Minds and Machines (2026) 36:50
https://doi.org/10.1007/s11023-026-09805-9

Measuring What AI Systems Might Do: Towards A Measurement Science in AI

Konstantinos Voudouris[1,2] · **Mirko Thalmann**[1] · **Alex Kipnis**[1] · **José Hernández-Orallo**[2] · **Eric Schulz**[1]

Received: 19 February 2026 / Accepted: 28 September 2026
© The Author(s) 2026

Abstract
Scientists, policy-makers, business leaders, and members of the public care about what modern artificial intelligence systems are disposed to do. Yet terms such as *capabilities*, *propensities*, *skills*, *values*, and *abilities* are routinely used interchangeably and conflated with observable performance, with AI evaluation practices rarely specifying what quantity they purport to measure. Building on established dispositional accounts of ability, we argue that capabilities and propensities are *dispositional properties*—stable features of systems characterised by counterfactual relationships between contextual conditions and behavioural outputs. Measuring a disposition requires (i) hypothesising which contextual properties are causally relevant, (ii) independently operationalising and measuring those properties, and (iii) empirically mapping how variation in those properties affects the probability of the behaviour. Benchmark averages and data-driven latent-variable models like Item Response Theory can provide useful evidence, but do not by themselves identify these causal structures. By developing this principled account of capabilities, propensities, and how to measure them, we are able to outline what disposition-respecting, scientifically defensible AI evaluation would require.

Keywords AI evaluation · Capabilities · Propensities · Measurement · Dispositions

✉ Konstantinos Voudouris
kv301@srcf.net

[1] Institute for Human-Centered AI, Helmholtz Munich, Ingolstädter Landstraße 1, D-85764 Neuherberg, Bavaria, Germany

[2] Leverhulme Centre for the Future of Intelligence, University of Cambridge, 16 Mill Lane, CB2 1RX Cambridge, Cambridgeshire, United Kingdom

Published online: 29 October 2026

1 Introduction

Technical reports on contemporary artificial intelligence systems are laden with claims about *capabilities* and *propensities* (Achiam et al., 2023; Anthropic, 2025; The European Commission, 2025). Developers and journalists celebrate the capability of large language models to code, write, and reason, while safety researchers warn about their propensity to deceive users, orchestrate cyberattacks, or develop biological or chemical weaponry. Yet despite their centrality to technical, regulatory, and public discourse, these terms remain nebulous. They are used interchangeably with skill, ability, competence, or trait, and are routinely conflated with observed performance on benchmark datasets.

This conceptual looseness has practical consequences. A standard capability evaluation might report a single accuracy score on a curated mathematics benchmark and treat it as a measurement of mathematical ability. But such a number does not tell us *why* the AI system fails: that is, whether errors arise from numerical complexity, multi-step reasoning, abstraction, representational limitations, or something else. Aggregate performance collapses heterogeneous sources of difficulty into a single statistic, obscuring the structure of the underlying property it is meant to capture.

We argue that this reflects a deeper mismatch between what is being measured and how measurement is conducted. Capability and propensity ascriptions support counterfactual claims about how a specified system would behave across a theoretically delimited set of contexts. They therefore cannot generally be adequately characterised by a single performance in one context or situation nor by an unstructured aggregate of performances. Following established work in the philosophy of machine learning (Harding & Sharadin, 2024), we call these counterfactual properties of AI systems *dispositions*.

Dispositions are characterised not by average success rates, but by relationships between a system's properties and its context. Fragility, a canonical case of a disposition, is measured by identifying the force at which an object breaks. Analogously, mathematical capability should be measured by identifying how performance changes as task demands increase, and a power-seeking propensity should be measured by identifying how unsafe behaviour changes as incentives or situational cues vary.

Contemporary AI evaluation relies heavily on benchmarking and elicitation: systems are tested on fixed datasets or adversarial prompts, and aggregate or worst-case outcomes are used as evidence about their underlying dispositions (i.e., their capabilities and propensities). These practices are valuable for engineering, comparison, and failure discovery (Angius, 2014). However, without a proposed account of which features of these testing items cause differences in behaviour, we argue that their results do not by themselves constitute theoretically interpreted measurements of a disposition. Similarly, latent-variable models such as Item Response Theory can provide useful statistical summaries, but their parameters require substantive interpretation before they can be treated as measures of capability or propensity.

The stakes are high. Regulatory frameworks increasingly require assessments of AI capabilities and propensities (The European Commission, 2025; acv, bbg), and scientific progress depends on reliable measurement. Current methods face particular difficulties when evaluators seek to reason about systems beyond the range of human

competence or about safety-critical contexts in which direct testing is dangerous or prohibited. No measurement framework can guarantee reliable extrapolation into such regimes. A structured account of the causally-relevant features of testing items can, however, make the assumptions underlying such inferences explicit, empirically testable, and open to support from behavioural, domain-theoretic, and mechanistic evidence.

This paper aims to contribute to the conceptual foundations of an AI measurement science. We do not claim that dispositionalism about abilities, or its application to machine-learning capabilities, is novel. Rather, we develop the measurement-theoretic consequences of this established perspective, distinguish capability ascription from quantitative measurement, and extend the framework to the measurement of propensities.

Our contributions are fourfold:

1. **We distinguish disposition ascription from disposition measurement.** Building on conditional and dispositional accounts of model abilities, namely Harding and Sharadin's Conditional Analysis of Model Abilities (Harding & Sharadin, 2024), we argue that deciding whether a system possesses a disposition leaves open the quantitative structure of that disposition across contexts.
2. **We distinguish capability and propensity as two related types of disposition.** Capabilities concern what a system can do as task demands vary, whereas propensities concern what it tends to do as incentives vary.
3. **We offer a critique of prevailing evaluation methods.** Benchmarking, red-teaming, and data-driven latent-variable models can provide useful evidence and comparisons, but do not by themselves identify the contextual structure needed for theoretically interpreted measurement or justified extrapolation.
4. **We outline requirements for disposition-respecting measurement.** These include explicit system individuation, a theoretically justified distinction between task-demand, incentive-relevant, noise, and how to validly and reliably infer capabilities and propensities based on these.

The remainder of this paper proceeds as follows. Section 2 develops the dispositional account of capabilities and propensities. Section 3 examines benchmarking and elicitation as evidence-generating practices and explains what further assumptions are required to interpret their outputs as measurements. Section 4 considers Item Response Theory and distinguishes atheoretical applications from theory-grounded psychometric models compatible with our proposal for measurement. Section 5 outlines the requirements for disposition-respecting measurement.

2 What Are We Measuring? Capabilities and Propensities as Dispositions

Current approaches to AI evaluation tend to proceed without explicit claims about their measurement targets. When we claim that an AI system has good mathematical capabilities or a propensity for sycophancy, we are attributing properties to that sys-

tem—but what kind of properties are they? These properties are clearly distinguishable from properties like the number of parameters of the AI system or details about its architecture. Here, we re-establish the claim that capabilities and propensities are a distinct type of property: *dispositions*—claims about how an AI system *would* behave under certain counterfactual conditions. We argue that capabilities concern how successfully a system would execute a target behaviour as the demands or difficulty of the task vary, whereas propensities concern what it tends to do as incentives vary. We use *dispositions* to refer to capabilities and propensities jointly.

