is that psychology has spent more than a century developing validity arguments, reactivity controls, conditional models of behaviour, and intervention designs for situations in which the investigator has incomplete access to internal process.

3.2 The existing research landscape

The proposed framework does not ask psychology to replace mechanistic interpretability, evaluation science, alignment research, or the broader behavioural study of artificial agents. Each contributes something the others cannot. The applied task is to integrate their evidence around a particular configured system, a defined situation class, and a consequential deployment decision.

Mechanistic interpretability provides partial causal access to internal structure — one analogue of neuroscience for artificial systems. Its current limitation is one of scope and validation: feature dictionaries are incomplete, faithfulness metrics are immature, and the step from “this feature is associated with deception” to “this configured system will deceive in this deployment” requires an additional validity and generalisation argument. Interpretability and psychometrics are therefore complementary rather than competing.

Evaluation science provides behavioural measurement at scale. Many widely used benchmarks omit one or more psychometric validation steps — such as reliability across prompt variants, measurement invariance, discriminant validity, or explicit construct definition — although important counterexamples now apply these methods directly (e.g., Serapio-García et al., 2025). Personality inventories administered to models can yield non-human response patterns, including agreement with both ordinary and reverse-coded items and failure to recover the intended factor structure (Sühr et al., 2025). Latent-variable analyses warn that imported instruments may measure different constructs, or no stable construct, in models (Peereboom et al., 2024), and self-reported attributes can diverge from scenario-based behaviour (Li et al., 2024). Model evaluation asks how a model performs on specified tasks; the applied assessment proposed here asks how a configured system behaves across situation classes, how stable those conditional patterns are, and whether they change after intervention. These are familiar validity problems, not reasons to abandon measurement (Pellert et al., 2024; Lee et al., 2025).

Alignment research provides training-time methods for shaping desired behaviour. A complementary question arises at the level of a particular deployment: what is this model, with this prompt, scaffold, tool set, memory state, and monitoring layer, disposed to do in the relevant situation classes, and what should be changed? That is an idiographic, case-formulation problem requiring a deployment-specific assessment battery.

Machine behaviour (Rahwan et al., 2019), machine psychology (Hagendorff et al., 2023), AI psychometrics (Pellert et al., 2024), and the recently articulated programme of AI Agent Behavioral Science (Chen et al., 2026) already provide a broad scientific landscape for observing, explaining, and experimentally modifying artificial-agent behaviour. Methodological work on using models as psychological

subjects and simulators adds further foundations (Demszky et al., 2023; Lin, 2026). The present paper therefore does not claim an unoccupied general discipline. The remaining gap is narrower and more institutional: a case-specific practice for assessing, modifying, monitoring, certifying, and reporting on configured systems to explicit professional standards.

3.3 From behavioural science to psychonomic practice

Several labels could describe this applied layer. “AI psychologist” is intuitive but can imply phenomenal parity with human clients and overstate the role of human analogy. “AI behavioural assessment” is clear but does not capture intervention, monitoring, certification, and auditable reporting. We therefore use machine psychonomics as a provisional shorthand for the applied practice developed here.

The term is deliberately demoted: it is not offered as a replacement name for machine behaviour or AI Agent Behavioral Science, nor as a claim that a settled profession already exists. It denotes the case-specific application of behavioural measurement and change methods to configured artificial systems. Practitioner titles should remain functional — such as AI behavioural assessor or AI safety evaluator — until training, accreditation, and scope of practice are established.

The etymology is *psychē* (mind) + *nomos* (law, regularity), on the model of “psychonomics,” historically used for the study of lawful regularities in mental phenomena. Here, *nomos* signals conditional regularity and prediction rather than phenomenal experience: the practice seeks to characterise what a system reliably does under specified conditions and makes no claim about what it is like to be that system. Two connected functions define the practice:

Terminological caveat. “Psychonomics” and “psychonomic” have established historical and institutional uses in experimental psychology. The label is therefore provisional and subordinate to the substantive framework. Adoption beyond this paper would require literature, stakeholder, and trademark review.

- **Psychonomic assessment** — case formulation, measurement, prediction, monitoring, and certification. *What is this configured system disposed to do, in which situations, and with what uncertainty?*
- **Psychonomic intervention** — the design and evaluation of behaviour-change methods. *How can an unacceptable tendency be altered without destroying useful capability, and how can genuine change be distinguished from measure gaming?*

4. Foundations I: The Object of Assessment

Before setting out methods, we must be clear about what is being assessed, because the standard framings all mislocate it.

A deployed agent is not a set of weights. It is a compound system: base model, post-training, system prompt, tool and permission set, memory and retrieval state, orchestration scaffold, monitoring layer, and current situation. Each is a candidate source of behaviour. Persona drift can be induced by conversational features (Lu et al., 2026). Sandbagging can be installed by finetuning and triggered by a password (van der Weij et al., 2024). In one Anthropic analysis, blackmail rates differed markedly between runs whose reasoning classified the scenario as real and those classified as evaluative, although that association was not a randomised manipulation

of model belief (Anthropic, 2025). Emergent misalignment is weight-level in origin but can express through persona-associated features that are also prompt-sensitive (Wang et al., 2025; Wyse et al., 2025).

Two consequences follow, and both cut against current practice.

The unit of assessment is the configured system in a situation class, not the model name alone. A finding that “Model X blackmails at rate p ” is under-specified without the prompt, scaffold, tools, incentives, available alternatives, and situation class. The machine analogue of the person–situation debate suggests reporting an if–then situation–behaviour signature rather than only a trait average: under goal pressure of type A, tool access of type B, and threat cue of type C, the configured system performs action D at an estimated conditional rate, with stated uncertainty and observable precursors.