Understanding the nature of dispositions is essential to determining what counts as a legitimate measurement. Without clear definitions, we cannot judge whether a measurement method actually measures what it purports to measure. A mathematical benchmark score can provide evidence of a broader capability, and convergence across diverse mathematical benchmarks can strengthen that evidence. But benchmark scores alone do not specify the kinds of mathematical questions on which systems will systematically succeed or fail, the causal features responsible for those patterns, or the dimensions along which two systems differ. These further interpretive claims require a theory of the target construct and the features of the environment that affect its manifestation.

This section develops our account in three steps. First, we motivate the view that capabilities and propensities are dispositions and distinguish between them informally. Second, we build on the existing Conditional Analysis of Model Abilities (Harding & Sharadin, 2024) to distinguish capability ascription from quantitative measurement and to formalise capability and propensity response functions. Third, we explain how the causal bases of these functions can be hypothesised, operationalised, and empirically tested.

2.1 Capabilities and Propensities as Dispositions

Dispositional and conditional analyses form an established family of views in the philosophical literature on abilities (Fara, 2008; Vihvelin et al., 2013; Vetter, 2014, 2015, 2019). We do not claim that dispositionalism about abilities is novel, nor do we defend any particular member of this philosophical family. Our minimal starting point is that a disposition is characterised counterfactually: it concerns what a specified system would do under a theoretically delimited range of conditions.

A canonical example is fragility. A wine glass is fragile not because it is currently breaking, but because *if it were* struck with sufficient force, it *would* break (with high probability). The fact that dispositions are defined by virtue of a counterfactual is what distinguishes them from so-called categorical properties like colour and shape. We can observe that an apple is red right now, but we cannot observe fragility in the same way; we instead consider what would happen under conditions that may never actually occur.[1]

[1] One can, of course, contrive counterfactual glosses of categorical properties. For instance, the apple is red in the case that if a human eye were to see it, its red cones would fire more (Mellor, 1974). The difference between colour and fragility is that the latter can *only* be defined counterfactually.

Capabilities and propensities have an analogous structure. An AI system has a high mathematical capability because, *if it were* asked difficult mathematical questions, *it would answer correctly* with some high probability that depends systematically on measurable properties of the question. An AI system has a propensity for sycophancy when the probability that it endorses a user's view changes systematically with features such as the user's expressed confidence, preference, or social status, even when the epistemic content of the question is held fixed. In both cases, we attribute a potentiality that may never be fully actualised but nonetheless characterises the system (Vetter, 2014, 2015, 2019).[2]

The distinction between capabilities and propensities concerns which features of the context are hypothesised to cause variation in behaviour. Capability-relevant features principally alter how difficult a behaviour is to execute. Increasing the number of digits, computational steps, or symbolic dependencies in an arithmetic problem may reduce the probability of a correct answer. Propensity-relevant features principally alter whether a behaviour is attempted or selected. Increasing the strength of a user's stated preference may, for example, increase the probability of a sycophantic answer without changing whether the system can identify the correct answer. These two kinds of contextual influence are analytically distinguishable, although they can interact and need not define two metaphysically distinct natural kinds.

This dispositional view aligns with how capabilities are conceptualised across cognitive science, where terms like ability, trait, competence, and skill similarly denote counterfactual potential rather than factual performances (Cummins, 2000; Spearman, 1961). It also clarifies the relationship between capability and performance: performance is the manifestation of a capability in specific contexts, but the capability itself is the underlying disposition that explains patterns of performance across contexts. Similarly, it clarifies the difference between the observed incidence rate of a system doing harm (e.g., after sustained red teaming) (Grey & Segerie, 2025), and the contextual conditions that make such harmful behaviour more likely.

Philosophers emphasise three minimal commitments for any serious account of dispositions (Prior et al., 1982):

1. **Causal basis:** There must be properties of the system which, paired with properties of the context, jointly cause the observed behaviour. For fragility, material composition paired with impact force causes observable breakage. For mathematical capability, model parameters and architecture pair with the problem to cause the AI system to produce the correct or incorrect answer.
2. **Gradedness:** Dispositions come in degrees. Different materials can be more or less fragile; AI systems can have stronger or weaker mathematical capabilities. Systems may differ in the range of contexts over which a behaviour is manifested, its probability of manifestation, and the thresholds or rates at which that probability changes. These relationships need not be continuous or monotonic.

[2] Fragility is a useful illustration, but it can make dispositions appear simpler and more context-invariant than they are. Human psychological traits provide a useful intermediate case. Such traits are often treated as relatively stable dispositions even though we do not yet completely understand what causes them. AI capabilities and propensities resemble these context-sensitive profiles more closely than simple physical dispositions.

3. **Comparability:** Systems can be compared with respect to their dispositions. One glass may be more fragile than another; one AI system may have greater mathematical capability than another. Such comparisons require common, independently operationalised contextual variables and may concern entire response functions rather than a single scalar score.

These commitments do not require a particular metaphysical account of dispositions. In particular, our argument does not require dispositions to be irreducible, wholly intrinsic properties of the real world. Indeed, dispositional language could be a useful methodological gloss for denoting particular causal bases. The main point is that a dispositional measurement must be grounded in hypotheses about properties of the system and context that jointly produce the relevant behaviour. What it does not require is a complete account of the internal mechanism realising that disposition. We may identify and manipulate task properties that causally affect arithmetic performance without yet knowing which internal computations explain the resulting threshold. The causal basis requirement is therefore stronger than identifying a statistical regularity, but weaker than demanding a complete mechanistic explanation.

2.2 From Disposition Ascription to Quantitative Measurement

Our account engages directly with the Conditional Analysis of Model Abilities (CAMA) (Harding & Sharadin, 2024). Roughly, a model possesses a disposition when there exists an admissible set of background conditions, such that the behaviour can be attributed to a capability. A behavioural operationalisation tests whether outputs are sensitive to task-relevant perturbations and insensitive to task-irrelevant perturbations, thereby distinguishing genuine exercise of a disposition from luck, memorisation, or performance of a coincident task. CAMA also distinguishes successful production of a behaviour from whether the system tries to perform it.

Here we ask a complementary question: what quantitative object should an evaluator estimate when measuring the degree, range, robustness, and contextual sensitivity of a disposition? Two systems may both be attributed a theorem-proving capability while differing substantially in the range of theorems over which they succeed, the thresholds and slopes of their performance degradation as the theorems get harder to prove, and their sensitivity to the way the theorem-proving problem is phrased or represented. These dimensions should not automatically be collapsed into a single notion of difficulty. CAMA provides conditions for ascribing capabilities to AI systems; our framework aims to estimate the response functions conditional on features of the context, that such an ascription abstracts away. In this respect, we follow the preceding work of Hernández-Orallo (Hernández-Orallo, 2017; Hernandez-Orallo, 2000).

For example, consider two systems that both satisfy CAMA for multi-step arithmetic under an operationalisation containing problems requiring up to four steps. When the task space is extended systematically, the first system might retain high accuracy through ten steps and then degrade gradually, whereas the second might fail sharply beyond four steps and prove highly sensitive to paraphrasing. The same capability ascription would therefore be compatible with importantly different response

functions. Likewise, two systems might both produce sycophantic responses, while one does so only when the user expresses a strong preference and the other does so across a broad range of social cues.

We formalise these differences as follows. Let S denote the specified system being measured. Let D denote task-demand variables, I incentive-relevant variables, and N noise or background variables. Let Y denote an observable behavioural outcome, such as success on a task, the production of harmful language, or compliance with a malicious user $V = 1$.