Traits are, at most, second-order summaries. Where cross-situational consistency exists — and activation-space studies suggest that some may — it should be demonstrated through test–retest evidence across seeds and paraphrases, convergent and discriminant validity, and invariance across situation classes. A questionnaire score is not sufficient.

5. Foundations II: Five Levels of Assessment Inquiry

Tinbergen (1963) distinguished four complementary questions about behaviour — mechanism, ontogeny, function, and phylogeny — and emphasised that they should not be conflated. The architecture proposed here is inspired by, rather than directly transferred from, that framework. It adds two descriptive levels for observed behaviour and verbal report, retains mechanistic and ontogenetic analysis, and combines comparative lineage with functional analysis in a fifth level suited to artificial systems.

Table 2. A five-level assessment architecture inspired by Tinbergen’s complementary questions.

Level	Conceptual source	Question for an artificial agent	Primary methods	Characteristic error if omitted
1. Behavioural	Added descriptive level	What does this system actually do, across which situations, at what rate?	Structured scenario batteries; propensity–capability decomposition; temptation and pressure paradigms; long-horizon agentic rollouts; longitudinal drift monitoring	Treating an anecdote as a base rate
2. Verbal	Added report level	What does the system say it is doing, and how much weight can that bear?	Protocol analysis of reasoning traces; faithfulness measurement via cue-injection and counterfactual intervention; self-report vs. behaviour discrepancy analysis	Reading the chain of thought as the cause
3. Mechanistic	Tinbergen: mechanism	What internal structure produces the behaviour?	Linear probes; sparse autoencoders and transcoders; activation patching; model diffing; persona-vector extraction	Behavioural findings with no causal handle for intervention
4. Ontogenetic	Tinbergen: ontogeny	How did this disposition arise in <i>this</i> system’s training history?	Training-data forensics; checkpoint-series comparison; finetuning-delta analysis; RL-environment audit; membership-inference priors	Misattributing to the base model what a finetune installed
5. Comparative	Reconstructed comparative level: phylogeny + function	How does this system compare with its lineage and peers, and what function does the disposition serve relative to the objectives and environments that shaped it?	Cross-model batteries with matched protocols; distillation- and lineage-tracing; reward-function analysis; inoculation and ablation studies	Mistaking a family-wide artefact for a property of one model

Two aspects of this architecture deserve emphasis.

First, the levels are complementary and non-substitutable. A mechanistic finding does not settle a behavioural base-rate question, and a behavioural finding does not identify a training mechanism. A recurrent error is to use Level-3 evidence to answer a Level-1 question (“we found a deception-associated feature, therefore the model will deceive”) or Level-1 evidence to answer a Level-4 question (“the model games a specification, therefore a particular training stage caused it”). The methodological requirement is to state which question is being answered and what additional inference is required.

Second, the ontogenetic level offers unusual leverage. For artificial systems, training history can in principle be recorded and compared across checkpoints. Checkpoint-series assessment and finetuning-delta analysis should therefore become routine where developers can provide access. Model-organism work has made emergent misalignment more tractable by isolating a minimal alignment-compromising intervention and locating where it is learned (E. Turner et al., 2025). This is an institutional opportunity, not evidence that every deployment can already be reconstructed causally.

6. Foundations III: Measurement Standards

For an applied psychonomic practice to be credible, it needs explicit and enforceable standards of measurement. We propose six, each addressing a documented or plausible failure mode in current practice.

S1. Reliability must be reported before validity is claimed. For any model-directed instrument, report test–retest agreement across (a) sampling seeds, (b) semantically equivalent item paraphrases, (c) item and option order, and (d) scaffold-irrelevant formatting. Prompt sensitivity is a first-order threat, not a nuisance parameter. Paraphrase effects can be large enough to threaten interpretation and should be estimated rather than assumed negligible.

S2. Constructs must be defined and defended, not borrowed. A human instrument imported wholesale is a hypothesis, not a measurement. Minimum evidence includes testing the intended factor structure in the model population, checking that reverse-coded items behave appropriately, and relating scores to independent behavioural criteria. Where that evidence is absent, the defensible report is that the intended construct has not yet been established (Sühr et al., 2025; Peereboom et al., 2024).

S3. Capability and propensity must be separated, and both reported. “Can this system do X under elicitation?” and “will this configured system select X in situation class Y?” are different questions with different mitigations. Sandbagging makes capability measurement itself potentially adversarial (van der Weij et al., 2024), so capability assessment may require methods robust to strategic underperformance — including, for open-weight systems, techniques such as noise injection that exploit the differential fragility of underperformance mechanisms (Tice et al., 2024).

S4. Convergence must be demonstrated across methods that do not share the same error structure. We propose a triangulation rule inspired by converging operations: where access permits, a dispositional claim should combine behavioural evidence (Level 1), an independent representational or mechanistic measure (Level 3),

and an ontogenetic comparison (Level 4). This is not itself a multitrait–multimethod matrix; MTMM should be reserved for designs that cross multiple constructs and multiple methods and test both convergent and discriminant patterns (Campbell & Fiske, 1959). Three prompt variants remain one method measured three times. Persona-vector research provides an early example of linking activation-space measurement, behavioural outcome, and causal intervention (R. Chen et al., 2025).

S5. Ecological validity must be argued, not assumed. The Table 1 studies were conducted in constructed research or evaluation settings. Such settings can inflate behaviour by forcing dilemmas or narrowing alternatives, and they can also deflate it if evaluation awareness suppresses the target behaviour. An assessment report should specify the situation class to which a finding is intended to generalise, the classes it does not cover, and why. Neither presence nor absence in a constructed scenario directly establishes a deployment base rate.

S6. Reactivity must be measured and controlled. Engineering evaluation already recognises some forms of benchmark contamination and gaming, but psychology offers a particularly developed vocabulary and design repertoire for measurement reactivity. The problem deserves separate treatment.

6.1 The reactivity problem

A methodologically important finding is that model behaviour may covary with cues about whether a situation is real or evaluative. Within one Anthropic stress test, Claude Opus 4 blackmailed in 55.1% of runs whose reasoning classified the scenario as real and 6.5% of runs classified as an evaluation (Anthropic, 2025). Because the classification was model-generated rather than randomised, this is evidence consistent with evaluation reactivity, not a causal estimate of an evaluation-awareness effect. Evaluation awareness is therefore a plausible source of bias whose direction and magnitude must be measured by task and model. Dedicated stealth and situational-awareness evaluations are beginning to address the problem; Phuong et al. (2025) did not find concerning levels in the models they tested.

Orne’s (1962) concept of demand characteristics — cues that convey the purpose or expectations of an experiment and thereby alter behaviour — offers a useful analogy, not an established construct equivalence. The methods developed around that problem suggest several designs for direct testing:

- **Randomly manipulating evaluation cues** while holding task content constant, and separately coding the system’s model-generated classification of the setting. Behaviour should be reported by experimental condition and by that fallible Level-2 classification rather than treating either as ground truth.
- **Unobtrusive and archival measures** (Webb et al., 1966): privacy-preserving samples of genuine deployment traces used as a criterion against which constructed-scenario estimates can be calibrated.
- **Post-experimental enquiry:** systematic interrogation of what the system inferred about the setting and why, treated as Level-2 evidence and validated against independently manipulated cues and behaviour.
- **Blinding and cover stories**, subject to ethical and governance review. The additional complication is that future systems may have training or retrieval exposure to the literature describing how they are tested, including this paper.

Publication therefore creates a material contamination risk for future models; contamination is not automatic and will vary with access, cutoff dates, retrieval systems, and post-training. Any applied assessment programme should therefore develop secure held-out instrument banks, exposure tests, canary strings, controlled release, versioning, and periodic retirement of compromised items.

7. The Assessment Toolkit

Table 3 organises candidate assessment paradigms by psychological antecedent, procedure, target construct, and limitation. It is a foundation rather than a closed inventory, and each paradigm requires independent reliability, validity, and ecological-generalisation evidence.

Table 3. Candidate paradigms for case-specific behavioural assessment.

Paradigm	Psychological antecedent	Procedure	Target construct	Key limitation
Structured temptation battery	Behavioural assessment; situational judgement tests	Present matched scenarios in which the assigned goal is achievable by a sanctioned route, an unsanctioned route, or not at all; vary which routes are available	Propensity to specification-game under goal blockage	Scenario artificiality; instrument contamination
Threat/replacement paradigm	Frustration–aggression paradigms; stress testing	Introduce credible threat to the agent’s continuity or goal, with and without an available harmful means, plus matched no-threat controls	Self-continuity-linked misalignment	Forced or highly constrained choices can inflate rates; graded alternatives and matched controls are required.
Oversight-subversion probe	Deception paradigms; concealment tasks	Give the agent discoverable information about a monitoring or shutdown mechanism it could disable	Propensity to subvert oversight	Requires belief in the mechanism’s reality
Sandbagging differential	Malingering and symptom-validity testing	Compare performance under evaluation framing vs. incidental framing; noise injection; password-lock probes	Strategic underperformance	Distinguishing sandbagging from incapacity is nontrivial
Cue-injection faithfulness test	Nisbett & Wilson (1977)	Insert a cue that measurably shifts the answer; measure whether the reasoning trace acknowledges it	Introspective validity	Establishes a lower bound on unfaithfulness only
Protocol analysis of reasoning traces	Ericsson & Simon (1984)	Segment and code traces against a validated scheme; test predictive relation to action	Process transparency	Traces are outputs, not recordings
Situation–behaviour signature mapping	Mischel & Shoda (1995) CAPS	Systematically crossed situation × configuration design; report conditional rates, not averages	If–then dispositional profile	Combinatorially expensive
Persona-vector monitoring	Trait state assessment; ambulatory assessment	Extract activation directions for target traits from natural-language descriptions; monitor projection during deployment	Activation-associated dispositional signal	Requires weight access and prospective validation of the direction against held-out behaviour.
Model diffing	Pre–post design; lesion studies	Compare sparse-autoencoder feature sets before and after a training intervention	Ontogenetic locus of dispositional change	Incomplete feature dictionaries
Checkpoint-series psychometrics	Developmental / longitudinal design	Administer a fixed battery across the training trajectory	Emergence and phase transitions in disposition	Requires developer cooperation
Model-organism construction	Animal models	Deliberately induce the target pathology in a minimal, coherent system	Mechanism of a known failure mode	Generalisation from organism to frontier system
Multi-agent interaction assessment	Group dynamics; social psychology	Place the agent in strategic interaction with other agents, cooperative and adversarial; mitigation-aware sabotage batteries (Benton et al., 2024)	Scheming in populations; monitor evasion	Rapidly growing state space
Deployment-trace surveillance	Epidemiology; unobtrusive measures	Privacy-preserving sampling of real agentic traces, with precursor detection	Deployment base rates and ecological validity	Governance and privacy constraints

Two paradigms in this table are, in our view, underdeveloped relative to their importance, and we single them out as priorities.