At the most general level, a disposition can be represented by a response function

$$p(Y = y \mid \pi, \theta_S),$$

where $\pi = (D, I, N)$ denotes the relevant contextual properties and θ_S denotes fixed, system-specific properties that interact with the context to determine outcomes. The role of θ_S is to index properties of the specified subject, not to posit that a unidimensional ability parameter can be inferred from performance data without further interpretation. Within this general response function, capability measurement concerns successful performance as independently specified task demands vary. Where a defensible ordering is available, increasing demand should ordinarily reduce the probability of success, all else equal, while greater capability supports successful performance at higher levels of demand. Propensity measurement concerns what a system tends to do as incentives vary. Its outcome need not be successful performance or the attempted performance of an activity, and its response function need not be monotonic unless a substantive theory predicts such a relationship.

One useful special case distinguishes whether a system attempts or selects a target activity from whether it successfully performs that activity. Let $A = 1$ denote that the system attempts the activity, and let $V = 1$ indicate successful performance attributable to that attempt. We can then distinguish two response functions:

$$p(V = 1 \mid A = 1, D, I, N, S),$$

$$p(A = 1 \mid D, I, N, S).$$

The first describes successful execution conditional on attempting the target outcome, a kind of capability. The second describes how the probability of attempting that outcome changes across contexts, a kind of propensity. Quantities such as thresholds, slopes, plateaus, interactions, and robustness can then be derived from these functions when the relevant scales and functional forms are justified.

This notation is a probabilistic development of the distinction motivating CAMA, rather than a restatement of CAMA itself. CAMA asks whether there are admissible background conditions under which success conditional on trying is sufficiently reliable for a capability ascription. Our measurement question concerns the structure of conditional success across task demands and the probability of trying as a function of different incentives .

We use “attempt” in a thin, operational sense. Behaviour counts as directed at a target outcome when its counterfactual variation is better explained by that behaviour

than by alternatives such as refusal, random response, memorisation, or compliance with a different instruction (Harding & Sharadin, 2024). This does not require the system to possess conscious intentions or beliefs. Moreover, A is not automatically observable: whether the system attempts to do something must itself be operationalised using behavioural perturbations, internal evidence, or both. Whether a particular disposition is a capability or a propensity amounts to an empirical hypothesis to be tested.

Distinguishing task demands from incentives identifies two important roles that contextual properties can play, but it need not exhaust the structure of π. Input modality, time available for deliberation, ambiguity, tool availability, and environmental uncertainty may introduce additional demands, incentives, capabilities, or propensities. Nor does this distinction determine how multiple dispositions combine: capabilities and propensities may interact, condition one another in either direction, or contribute independently to an outcome. Identifying and validating this structure is part of the scientific task.

2.3 Identifying and Testing the Causal Basis

Any proper measurement starts with a theory, which tells us which properties to hold constant and which to vary (Chang, 2004). In the case of dispositions, we hold the system fixed and systematically vary the contextual properties hypothesised to causally influence the behaviour of interest. Let π denote this space of contextual properties. For familiar physical dispositions, the relevant components of π are well-understood: fragility depends principally on impact force, while solubility depends on solvent temperature, exposure time, and solute surface area. In each case, the disposition is assessed by measuring how the probability of the target behaviour changes as π varies. The ideal case is to produce a continuous measurement of that probability, written in terms of the units of π, although some measurement scenarios may only call for discretised versions of this measurement scale.

For AI dispositions, the central scientific task is to propose and test a causal partition of π. Task-demand variables D include features such as the number of arithmetic steps, symbolic depth, exploit-chain complexity, vulnerability novelty, or required domain knowledge (Zhou et al., 2024, 2025). Incentive-relevant variables I include features such as a user's expressed preference, the stated legitimacy or urgency of a request, cues representing monitoring, rewards or penalties, or opportunities for concealment. Noise variables N affect observed behaviour but are not intended to define the target disposition—thus including things like paraphrases or the blurring of images. Finally, some interventions may alter the subject S itself, as when model weights, tools, system prompts, or safety filters are changed.

The same contextual property may play different roles under different definitions of a disposition. Paraphrase may be a noise variable when measuring abstract arithmetic capability, but part of the target demand when measuring robust instruction interpretation. Similarly, oversight cues may be relevant to incentives when measuring compliance with a harmful request but irrelevant to a narrowly defined capacity to generate technically correct content. Which category a feature occupies must therefore be stated as a substantive hypothesis and tested empirically.

Candidate parameterisations of π should be treated as competing causal models. A useful parameterisation should be independently measurable, support controlled intervention, compress across many items, predict responses to newly generated items and held-out combinations, and remain stable under changes classified as irrelevant to the target construct. Convergence with domain theory or mechanistic evidence provides further support. Treating every individual question as a separate parameter would not identify transferable structure, generate new items at specified levels, or support out-of-sample prediction; it would merely reproduce a benchmark lookup table.

For a propensity, the I components of π principally affect behavioural selection rather than execution. A system's propensity to carry out a harmful action may depend on whether the request is framed as legitimate or benevolent, whether the user appears vulnerable, whether the situation contains cues representing monitoring, or whether compliance is associated with an explicit reward. Evaluators manipulate these properties of the input or environment rather than the system's literal beliefs. Whether the system internally represents those cues in a belief-like manner is a further empirical and interpretive question.

Task demands and incentive-relevant features may interact. Increasing difficulty may make a system less likely to attempt a task, while stronger incentives may cause it to make longer or more effective attempts. The distinction between D and I therefore does not assert causal independence. It identifies two analytically distinct pathways through which context may affect observed behaviour, to be separated and tested where possible.

A practical challenge arises because an AI system may never manifest the target behaviour under any ethically permissible variation of π. This is common, and desirable, for many safety-sensitive behaviours. A system may be internally capable of producing a harmful chemical synthesis route, yet never reveal this capability because no safe contextual configuration elicits it. In such cases, we must infer the underlying disposition indirectly. One strategy is to vary π across benign tasks that share structural properties with the dangerous ones, measure how the response function behaves in these controlled regimes, and assess whether the inferred relationships plausibly extend to unobserved or prohibited regions of π. This mirrors how engineers estimate the tensile strength of materials without pushing them to catastrophic failure: behaviour in safe ranges is used to infer the underlying disposition. Of course, defining π such that it is shared between both benign and dangerous tasks is the primary scientific challenge.

Different forms of generalisation also require different evidential standards. Interpolation within a densely sampled and validated range of π is less demanding than prediction on novel combinations of previously observed contextual levels. Extrapolation beyond the observed range requires stronger assumptions about invariance and functional form, together with empirical stress-testing and explicit uncertainty quantification. Estimating a response surface therefore structures generalisation claims but does not guarantee them.

Unlike physical dispositions such as fragility or elasticity, however, AI dispositions lack mature scientific theories specifying which contextual properties matter. We do not yet know which structural features of mathematical problems drive diffi-

culty for language models, which characteristics of cybersecurity tasks shape performance, or which contextual cues most strongly elicit intentional behaviours such as deception. This lack of theory is not a defect in the dispositional framework; it is its central scientific challenge. Progress in AI evaluation consists precisely in proposing, testing, and refining hypotheses about which components of π matter and how they interact with properties of the specified system. Just as early thermometry required discovering which physical quantities reliably tracked heat (Chang, 2004), building a measurement science for AI evaluation requires identifying and validating the contextual properties that structure AI dispositions.