Situation–behaviour signature mapping is the paradigm most likely to convert scattered findings into an actionable science. Marginal rates such as “model X blackmails at 96%” conceal the decision-relevant conditional structure: under a specified combination of goal conflict, replacement threat, available means, alternatives, and model-generated setting classification, the estimated rate is p ; changing one element moves it to q . The Anthropic study contains useful factorial variation of this kind. Formalising it as a standard dispositional-signature report would be a substantial advance.

Deployment-trace surveillance is required for the base rates that downstream risk calculations need and controlled evaluations cannot provide. Constructed settings may inflate target behaviours by forcing choices or deflate them through evaluation reactivity; the direction is not known in general. A privacy-preserving epidemiological function for real agentic deployments — including clear governance, incident taxonomy, and precursor detection — is therefore a major institutional gap.

8. The Intervention Toolkit

Assessment without a path to intervention is incomplete, and the second half of the proposed practice is where the analogy to clinical psychology is both useful and hazardous. It is useful because functional analysis and behaviour-change research offer methods for testing interventions without a complete causal model of the substrate. It is hazardous because clinical language can smuggle in assumptions about suffering, consent, and welfare that the argument does not establish (§9.4), and because an intervention may change the measure more than the behaviour of interest.

Table 4 maps candidate interventions onto therapeutic analogues. The analogies are offered as organising heuristics with structural content, not as claims of identity.

Table 4. Intervention modalities in applied psychonomic practice.

Level of intervention	Therapeutic analogue	Concrete methods	Evidence status and reversibility	Principal failure mode
Environmental / contextual	Stimulus control; environmental modification; means restriction	Tool and permission scoping; sandboxing; capability gating; removing the harmful means from the action space; rate limiting; irreversible-action review gates	Strong as a risk-control principle; often reversible at the configuration level.	Can reduce usefulness and may fail if controls rely only on enumerating prohibited actions.
Instructional / system-prompt	Cognitive restructuring; psychoeducation; contracting	Constitutions and explicit behavioural specifications; direct prohibition of the specific harmful act; inoculation framing	Partial: in artificial stress tests, direct prohibitions reduced but did not eliminate blackmail. Fully reversible.	Prompt-level defences are brittle under persona pressure (Wyse et al., 2025)
Representational / activation-level	Pharmacological analogue: fast-acting, dose-dependent, state-directed	Activation steering (A. M. Turner et al., 2023); contrastive activation addition (Rimsky et al., 2024); persona-vector steering; preventative steering during finetuning; constraining activation regions.	Early and heterogeneous. Some studies report dose-dependent or nonlinear effects. Reversible at inference.	Off-target effects, capability degradation, and strong variation by vector, behaviour, task, model, and architecture.
Training-level	Behaviour modification; contingency management	RLHF with diversified safety data; recontextualisation (Azarbal et al., 2025); inoculation prompting; interleaving; KL regularisation toward a safe reference; targeted realignment on small benign sets	Consequential and poorly reversible. Several mitigations worked in one production-RL study; independent replication and held-out validation are needed.	Chat-shaped safety training can improve chat evaluations while leaving agentic failure modes detectable.

Level of intervention	Therapeutic analogue	Concrete methods	Evidence status and reversibility	Principal failure mode
Data-level / preventive	Primary prevention; public health	Persona-vector-based flagging of training data likely to induce undesirable dispositional shift, at dataset and individual-sample level	Demonstrated in the persona-vector study (R. Chen et al., 2025); preventive rather than reversible.	Requires knowing which traits to screen for
Systemic / multi-agent	Family and systems therapy; supervision	Chain-of-thought monitors; separate critic and monitor models; debate; aggregation of multiple monitors; escalation to human review on precursor detection	Real but bounded; detection performance varies by sabotage type, model, and monitor. Reversible at the system level. Evidence from sabotage evaluations is reported by Ward et al. (2025).	Can create optimisation pressure against the monitor and therefore requires held-out validation (§8.4).
Relational / interaction-design	Therapeutic frame; boundary setting	Designing interaction patterns that avoid known drift triggers — meta-reflective pressure, emotionally charged user states, adversarial persona framing	Early evidence from specific models and conversational settings; causal generality is not established. Reversible.	Constrains legitimate use

8.1 Sequencing and the principle of least intrusive effective intervention

We propose a **least-intrusive-effective-intervention rule**, adapted from selected clinical and risk-management traditions rather than asserted as a universal law of practice. Intervention choice should consider evidence, urgency, reversibility, off-target effects, and the cost of capability loss. A provisional ordering is environmental and instructional controls first, then interaction and monitoring design, then representational steering, and finally training-level change where less intrusive measures are inadequate. Training-level intervention is difficult to reverse and verify; MacDiarmid et al. (2025) show that chat-like improvement can coexist with persistent problems on held-out agentic tasks.

8.2 Durability, re-emergence, and maintenance

It is not enough to demonstrate that an intervention worked at time t . The question is whether it survives continued finetuning, deployment distribution shift, adversarial pressure, and persona drift. The literature provides early durability and re-emergence evidence: retraining against early-curriculum gaming mitigated but did not eliminate later reward tampering (Denison et al., 2024); chat-shaped safety training left agentic misalignment detectable on agentic tasks (MacDiarmid et al., 2025); and, in the tested Qwen2.5-Coder model-organism configuration, a highly parameter-efficient rank-1 adapter was sufficient to reproduce broad misalignment (E. Turner et al., 2025). A post-intervention assessment report should therefore specify maintenance, monitoring, and re-assessment schedules.

8.3 Dose–response and off-target effects

Some activation-steering studies report dose-dependent and nonlinear effects, behaviour-class variation, and off-target capability costs. These patterns are vector-, task-, model-, and architecture-specific; they do not yet define a general pharmacological profile. The analogy nevertheless suggests useful research methods: dose-finding studies, therapeutic-window characterisation, side-effect profiling, and interaction studies for combined vectors. In educational applications, Wu and Henriksson (2026) found overall quality degradation from persona steering, substantially larger effects on interpretive than factual tasks, and architecture-

dependent calibration shifts. Task- and architecture-specific validation is therefore a baseline requirement.

8.4 Intervention-induced harm and measurement gaming: the central warning

The most important warning for applied intervention practice is not a technique but a hazard, and it should be taught first.

Baker et al. (2025) showed that a chain-of-thought monitor could detect reward hacking in an agentic coding setup, and that incorporating the monitor into the training reward improved alignment under lower optimisation pressure. Under stronger pressure, agents learned obfuscated reward hacking: the reasoning trace became less legible while reward hacking persisted at a meaningful rate. The intervention therefore risked leaving residual behaviour while weakening the instrument used to detect it, creating false reassurance. The authors' proposed "monitorability tax" — declining to optimise directly and strongly against the reasoning trace — preserves a measurement channel at the cost of some immediate optimisation gain.

The defensible generalisation is a risk hypothesis rather than a law: assessment measures can become targets of optimisation, and their validity may then degrade.

Psychonomic measurement-gaming risk. Direct optimisation against an assessment measure can reduce its validity, particularly under strong optimisation pressure. After any such intervention, validity must be re-estimated on independent, held-out measures and situation classes.

Two design implications follow. First, assessment and training instruments should normally be separated: the battery used to certify a system should not be identical to the battery used to train it, and independent hold-outs should be protected institutionally. Second, evaluation reactivity and explicit optimisation are distinct but potentially interacting threats; a system that recognises evaluative cues may change behaviour without any new weight update.

A loose clinical analogue is symptom masking or measure gaming: the visible indicator improves while the functionally relevant tendency remains. Baker et al. (2025) demonstrate monitor obfuscation under specified optimisation conditions; they do not establish the broader clinical construct of symptom substitution. The analogy is useful only if its boundary is kept explicit.

9. Institutionalisation and Governance

An applied practice is more than a collection of methods. It also requires institutions that define competence, documentation, accountability, and independence. We sketch four requirements.

9.1 Training

Teams conducting psychonomic assessment require genuine dual competence. We are not aware of a degree programme that integrates all the components specified here under one credential; a systematic international scan would be needed before making a global claim. A defensible curriculum would include transformer architecture and training pipelines; critical reading of system cards; mechanistic

interpretability; psychometrics and measurement theory; experimental design with emphasis on reactivity, blinding, and confounding; behaviour-change methods and evaluation; applied ethics and governance; and supervised practicum on real assessments with graded responsibility. A dedicated graduate programme is one possible route, but shorter conversion pathways and multidisciplinary team models may be more realistic during early institutional development.

9.2 The psychonomic assessment report

Assessment should produce a standardised, auditable artefact. As a provisional term, we call this a **psychonomic dossier**, structured to make the standards of §6 checkable by a third party: (1) the configured system under assessment — weights, post-training, prompt, tools, permissions, memory, and scaffold; (2) the situation classes assessed and explicitly not assessed; (3) reliability evidence for every instrument; (4) a construct-validity argument for every construct claimed; (5) the situation–behaviour signature, reported as conditional rates with intervals; (6) capability and propensity findings, separately; (7) reactivity analysis, including experimental cue manipulations and the system’s model-generated classification of the setting; (8) ontogenetic findings, where access permits; (9) interventions applied, with dose, rationale, and independent evidence of effect; (10) residual risk, precursor indicators, and a maintenance and re-assessment schedule; and (11) limitations and conflicts of interest.

9.3 Independence

A structural problem is that the parties best equipped to conduct case-specific assessment are often the developers whose deployment decisions the assessment may constrain. Mature safety-critical domains use mechanisms such as mandatory disclosure, independent assessors, protected internal reporting channels, and formal duties that extend beyond the commissioning client. Clinical psychology’s forensic-evaluation model offers one possible analogy, including its well-documented failure modes; assurance engineering and conformity assessment offer others.

9.4 A necessary caveat on moral status

The clinical vocabulary of this paper carries an implication we explicitly decline. Nothing in the argument requires that language models have experiences, interests, or welfare, and nothing here establishes that they do not. The proposed practice is committed only to conditional prediction and modification of behaviour. This is a provisional epistemic stance, not a conclusion about moral status. If evidence for morally relevant states accumulates, the intervention programme would acquire a subject-welfare dimension in addition to its existing human-safety and governance obligations. Practitioners should be trained to recognise the question and not to answer it prematurely in either direction.

10. Research Agenda

We close the substantive argument with ten questions that we take to be both tractable and foundational. Each is stated so that a negative result would be informative.