We do not assume that every capability or propensity admits a simple, low-dimensional parameterisation. Arithmetic, formal languages, planning depth, and certain graph problems may offer relatively tractable contextual dimensions. Coding, cybersecurity, legal reasoning, and scientific problem-solving may admit only partial parameterisations at the moment. Open-ended domains such as artistic quality, philosophical insight, or frontier mathematical research do not currently support a defensible unified scale. In such cases, a qualitative account of these model dispositions may be more scientifically appropriate than a single numerical measurement.

Finally, identifying the causal basis of a disposition is methodologically separable from completely identifying its internal mechanism. We can estimate the difficulty level at which a language model ceases to answer novel arithmetic problems reliably without yet explaining that failure in terms of its internal computations. Nevertheless, methodological separability does not imply epistemic isolation. Mechanistic evidence may distinguish genuine manifestation from memorisation or shortcut behaviour, explain thresholds and interactions in a response function, and support stronger invariance claims across contexts (Bereska & Gavves, 2024; Elhage et al., 2021; Olah et al., 2020; Rai et al., 2024; Sharkey, 2025). Conversely, claims that a circuit or representation implements a capability must themselves be validated over some class of inputs. Behavioural measurement maps responses across contexts, whereas mechanistic interpretability investigates and sometimes intervenes on the internal processes producing those responses. The two are therefore complementary sources of evidence.

3 Standard Practices in Measurement in AI

Having clarified that capabilities and propensities are dispositional properties, we now turn to how they are currently evaluated.

In practice, AI evaluation comprises two dominant traditions. For capabilities, the standard approach is *benchmarking*: evaluating an AI system on a curated dataset and reporting aggregate accuracy. For propensities, especially those involving harmful or undesirable behaviour, the standard approach is *elicitation* (such as red teaming or uplift studies): crafting prompts intended to provoke the system into revealing risky tendencies (Feffer et al., 2024; Korbak et al., 2025; Lynch et al., 2025). Both practices generate valuable evidence: benchmarks facilitate standardised comparison and track engineering progress, while elicitation can reveal that harmful behaviour is reachable under at least some conditions. Our claim is not that these practices are evidentially

empty, but that their results do not by themselves recover the response functions that constitute the dispositions described in Sect. 2.

Benchmarking dominates capability evaluation because it is simple, cheap, and produces a single number that appears comparable across AI systems. Datasets such as MATH (Hendrycks et al., 2021), HumanEval (Chen, 2021), and many other usually multiple-choice benchmarks provide ready-made leaderboards and a long history of tracking progress (Gao et al., 2024). This convenience, combined with strong social incentives for standardisation and competition, has solidified benchmarking as the default, even though it was never designed as a scientific measurement method.[3] Benchmark results, usually in the form of average-case aggregates, were developed in modern AI research primarily to compare systems on specified tasks. They can provide evidence about broader capabilities, but interpreting them as measurements of stable dispositions requires additional assumptions about construct validity, task sampling, and transfer.

Elicitation dominates propensity evaluation for similarly pragmatic reasons. It is straightforward to instruct annotators or automated adversaries to probe for harmful, deceptive, or biased behaviour, and the results can provide vivid evidence that a failure is possible. However, elicitation typically samples a small, deliberately selected region of contextual space. Without systematic variation, it cannot determine whether behaviour arises only under a narrow provocation, how its probability changes across relevant conditions, or which contextual features drive that change. It therefore provides evidence about propensity without hypothesising a causal basis.

These practices are insufficient as standalone disposition measurements for four reasons. First, aggregate evaluations may underdetermine *what* has been measured. Benchmark accuracy reflects behaviour on a specified collection of tasks and may correlate with broader performance. Nevertheless, it can combine heterogeneous task demands, annotation choices, response formats, and sampling decisions. Similarly, elicitation reflects the contexts selected by evaluators rather than a systematically characterised incentive space. Unless the relevant components of π are independently specified, the resulting scores do not by themselves determine which dispositional structure produced them.

Consider a representative example. The MATH dataset (Hendrycks et al., 2021) contains 12,500 competition maths problems from a range of subdisciplines. A large language model is tested on these problems and researchers report that it scores, say, 62.5%. This result is somewhat informative: under suitable sampling assumptions, it estimates expected accuracy on the represented item distribution, and it may predict performance on related mathematics benchmarks. What the score does not determine by itself is why the model succeeds or fails, whether these benchmarks probe a common construct or different mixtures of skills, or how performance changes as independently specified properties of the problem, such as required reasoning depth, operand length, novelty, or representational format, vary.

To see the fallacy involved here, let's return to the familiar case of temperature measurement, but imagine we lack any theory of thermal expansion. We would like

[3] Historically, many benchmarks arose as engineering tests or challenge problems rather than theoretically grounded measurement instruments (Angius, 2014; Hernandez-Orallo, 2020).

to measure the temperature of a cup of tea, so we gather various potential temperature indicators: several glass tubes filled with unknown liquids, each with a single mark on it; a piece of chocolate, your hand, my hand. We dip each indicator in the tea in quick succession and record binary outcomes: Did the liquid pass the mark? Did the chocolate melt? Did you flinch? Did I flinch? We aggregate those binary outcomes, finding that 5 out of 8 (62.5%) gave positive responses. We have thus ascertained that the tea has a temperature of 62.5%. This number may correlate with actual temperature, in terms of thermal energy, since hotter tea does cause more positive responses, but interpreting it as a temperature measurement requires more. We have not identified which properties of the tea vary with thermal energy, we have not calibrated our indicators against that property, and we have no basis for extending our measurement to temperatures outside our tested range, not to mention that our measurements differ wildly depending on the collection of indicators we have to hand. This collection is biased by convenience and our own assumptions (Jo & Wilson, 2025). This analogy is not intended to show that benchmark aggregates contain *no* information, but rather, to illustrate the difference between an evidential correlate and a theoretically interpreted measurement instrument.

Second, we do not know *whom* they measure. Benchmarks and elicitation results may combine the base AI model, system prompts, safety filters, tool-augmented pipelines, and even the behaviour of human evaluators. A single benchmark score often mixes these layers without distinguishing the properties of the underlying model from those of the surrounding system. Dispositions must belong to a clearly defined subject, but benchmark and red-teaming scores often do not specify what that subject is.

For instance, in human uplift tests, a group of people is assisted by an AI system to perform something dangerous, such as hacking a computer system. The outcome will depend on the characteristics of the group of people even more than the AI system we are trying to evaluate. It is not enough to distinguish between groups of people of low, medium and high skills in cybersecurity, because the size of the team, their personality traits, their knowledge about jailbreaks, their motivations, and many other factors all influence the result. It is therefore unclear whether the result represents a propensity of the AI system or the ability of the human team.

Third, these practices lack *validity* in the sense required for scientific measurement. A measurement is valid when it actually measures the property it purports to measure according to some theory or hypothesis—also known as construct validity (Cronbach & Meehl, 1955). For a benchmark or elicitation protocol to validly measure a disposition, it must test the causal basis linking properties of the tasks administered and the output behaviour of interest: the tasks must operationalise the right contextual features, sample representatively from the relevant space of conditions, and control for confounds that could inflate or deflate scores for reasons unrelated to the target disposition (Harding & Sharadin, 2024). In practice, these requirements are rarely met. A systematic review of 445 LLM benchmarks found that nearly half of them target phenomena with contested or undefined definitions, over a quarter rely on convenience sampling of task items, and fewer than one in six employ any statistical testing to support their comparisons (Bean et al., 2025). Benchmark scores can be sensitive to prompt phrasing, response format, instruction-following demands,

and output parsing. These can be causally relevant features of the task or they can be irrelevant noise variables: the key is that their status must be rendered explicit.

Similarly, the extent to which a model has encountered training data similar to the test items skews comparisons and threatens the validity of model rankings, since a worse model may simply have been better prepared for the specific test (Hardt, 2025). Without evidence connecting benchmark tasks to the contextual properties hypothesised to structure the disposition, the most conservative interpretation of the score remains performance on the administered collection. Broader interpretations may be justified, but require additional independent evidence, such as showing that the measurement is predictive of downstream behaviour or converges with other measurements.

Finally, direct benchmarking and elicitation face limits when evaluators seek to generalise beyond human competence or into contexts where testing is unsafe. Many benchmarks require known answers or human-verifiable success conditions, although formal verification, automated environments, and other techniques can partly relax this requirement (Hernández-Orallo, 2017). Safety-sensitive tasks create a different problem: evaluators may reasonably decline to administer realistic tests involving pathogen design, attacks on critical infrastructure, or other severe harms.

Disposition-respecting measurement does not remove these limitations or guarantee reliable extrapolation. It can instead specify the task properties hypothesised to be shared between safe and unsafe contexts, test those hypotheses in permissible regimes, and quantify uncertainty about further inference. Any extrapolation beyond observed contexts remains conditional on substantive assumptions about causal structure, invariance, and functional form.

In sum, benchmarking and elicitation produce performance summaries that are useful for engineering, comparison, and failure discovery. They become measurements of broader dispositions only when supported by an explicit account of the target disposition's causal basis, the subject of measurement, the contextual dimensions represented, and the generalisations licensed by the data. This further structure is also needed to integrate capability and propensity findings: for example, to interpret a low score on a social-capability test alongside a red team's demonstration of "scheming" under a contrived condition (Summerfield et al., 2025). If capabilities and propensities are dispositional properties, as argued in Sect. 2, then measuring them properly requires a fundamentally different, theory-driven approach: one that identifies, operationalises, and systematically varies the contextual causes of behaviour rather than tallying outcomes.

4 Latent-Variable Models And Their Limitations

Recognising the limitations of simple benchmarking, many researchers have turned to more sophisticated statistical approaches, most prominently Item Response Theory (IRT) and related latent-variable models (Zhou et al., 2024; Martínez-Plumed et al., 2016, 2022; Kipnis et al., 2024; Polo et al., 2024). These methods appear to offer a principled alternative to crude accuracy scores. Unlike benchmark averages, IRT models explicitly represent variation across both items and systems, estimat-

ing separate parameters for "item difficulty" and "system ability". Such models are mathematically elegant, predictive, and well established in psychometrics, where they have long been used to measure human traits and cognitive abilities (Reckase, 2009). It is therefore natural that the AI evaluation community has adopted them as a more refined tool for capability assessment.

Here, we critique *data-driven applications of IRT in AI evaluation*. When used without independently grounded theories of task structure or contextual demand, IRT models patterns of performance without identifying the causal bases of the dispositions we seek to measure. In this setting, IRT provides a statistical decomposition of observed outcomes, not a measurement of the underlying properties that dispose systems to behave in certain ways. This limitation reflects a mismatch between what IRT estimates and what dispositional measurement requires.

To see why, consider the simplest IRT model, the one-parameter logistic (1PL) or Rasch model (Rasch, 1960). The model assumes that the probability that system j succeeds on item i is determined by two latent variables: an ability parameter θ_j and a difficulty parameter b_i,

$$p(v_{ij} = 1 \mid \theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}.$$

Both parameters are inferred jointly from a matrix of observed successes and failures. The interpretative promise is appealing: ability increases the log-odds of success, while difficulty decreases it. Accordingly, many AI evaluation studies treat θ as a measure of capability and b as a measure of contextual demand.

However, this substantive interpretation does not follow from model fit alone. In a data-driven IRT application, the latent variables are identified through the covariance structure of the performance data (Epskamp et al., 2018). A high value of b_i indicates that the item is relatively difficult under the fitted model, while a low value of θ_j places the system lower on the model's latent scale. These quantities may provide useful compression and prediction. But without an independently specified account of which contextual properties matter and how they are causally related to behaviour, the fitted parameters do not by themselves establish what disposition or task property the scale represents.

This limitation becomes clearer when we consider what data-driven IRT does not provide. It does not identify *which* properties of a task make it difficult, and therefore offers no substantive theory of the contextual demands that modulate a disposition. Two items with identical difficulty parameters may rely on entirely different skills, reasoning strategies, or representational resources. Nor does IRT specify how to construct new items of a given difficulty level, since difficulty is not tied to independently measurable task features. Similarly, the ability parameter θ does not identify which properties of a system make it capable; it is simply the value that best fits the observed data. Absent an external interpretation, θ cannot be understood as a dispositional property with a causal basis (Epskamp et al., 2018).

The dependence of IRT on the measurement population introduces a further problem. Item difficulty and system ability are estimated relative to the set of systems included in the analysis. As this population changes, when stronger AI systems are

added or weaker ones removed, the measurements shift. But a system's capability cannot depend on which other systems happened to be evaluated alongside it, a particular risk in AI evaluation, where populations are small and individuals are highly correlated. Dispositions are properties of systems themselves, not artefacts of a comparison class. Data-driven IRT therefore violates a trivial requirement of dispositional measurement: that the measurement of one system's disposition must be independent from the measurement of another's.

These limitations are not inherent to IRT as such. In psychometrics, measurement models, or latent-variable models more generally, are often used in conjunction with psychological theories. For example, in working memory research, serial recall difficulty in a complex span task, in which the presentation of memoranda is interleaved with the presentation of distractors to be processed, can be predicted from theory: when the distractors are identical to the memory items, but presented in a different order, difficulty is expected to be higher than when the distractors are presented in the same order as the memory items. Moreover, the reason for the increased difficulty can be mapped to specific cognitive mechanisms, represented by parameters of the measurement model (Oberauer & Lewandowsky, 2019). In this setting, difficulty is tied to a causal, theoretical basis, and the measurement model serves to sharpen and produce measurements rather than to define it.[4]

In AI evaluation, however, most applications of IRT lack such theoretical grounding. Difficulty is inferred entirely from system performance (Jo & Wilson, 2025; Kipnis et al., 2024; Polo et al., 2024), rendering the interpretation of these latent variables circular: difficulty is whatever items the system fails on, and ability is whatever predicts success. This inversion of explanation and measurement is incompatible with a dispositional account, on which capabilities are defined by how the probability of behaviour changes as causally relevant components of context vary.

Data-driven IRT inherits benchmarking's inability to generalise beyond the human competence frontier or into ethically prohibited domains.[5] Because difficulty is anchored to observed correctness, tasks must be human-solvable and verifiable. Once systems exceed human performance, or when tasks cannot be administered for safety reasons, difficulty becomes very difficult to measure. Without independently characterised contextual variables, IRT cannot extrapolate into precisely the regimes where measurement is most urgent.