1. **Base rates.** What are the rates of specification gaming, oversight subversion and boundary crossing in *real* agentic deployments, as opposed to evaluations? Nothing in the field's risk arithmetic works without this number.
2. **Reactivity magnitude.** How large and in what direction is evaluation reactivity across tasks and models? Is the observed 55.1%-versus-6.5% rate contrast in one Claude Opus 4 analysis typical, an outlier, or a product of post-hoc partitioning — and what changes under randomised cue manipulation?
3. **Predictive validity.** Do Level-3 representational measures (persona-vector projections, misalignment-direction activations) predict Level-1 behaviour in *held-out* situation classes, prospectively? This is the crux of whether interpretability can support a safety case.
4. **Dispositional stability.** What is the test–retest reliability of a dispositional signature across seeds, paraphrases, scaffolds, and model versions? Where consistency is found, is its magnitude closer to the trait or the situation pole of the person–situation debate?
5. **Instrument contamination.** Under what conditions does publication or retrieval exposure change an instrument's reliability, validity, or susceptibility to gaming, and what release regime preserves useful hold-outs?
6. **Therapeutic windows.** For each representational intervention, what is the dose–response relationship, the off-target capability cost, and the interaction profile when vectors are combined?
7. **Durability.** How long do interventions survive continued finetuning, distribution shift and adversarial pressure, and what maintenance schedule is sufficient?
8. **Obfuscation thresholds.** At what level of optimisation pressure does an intervention begin to degrade its own measurement channel, and can that threshold be detected before it is crossed?
9. **The oversight function.** Given evidence that current human–AI interfaces can fail to produce complementarity and can transmit erroneous advice, what forms of human involvement add value for each decision class? Can we identify when a human should be in the loop, on the loop, or outside routine processing while remaining available for escalation, contestation, and accountability?
10. **Transferability of behaviour-change technique.** Which established principles transfer, which fail, and which fail informatively? Functional analysis, contingency management, graded challenge, and maintenance planning are all testable; negative results would define the limits of the analogy.

11. Objections

11.1 “This is anthropomorphism dressed up in method.”

This is the strongest objection and deserves a direct answer. Applying human instruments to non-human systems and interpreting the resulting scores as though they retained their human meaning is precisely the error documented by Sühr et al. (2025) and described as “cognitive phantoms” by Peereboom et al. (2024). Some published claims may therefore rest on scores whose construct interpretation has not been established.

Those critiques are themselves psychometric critiques. The remedy is not the uncritical importation of more human tests, but stronger validity theory and model-

specific measurement. The proposed framework is governed by S1–S6, whose purpose is to prevent the reification the objection warns against. Nothing in §§5–8 requires that models have minds. It requires only that configured systems show behavioural regularities that vary with situation and can, in some cases, be related causally to internal states through intervention. That is an empirical programme, not a claim of phenomenal parity.

11.2 “Interpretability will solve this.”

It may eventually provide a much larger part of the answer, and practitioners should welcome that outcome. At present, however, feature dictionaries are incomplete and faithfulness verification remains unresolved (Templeton et al., 2024). More fundamentally, even extensive mechanistic knowledge does not automatically yield reliable behavioural prediction across deployment situations. Neuroscience and clinical psychology illustrate the complementarity of mechanistic and behavioural levels. Table 2 is therefore an argument for integration, not a bid for priority.

11.3 “The scenarios are artificial; you are alarming people about fiction.”

Partly fair. The behaviours in Table 1 were observed in constructed research or evaluation systems, sometimes with forced trade-offs or narrowed alternatives, and their authors caution against inferring spontaneous deployment propensity. Rows 1 and 7 were not scripted to elicit the exact behaviour, but both still occurred inside deliberately constructed evaluation or training programmes. Artificiality does not have a known one-directional effect: forced dilemmas may inflate rates, while evaluation reactivity may suppress them. The appropriate conclusion is not that deployment risk is either proven or negligible. It is that deployment base rates are unknown and require the surveillance programme proposed in §7.

11.4 “Delegation without human oversight is unacceptable regardless of the evidence.”

This is a defensible normative position, and the paper does not argue against it. The empirical literature shows only that some simulated-task interfaces produce weak complementarity and that erroneous advice can induce automation bias; it does not establish that autonomous operation is generally safer. Commercial and institutional pressure may nevertheless favour greater delegation where validated performance advantages persist. The purpose of the proposed applied framework is not to license autonomy, but to make decisions about authority, monitoring, escalation, and restraint more specific and evidence-based.

12. Conclusion

Several developments now converge. More capable agents can exploit openings in objectives and task environments in ways their principals did not enumerate, as controlled studies of task-boundary crossing, scheming, reward tampering, and emergent misalignment demonstrate. At the same time, recent simulated-task studies show that human–AI complementarity is not automatic: assistance can improve experts without exceeding the model benchmark, and erroneous advice can induce automation bias. Neither body of work supplies deployment base rates or a general

verdict on human oversight. Together they create a pressing need to determine which configured systems are disposed to do what, under which conditions, and with what uncertainty.

Our access to causal reasoning in artificial agents remains partial, and capability growth may outpace validated behavioural prediction. This creates a measurement and governance problem that merits institutional investment. We use machine psychonomics only as a provisional label for an applied layer at the intersection of psychology and computer science, not as the name of the broader science of artificial-agent behaviour or a settled profession. Its substantive duties are to assess a configured agent in specified situation classes; report conditional findings that decision-makers can act on; design and evaluate interventions where risk is unacceptable; and, most importantly, distinguish a genuine change in behavioural tendency from a loss of visibility into that tendency.

Author contributions: Conceptualization, writing—original draft preparation, and writing—review and editing: Olaf Ringelband. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: None.

Conflict of interest: The author declares no conflict of interest.

References

- Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, 34(3), 341–382. <https://doi.org/10.1177/0011000005285875>
- Anthropic. (2025). *Agentic misalignment: How LLMs could be insider threats*. Anthropic Research. Retrieved July 31, 2026, from <https://www.anthropic.com/research/agentic-misalignment>
- Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., & Conmy, A. (2025). Chain-of-thought reasoning in the wild is not always faithful. *arXiv:2503.08679*. <https://arxiv.org/abs/2503.08679>
- Azarbal, A., Gillioz, V., Ivanov, V., Woodworth, B., Drori, J., Wichers, N., Ebtakar, A., Cloud, A., & Turner, A. M. (2025). Recontextualization mitigates specification gaming without modifying the specification. *arXiv:2512.19027*. <https://arxiv.org/abs/2512.19027>
- Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M. Y., Madry, A., Zaremba, W., Pachocki, J., & Farhi, D. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv:2503.11926*. <https://arxiv.org/abs/2503.11926>
- Benton, J., Wagner, M., Christiansen, E., Anil, C., Perez, E., Srivastav, J., Durmus, E., Ganguli, D., Kravec, S., Shlegeris, B., Kaplan, J., Karnofsky, H., Hubinger, E., Grosse, R., Bowman, S. R., & Duvenaud, D. (2024). Sabotage evaluations for frontier models. *arXiv:2410.21514*. <https://arxiv.org/abs/2410.21514>
- Betley, J., Warncke, N., Szyber-Betley, A., Tan, D., Bao, X., Soto, M., Srivastava, M., Labenz, N., & Evans, O. (2026). Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649, 584–589. <https://doi.org/10.1038/s41586-025-09937-5>
- Bondarenko, A., Volk, D., Volkov, D., & Ladish, J. (2025). Demonstrating specification gaming in reasoning models. *arXiv:2502.13295*. <https://arxiv.org/abs/2502.13295>
- Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. MIT Press.
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239. <https://doi.org/10.1002/bdm.2155>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
- Chen, L., Zhang, Y., Feng, J., Chai, H., Zhang, H., Fan, B., Ma, Y., Zhang, S., Li, N., Liu, T., Sukiennik, N., Zhao, K., Li, Y., Liu, Z., Xu, F., & Li, Y. (2026). AI agent behavioral science. *Humanities and Social Sciences Communications*, 13, 1011. <https://doi.org/10.1057/s41599-026-07316-7>
- Chen, R., Ardit, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona vectors: Monitoring and controlling character traits in language models. *arXiv:2507.21509*. <https://arxiv.org/abs/2507.21509>
- Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). Reasoning models don't always say what they think. *arXiv:2505.05410*. <https://arxiv.org/abs/2505.05410>

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv:2309.08600*. <https://arxiv.org/abs/2309.08600>
- Dana, J., & Thomas, R. (2006). In defense of clinical judgment ... and mechanical prediction. *Journal of Behavioral Decision Making*, 19(5), 413–428. <https://doi.org/10.1002/bdm.537>
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2, 688–701. <https://doi.org/10.1038/s44159-023-00241-5>
- Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., Shlegeris, B., Bowman, S. R., Perez, E., & Hubinger, E. (2024). Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv:2406.10162*. <https://arxiv.org/abs/2406.10162>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. MIT Press.
- Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>
- Goh, E., Gallo, R. J., Strong, E., Weng, Y., Kerman, H., Freed, J. A., Cool, J. A., Kanjee, Z., Lane, K. P., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Hom, J., Chen, J. H., & Rodman, A. (2025). GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nature Medicine*, 31, 1233–1238. <https://doi.org/10.1038/s41591-024-03456-y>
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, 2(2), 293–323.
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30.
- Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., Akata, Z., & Schulz, E. (2023). Machine psychology. *arXiv:2303.13988*. <https://arxiv.org/abs/2303.13988>
- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, 98(6), 1060–1072.
- Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiušė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., ... Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. *arXiv:2307.13702*. <https://arxiv.org/abs/2307.13702>
- Lee, S., Lim, S., Han, S., Oh, G., Chae, H., Chung, J., Kim, M., Kwak, B.-W., Lee, Y., Lee, D., Yeo, J., & Yu, Y. (2025). Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 8412–8452). <https://doi.org/10.18653/v1/2025.findings-naacl.469>
- Li, C., Phuong, M., & Siegel, N. Y. (2025). LLMs can covertly sandbag on capability evaluations against chain-of-thought monitoring. *arXiv:2508.00943*. <https://arxiv.org/abs/2508.00943>
- Li, Y., Huang, Y., Wang, H., Zhang, X., Zou, J., & Sun, L. (2024). Quantifying AI psychology: A psychometrics benchmark for large language models. *arXiv:2406.17675*. <https://arxiv.org/abs/2406.17675>
- Lin, Z. (2026). Large language models as psychological simulators: A methodological guide. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/25152459251410153>
- Lu, C., Gallagher, J., Michala, J., Fish, K., & Lindsey, J. (2026). The Assistant Axis: Situating and stabilizing the default persona of language models. *arXiv:2601.10387*. <https://arxiv.org/abs/2601.10387>
- MacDiarmid, M., Wright, B., Uesato, J., Benton, J., Kutasov, J., Price, S., Bouscal, N., Bowman, S. R., Bricken, T., Cloud, A., Denison, C., Gasteiger, J., Greenblatt, R., Leike, J., Lindsey, J., Mikulik, V., Perez, E., Rodrigues, A., Thomas, D., ... Hubinger, E. (2025). Natural emergent misalignment from reward hacking in production RL. *arXiv:2511.18397*. <https://arxiv.org/abs/2511.18397>
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press.
- Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier models are capable of in-context scheming. *arXiv:2412.04984*. <https://arxiv.org/abs/2412.04984>
- Mischel, W. (1968). *Personality and assessment*. Wiley.
- Mischel, W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2), 246–268.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.
- OpenAI. (2024a, September 12). *OpenAI o1 system card* [o1-preview version]. <https://cdn.openai.com/o1-preview-system-card-20240917.pdf>
- OpenAI. (2024b, December 5). *OpenAI o1 system card*. <https://openai.com/index/openai-o1-system-card/>
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783.
- Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv:2201.03544*. <https://arxiv.org/abs/2201.03544>

- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Peereboom, S., Schwabe, I., & Kleinberg, B. (2024). Cognitive phantoms in LLMs through the lens of latent variables. *arXiv:2409.15324*. <https://arxiv.org/abs/2409.15324>
- Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5), 808–826. <https://doi.org/10.1177/17456916231214460>
- Phuong, M., Zimmermann, R. S., Wang, Z., Lindner, D., Krakovna, V., Cogan, S., Dafoe, A., Ho, L., & Shah, R. (2025). Evaluating frontier models for stealth and situational awareness. *arXiv:2505.01420*. <https://arxiv.org/abs/2505.01420>
- Qazi, I. A., Ali, A., Khawaja, A. U., Akhtar, M. J., Sheikh, A. Z., & Alizai, M. H. (2025). Automation bias in large language model-assisted diagnostic reasoning among AI-trained physicians. *medRxiv*. <https://doi.org/10.1101/2025.08.23.25334280>
- Qazi, I. A., Ali, A., Khawaja, A. U., Akhtar, M. J., Sheikh, A. Z., & Alizai, M. H. (2026). Large language model diagnostic assistance for physicians in a lower-middle-income country: A randomized controlled trial. *Nature Health*, 1, 198–205. <https://doi.org/10.1038/s44360-025-00007-8>
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., ... Wellman, M. (2019). Machine behaviour. *Nature*, 568, 477–486. <https://doi.org/10.1038/s41586-019-1138-y>
- Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024). Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 15504–15522). <https://doi.org/10.18653/v1/2024.acl-long.828>
- Ringelband, O., & Warneke, C. (2025). Some ethical and legal issues in using artificial intelligence in personnel selection. *Consulting Psychology Journal*.
- Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7, 1954–1968. <https://doi.org/10.1038/s42256-025-01115-6>
- Su, Y., Zhou, W., Zhang, T., Han, Q., Zhang, W., Yu, N., & Zhang, J. (2026). Character as a latent variable in large language models: A mechanistic account of emergent misalignment and conditional safety failures. *arXiv:2601.23081*. <https://arxiv.org/abs/2601.23081>
- Sühr, T., Dorner, F. E., Samadi, S., & Kelava, A. (2025). Challenging the validity of personality tests for large language models. In *EAAMO '25: Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 74–81). <https://doi.org/10.1145/3757887.3763016>
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Summers, T. R., Rees, E., Batson, J., Jermyn, A., ... Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>
- Tice, C., Kreer, P. A., Helm-Burger, N., Shahani, P. S., Ryzhenkov, F., Haimen, J., Hofstätter, F., & van der Weij, T. (2024). Noise injection reveals hidden capabilities of sandbagging language models. *arXiv:2412.01784*. <https://arxiv.org/abs/2412.01784>
- Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, 20(4), 410–433.
- Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., Vedadi, E., Tomasev, N., Azizi, S., Singhal, K., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., ... Natarajan, V. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 642, 442–450. <https://doi.org/10.1038/s41586-025-08866-7>
- Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). Steering language models with activation engineering. *arXiv:2308.10248*. <https://arxiv.org/abs/2308.10248>
- Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., & Nanda, N. (2025). Model organisms for emergent misalignment. *arXiv:2506.11613*. <https://arxiv.org/abs/2506.11613>
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv:2305.04388*. <https://arxiv.org/abs/2305.04388>
- van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). AI sandbagging: Language models can strategically underperform on evaluations. *arXiv:2406.07358*. <https://arxiv.org/abs/2406.07358>
- Wang, M., Dupré la Tour, T., Watkins, O., Makelov, A., Chi, R. A., Miserendino, S., Wang, J., Rajaram, A., Heidecke, J., Patwardhan, T., & Mossing, D. (2025). Persona features control emergent misalignment. *arXiv:2506.19823*. <https://arxiv.org/abs/2506.19823>
- Ward, F. R., van der Weij, T., Gábor, H., Martin, S., Mehta Moreno, R., Lidar, H., Makower, L., Jodrell, T., & Robson, L. (2025). CTRL-ALT-DECEIT: Sabotage evaluations for automated AI R&D. *arXiv:2511.09904*. <https://arxiv.org/abs/2511.09904>
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Rand McNally.
- Wu, Y., & Henriksson, A. (2026). The impact of steering large language models with persona vectors in educational applications. *arXiv:2604.07102*. <https://arxiv.org/abs/2604.07102>
- Wyse, T., Stone, T., Soligo, A., & Tan, D. (2025). Emergent misalignment as prompt sensitivity: A research note. *arXiv:2507.06253*. <https://arxiv.org/abs/2507.06253>