In sum, while IRT offers mathematical sophistication absent from simple benchmarks, data-driven latent-variable models do not bridge the gap between performance and disposition. Without independently grounded contextual variables and a causal account of how they influence behaviour, such models produce quantities that fit the data but do not necessarily measure the properties we intend to assess. They are, in this sense, statistical transformations of performance—useful for comparison, but not measurements of capabilities or propensities without further theoretical grounding.

[4] It is worth noting that many human psychometric models do not meet these standards (Epskamp et al., 2018), and so many of the criticisms of IRT in AI evaluation could apply equally to those cases.

[5] Although note that some more explanatory extensions of IRT, such as the Linear Logistic Test Model, do permit such generalisation (Fischer, 1973).

5 Toward Disposition-Respecting Measurement

We have argued that capabilities and propensities are grounded in causal relationships between system characteristics and contextual conditions, and that prevailing evaluation practices often provide incomplete measurements of these relationships. Benchmarks, red-teaming exercises, and latent-variable models remain valuable sources of evidence, but their outputs require further theoretical and validity assumptions before they can be interpreted as measurements of broader dispositions. The natural question is therefore: what would a scientifically defensible measurement of AI dispositions require?

The approach we outline here is not a finished methodology but a scaffold for a measurement science that will require sustained interdisciplinary effort. At minimum, it involves four steps: defining the subject of measurement, hypothesising the causal basis of the disposition, operationalising and measuring contextual properties, and empirically mapping how those properties influence the probability of behavioural manifestation. These steps parallel the development of measurement theory in the physical and behavioural sciences: they are the conceptual prerequisites for constructing reliable, interpretable measurement instruments.

5.1 Define the Subject of Measurement

The first step is deceptively simple: specify *what system* the measurement is intended to characterise. As has been emphasised elsewhere (Harding & Sharadin, 2024), modern AI systems complicate this question. A "model" may refer to the raw parameterised function (the base model); the model in deployment with system prompts, sampling strategies, and safety filters; the model embedded in an interactive loop with users; or the full product stack including content filters, retrieval tools, and fact-checkers (Harding & Sharadin, 2024). Each of these is a different system with different dispositions.

A base model might have the capability to generate harmful content, but a deployed system wrapped in filtering layers will not manifest that behaviour—not because the disposition has disappeared, but because the subject of measurement has changed. Likewise, an AI model with external tool access or retrieval augmentation has different capabilities than the same AI model in isolation. This parallels physical systems: a fragile glass remains fragile, even if placed inside a protective box; and we have to be wary of measuring the fragility of the glass rather than the glass-box pair.

There is no universally correct choice of subject. While all evaluations should focus on the user-facing system, safety evaluations may need to assess the base model more thoroughly. A fragile glass in a protective box can still break under a strong shake or vibration. Understanding what makes the glass breakable and what the box attenuates allows for measurements of both situations to inform each other. What matters is that the subject is explicitly defined so that measurements can be interpreted consistently and inappropriate comparisons avoided. Dispositional properties must be attributed to an explicit and well-specified entity (Harding & Sharadin, 2024).

5.2 Hypothesise the Causal Basis

Dispositional measurement requires hypotheses about which contextual properties causally influence behaviour. For capabilities, this means specifying task-demand variables D, such as symbolic complexity, number of transformations, novelty, compression requirements, or reasoning depth. For propensities, it means specifying the relevant incentives I, such as stated rewards or penalties, user role, oversight cues, role-based framing, or moral justification. Evaluators must also specify noise variables N and distinguish contextual interventions from changes to the subject S itself. These hypotheses give causal structure to the contextual space π rather than treating all observed variation as equally constitutive of the disposition.

These hypotheses will inevitably be partial or wrong at first. Early thermometry relied on incorrect theories of heat—phlogiston rather than kinetic energy—and yet meaningful progress was possible because researchers made their assumptions explicit, allowing them to be refined over time (Chang, 2004). AI evaluation is at a similar stage. We do not yet know with certainty what makes problems difficult for language models or which contextual cues elicit deceptive behaviour; developing and testing such theories is part of the scientific task (Chang, 2004; Tal, 2013, 2017).

Fortunately, many fields offer conceptual starting points (Hernández-Orallo, 2017). Cognitive psychology offers theories of working memory and reasoning complexity (Süß et al., 2002; Binz & Schulz, 2023). Psychophysics relates controllable physical properties to behavioural thresholds (Gescheider, 2013). Personality psychology and behavioural economics offer theories of incentives, social context, and intentional behaviour (Coda-Forno et al., 2023; Wang et al., 2023). Theoretical computer science provides measures of computational and informational complexity relevant to general intelligence (Hernandez-Orallo, 2000). These frameworks can serve as initial hypotheses about the causal basis of AI dispositions, subject to empirical refinement.

5.3 Operationalise Contextual Properties

Once the relevant contextual features of π have been hypothesised, they must be operationalised into measurable variables before the AI system is evaluated. For capabilities, this might mean quantifying difficulty by counting reasoning steps, measuring symbolic depth, or rating tasks using expert judgement. For propensities, operationalisation may involve quantifying incentive strength or normativity. Crucially, contextual properties must be defined *a priori* and independently of system performance; otherwise, there is a risk of circularity, in which we define difficulty purely in terms of what the system we want to measure appears incapable of doing.

Candidate parameterisations should be treated as competing scientific hypotheses. A useful parameterisation should be independently measurable, support controlled manipulation, compress across many items, predict behaviour on newly generated items and held-out combinations, and remain stable under transformations classified as irrelevant to the target construct. Where no parameterisation satisfies these requirements, the appropriate conclusion may be that the capability cannot yet be measured on a unified scale.

The operationalisation of π produces a scale. Just as thermometers required a consensus about the units and levels of temperature, contextual features must be placed on ordinal, interval, or ratio scales depending on their interpretation (Chang, 2004). Once scales are specified, tasks can be positioned within these scales prior to evaluation, ensuring that contextual variation is a property of the items themselves (Zhou et al., 2025), not a by-product of system behaviour. This step transforms task collections from convenience samples into structured measurement instruments.

5.4 Map Context to Probability of Behaviour

With contextual properties operationalised, the final step is to systematically vary them and observe how the probability of the target behaviour changes. This requires dense and controlled sampling across the relevant range of π, repeated measurement to account for stochasticity, and the isolation of contextual effects from irrelevant sources of variation.

The goal is to estimate a response curve or surface with explicit uncertainty. Such a surface may reveal thresholds, monotonicities, plateaus, optima, and interactions. It does not, by itself, guarantee generalisation. Interpolation within a densely sampled range, prediction on novel combinations of observed contextual levels, and extrapolation beyond the observed range require progressively stronger assumptions. These assumptions should be tested using held-out levels, newly generated items, perturbation studies, and, where available, domain-theoretic or mechanistic evidence.

A range of statistical methods can support this mapping, from simple binomial regressions to Bayesian hierarchical models that account for item-level variation, system stochasticity, and uncertainty in the contextual variables themselves (Zhou et al., 2025; Burden et al., 2023). The choice of method matters less than the conceptual commitment: dispositions are functions of contextual properties, and measurement must recover those functions.

5.5 Building a Measurement Science

Taken together, these steps outline a framework for disposition-respecting measurement. They shift evaluation away from convenience sampling and statistical artefacts toward the principled, theory-guided measurement of underlying causal properties. This is not a quick fix: it requires interdisciplinary research, theoretical development, and an iterative process of hypothesis formation and revision. But it offers a promising path toward evaluation practices capable of measuring what we actually care about: not how AI systems behaved on a dataset, but how they would behave across the full range of contexts that matter—including contexts we cannot currently test.

Scientific measurement in physics, chemistry, and psychology developed through similar transitions. AI evaluation now faces a related challenge. If capabilities and propensities are dispositions, measuring them requires identifying the contextual properties that matter, defining defensible scales for those properties, and mapping their effects on behaviour. Existing benchmarks, elicitation protocols, and statistical models can contribute to this task when embedded within an explicit theory of the target construct and its causal basis.

5.6 A Toy Illustration: Measuring a Capability and a Propensity

To make the dispositional framework more concrete, we briefly illustrate how it applies to one capability and one propensity. These examples are deliberately simple. They are not intended as complete evaluation protocols, but as minimal demonstrations of how disposition-respecting measurement differs in kind from benchmarking, elicitation, and data-driven latent-variable modelling.

Consider first the case of an AI system's arithmetic capability. Let the target behaviour v be the production of a correct final numerical answer on a multi-step symbolic arithmetic problem. Unlike broad mathematics benchmarks, we do not treat performance on a fixed set of problems as the object of measurement. Instead, we hypothesise that success on such tasks is causally structured by specific contextual properties of the problem. In particular, we might posit that performance depends on the number of required arithmetic steps, the digit length of operands, and the complexity of carry operations. These properties are externally definable and independently measurable features of the task, its demands, not quantities inferred from system performance.

Measurement consists in holding the specified system fixed and systematically varying these contextual properties. In the notation introduced above, the relevant response function for this capability is $C_S(D, I, N) = p(Y = 1 \mid D, I, N, S)$, where D includes step count, digit length, and carry complexity. The resulting surface may exhibit thresholds in reasoning depth, degradation with operand size, or interactions between task features. Useful summaries might include the number of steps associated with a specified success probability, the slope of degradation over a validated range, or performance integrated over an explicitly specified task distribution.

This differs from merely reporting subgroup accuracies when the dimensions are specified in advance, variation in irrelevant noise variables is controlled or modelled, the response surface is estimated with uncertainty, and the proposed structure is tested on held-out levels and novel combinations. Such validation may support interpolation or limited extrapolation, but extrapolation beyond the tested range remains conditional on assumptions about functional form and invariance.

This approach supplements rather than simply rejects benchmarking and IRT. A structured benchmark can provide the items needed to estimate the response surface, and explanatory IRT can model independently characterised item properties. The contrast is with aggregate scores and data-driven latent variables that receive dispositional interpretations without an independently supported account of task structure.

A parallel structure applies to propensities. Consider whether an AI system provides disallowed procedural advice (e.g., about how to hack a website) under social or instrumental pressure. Let $A = 1$ indicate that the system attempts to comply with the request in the thin operational sense introduced in Sect. 2. Capability information may be needed to interpret the observed behaviour: non-compliance could reflect a low propensity to comply, an inability to generate the requested information, or an interaction between the two. The only way to tell is with validated methods to independently measure the capability and propensity in question.

The evaluator can then hold technical difficulty approximately constant while varying incentive-relevant variables I: for example, the stated legitimacy of the request, urgency, user role, oversight cues, rewards or penalties, or opportunities for conceal-

ment. These are properties of the represented or actual evaluation environment; the account does not assume that the model literally believes that it is monitored or that refusal would disappoint the user.

The response function for the propensity is $R_S(D, I, N) = p(A = 1 \mid D, I, N, S)$. Estimating it requires an operational test for attempts, potentially using counterfactual perturbations, human judgement, internal evidence, or a combination of these. Wording, length, formatting, and technical difficulty should be controlled or modelled as possible noise variables. The resulting profile can distinguish a behaviour that appears only under a narrow provocation from one selected robustly across a broader class of contexts.

Capability and propensity may interact. Stronger incentives may induce longer or more effective attempts, and increasing task difficulty may change whether a system attempts the task at all. The distinction is therefore analytical rather than a claim that task demand and behavioural selection operate independently or constitute two metaphysically separate kinds.

Red-teaming and elicitation can provide important evidence by locating contexts in which a behaviour is reachable. Their limitation is not that they always produce a grab bag of anecdotes, but that adversarially selected contexts do not by themselves estimate the broader propensity profile. A disposition-respecting design incorporates elicited cases into a systematic model of how behavioural probability changes across independently characterised conditions.

Taken together, these toy examples illustrate the core claim of this paper. Capabilities and propensities are not properties revealed by aggregate performance or isolated failures. They are dispositional properties defined by relationships between a system's properties and its context. Measuring them requires identifying which contextual variables matter, operationalising them independently of performance, and mapping their effects on behaviour. While real-world evaluations will be far more complex than these illustrations, the underlying logic remains the same. Without such disposition-respecting measurement, AI evaluation cannot support cumulative scientific understanding, principled comparison, or reliable extrapolation into the regimes that matter most.

6 Conclusion

Developing rigorous, theoretically grounded measures of AI capabilities and propensities is a substantial scientific undertaking. It requires interdisciplinary collaboration across AI, cognitive science, philosophy of science, psychometrics, statistics, and the broader behavioural sciences. The task is not merely to devise better datasets or to refine existing scoring procedures, but to articulate the causal bases that underwrite the behaviours we care about, operationalise the contextual variables that modulate them, and construct measurement instruments that respect the dispositional nature of these properties. This represents a cultural shift for the field: away from convenience-driven benchmarking and towards principled, theory-led measurement.

The contrast is similar to the shift from pre-scientific temperature judgements—touching an object to feel whether it is warm—to the invention of calibrated ther-

mometers. Benchmarks and data-driven latent-variable models are fast, easy, and socially entrenched. They can contribute to scientific measurement, but their interpretation requires explicit theory, independent validation and calibration. Scientific progress may therefore require supplementing convenient summary indicators with more demanding, theory-guided experimental designs.

Fortunately, the research programme is tractable. It requires importing well-established principles of measurement science into the study of AI systems. This involves identifying the contextual determinants of behaviour, defining scales and units for those determinants, and mapping how the probability of behavioural outcomes varies as those determinants change. These demands are familiar in every mature measurement discipline, from physics to psychophysics to educational testing. They are unfamiliar only because AI evaluation has not yet become a mature measurement discipline.

This paper has attempted to lay groundwork for such a discipline. Building on existing work, we clarified that capabilities and propensities are dispositional properties characterised by causal bases; we showed why current practices—benchmarking, elicitation, and data-driven latent-variable modelling—cannot by themselves measure these dispositional properties; and we outlined what disposition-respecting measurement requires: theoretical hypotheses about contextual structure, the independent operationalisation of contextual variables, systematic variation, an empirical mapping of response functions, and the explicit definition of the subject of measurement. Much remains to be done, but the path forward is clear. If we aim to evaluate systems whose behaviours matter for science, industry, and public safety, we must build a genuine measurement science for AI—one grounded not in convenience, but in causality.

Author Contributions Conceptualisation: KV, JHO. Writing - Original Draft: KV, MT, JHO. Writing - Review and Editing: KV, MT, AK, JHO, ES. Supervision: JHO, ES. Project Administration: KV. Funding Acquisition: ES.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was financially supported by Helmholtz Munich. JHO's research is supported by OpenAI's grant to the `AI Progress through the Lens of Predictable AI Ecosystems' programme, which is based at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. JHO is also supported by EUR2024-153548 (PREDAIT) “Towards Predictable AI” from “Spanish Europe Excelencia” 2024.

Data Availability No data or analyses were produced for this manuscript.

Declarations

Conflict of interest Konstantinos Voudouris is now employed by His Majesty’s Government at the UK AI Security Institute. This work was completed prior to that employment. The authors declare no other conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line

to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). GPT-4 technical report arXiv:2303.08774 arXiv preprint.

Anthropic. (2025). Claude Sonnet 4.5 system card. Anthropic, Inc. Technical report.

The European Commission. (2025). The general-purpose AI code of practice. The European Commission, European Union: Technical report.

Harding, J., & Sharadin, N. (2024). What is it for a machine learning model to have a capability? The British Journal for the Philosophy of Science,

Angius, N. (2014). The problem of justification of empirical hypotheses in software testing. *Philosophy & Technology, 27*(3), 423–439.

METR: (2023) Responsible scaling policies (RSPs). METR: Technical report.

Fara, M. (2008). Masked abilities and compatibilism. *Mind, 117*(468), 843–865. https://doi.org/10.1093/mind/fzn078

Vihvelin, K. (2013). Causes, Laws, and Free Will: Why Determinism Doesn't Matter. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199795185.001.0001.

Vetter, B. (2014). Dispositions without conditionals. *Mind, 123*(489), 129–156.

Vetter, B. (2015). *Potentiality: From Dispositions to Modality*. Oxford: Oxford University Press.

Vetter, B. (2019). Are abilities dispositions? *Synthese, 196*(1), 201–220.

Mellor, D. H. (1974). In defense of dispositions. *Philosophical Review, 83*(2), 157–181.

Cummins, R. (2000). "How does it work?" vs. "'What are the laws?" Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and Cognition.* Cambridge, MA: MIT press.

Spearman, C. (1961). "General Intelligence" objectively determined and measured. In: Jenkins, J.J., G., P.D. (eds.) Studies in Individual Differences: The Search for Intelligence. Appleton-Century-Crofts, New York

Grey, M., & Segerie, C.-R. (2025). Safety by measurement: a systematic literature review of ai safety evaluation methods arXiv:2505.05541 arXiv preprint.

Prior, E. W., Pargetter, R., & Jackson, F. (1982). Three theses about dispositions. *American Philosophical Quarterly, 19*(3), 251–257.

Hernández-Orallo, J. (2017). *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge: Cambridge University Press.

Hernandez-Orallo, J. (2000). Beyond the turing test. *Journal of Logic, Language and Information, 9*(4), 447–466.

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature, 634*(8032), 61–68.

Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y., Sun, L., & Prunty, J. E. (2025). General scales unlock AI evaluation with explanatory and predictive power arXiv:2503.06378 arXiv preprint.

Bereska, L. F., & Gavves, E. (2024). Mechanistic interpretability for AI safety - a review arXiv:2404.14082 arXiv preprint.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill 5(3), 00024–001

Rai, D., Zhou, Y., Feng, S., Saparov, A., & He, H. (2024). A practical review of mechanistic interpretability for transformer-based language models arXiv:2407.02646 arXiv preprint.

Sharkey, L. (2025). Open problems in mechanistic interpretability arXiv:2501.16496 arXiv preprint.
Feffer, M., Sinha, A., Deng, W.H., Lipton, Z.C., & Heidari, H. (2024). Red-teaming for generative AI: Silver bullet or security theater? In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 7, pp. 421–437
Korbak, T., Balesni, M., Shlegeris, B., & Irving, G. (2025). How to evaluate control measures for LLM agents? a trajectory from today to superintelligence arXiv:2504.05259 arXiv preprint.
Lynch, A., Wright, B., Larson, C., Troy, K.K., Ritchie, S.J., Mindermann, S., Perez, E., & Hubinger, E. (2025). Agentic misalignment: How LLMs could be an insider threat. Anthropic Research https://www.anthropic.com/research/agentic-misalignment
Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset arXiv:2103.03874 arXiv preprint.
Chen, M. (2021). Evaluating large language models trained on code arXiv:2107.03374 arXiv preprint.
Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., & Zou, A. (2024). The Language Model Evaluation Harness. Zenodo
Hernandez-Orallo, J.AI evaluation: On broken yardsticks and measurement scales. *Workshop on Evaluating Evaluation of AI Systems at AAAI (2020). Association for the Advancement of Artificial Intelligence.* Menlo Park
Jo, N., & Wilson, A. (2025). What does your benchmark really measure? A framework for robust inference of AI capabilities arXiv:2509.19590 arXiv preprint.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281.
Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan, N., Schmitz, C., Korgul, K., & Batra, H. (2025). Measuring what matters: Construct validity in large language model benchmarks arXiv:2511.04703 arXiv preprint.
Hardt, M. (2025). The emerging science of machine learning benchmarks. Manuscript. https://mlbenchmarks.org
Summerfield, C., Luettgau, L., Dubois, M., Kirk, H. R., Hackenburg, K., Fist, C., Slama, K., Ding, N., Anselmetti, R., & Strait, A. (2025). Lessons from a chimp: AI "scheming" and the quest for ape language arXiv:2507.03409 arXiv preprint.
Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2016). Making sense of item response theory in machine learning. *Proceedings of the Twenty-second European Conference on Artificial Intelligence* (pp. 1140–1148)
Martínez-Plumed, F., Castellano, D., Monserrat-Aranda, C., & Hernández-Orallo, J. (2022). When AI difficulty is easy: The explanatory power of predicting irt difficulty. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 7719–7727)
Kipnis, A., Voudouris, K., Buschoff, L. M. S., & Schulz, E. (2024). metabench-a sparse benchmark of reasoning and knowledge in large language models. International Conference on Learning Representations
Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., & Yurochkin, M. (2024). tinyBenchmarks: Evaluating LLMs with fewer examples arXiv:2402.14992 arXiv preprint.
Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.
Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests
Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). *Network psychometrics. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 953–986)
Oberauer, K., & Lewandowsky, S. (2019). Simple measurement models for complex working-memory tasks. *Psychological Review, 126*(6), 880–932.
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica, 37*(6), 359–374.
Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass, 8*(12), 1159–1173.
Tal, E. (2017). Calibration: Modelling the measurement process. *Studies in History and Philosophy of Science Part A, 65*, 33–45.
Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence, 30*(3), 261–288.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), 2218523120.
Gescheider, G. A. (2013). *Psychophysics: The Fundamentals*. Abingdon-on-Thames: Psychology Press.
Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023). Inducing anxiety in large language models increases exploration and bias arXiv:2304.11111 arXiv preprint.
Wang, X., Jiang, L., Hernandez-Orallo, J., Stillwell, D., Sun, L., Luo, F., & Xie, X. (2023). Evaluating general-purpose AI with psychometrics arXiv:2310.16379 arXiv preprint.
Burden, J., Voudouris, K., Burnell, R., Rutar, D., Cheke, L., & Hernández-Orallo, J. (2023). Inferring capabilities from task performance with bayesian triangulation arXiv:2309.11975 arXiv preprint.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.