International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(18s) 2026 pp. 816-830
https://doi.org/10.70917/ijcisim-2026-4926
Article

When Logical Correctness Is Not Rational Judgment: The Limits of Formal Reasoning in Algorithmic Decision Systems

Zhe Ji

Nanjing Foreign Language School, No. 60 Huixiu Road, Qinhuai District, Nanjing 210000, Jiangsu, China
Corresponding Author Email: 18051645618@163.com
Orcid: https://orcid.org/0009-0006-4607-9174

Abstract: Algorithmic failures are conventionally diagnosed as deviations from a correct specification — bias to be corrected, error to be patched. This article argues that the more consequential and less visible failure mode runs the other way: a system executes its specification exactly, and the specification was inadequate to the situation it governed. The distinction is developed through a formal argument rather than an analogy. Treating the feature set available to a learning system as a -algebra , the article shows that no -measurable function can represent a value-relevant event excluded from , and — more precisely — that no -measurable self-assessment functional can take the adequacy of itself as an argument. Every reliability measure a model reports, from calibration to conformal coverage, is computed inside the representation whose adequacy is in question. Sections 2 and 3 locate this claim against the literatures on bounded rationality, rule-following, and algorithmic fairness, arguing that each stops short of the reflexive point at issue. Sections 4 through 6 develop the argument through decision theory, the Rashomon set, Goodhart-type selection effects, and an impossibility theorem for fairness criteria, and confront the two strongest objections available — that scope-monitoring is already being formalised, and that closure under a fixed representation cannot be a principled limit if human reasoners are themselves physical systems. The article concludes that formal validity and practical rationality answer to different criteria of success, that the empirical superiority of statistical prediction over human judgment does not collapse this distinction, and that the operative difference lies in the structure of failure rather than in its frequency: formal errors are quiet, mechanically correlated, and reproduced at scale, while human error — correlated though it often is by institutional bias, professional norms, and bureaucratic routine — remains comparatively noisy, locally visible, and costly to scale.
Keywords: algorithmic decision-making; formal rationality; practical reason; measurability and representation; specification failure; Goodhart's law; algorithmic fairness

1. Introduction

A pneumonia triage model trained on hospital records once learned that patients with a history of asthma died less often than other admitted patients, and scored them as low risk accordingly; the regularity was entirely real, manufactured by a protocol that routed asthmatic patients directly into intensive care (Caruana et al. 2015). Every step of the induction held. The correlation existed in the data, the estimator was consistent, the fitted function generalised to held-out cases, and the ranking would have survived any audit conducted in the vocabulary of the model itself. Deployed, it would have sent the most fragile patients home, and it would have done so without committing a single inferential error.

Anomalies of this kind are usually filed under bias, leakage, or distribution shift, as though the machine had slipped somewhere and the task were to find the slip. A more uncomfortable reading is available. Nothing slipped. The system did what a formal system does: it drew what followed from what it had been given, with a fidelity no

human clinician could match, and the catastrophe lived entirely in what it had been given. Fidelity to a specification and adequacy of judgment came apart, and came apart in a way that no amount of additional fidelity could repair.

That gap is the subject here. Formal validity is a property that a piece of reasoning has relative to a fixed vocabulary, a fixed set of premises, and a fixed criterion of success; rationality, in the older and broader sense that survives in ordinary talk about someone reasoning well, includes the capacity to notice that the vocabulary is wrong. Whether the second is a mere extension of the first, or something the first structurally cannot contain, decides how one should think about machine decision-making — and, less obviously, about what one is doing when one calls a human judgment unreasonable.

Three separations organise what follows. The first is conceptual: formal validity and practical rationality answer to different criteria of success and cannot be ranked on a single scale, which already blocks the common inference from "the model is more accurate" to "the model judges better." The second is structural and constitutes the central claim: the characteristic failure of algorithmic decision is not deviation from formal reasoning but excessive fidelity to it, and this failure has a precise mathematical shape that can be stated rather than merely gestured at. The third is normative and dialectical: granting all of that, human judgment is measurably worse than statistical rules across an enormous range of tasks (Meehl 1954), so the argument cannot end in a defence of intuition, and the position that survives is narrower and stranger than either side of that familiar quarrel.

The contribution is threefold. First, the article gives the intuition behind specification failure a precise formal statement rather than a metaphor, using the measurability of value-relevant events to show what a fixed representation structurally cannot raise as a question about itself (Section 6). Second, it separates two failure modes that the dominant bias-and-correction framing treats as one — execution failure and over-fidelity — and shows that they require opposite remedies, which matters because a repair aimed at the wrong one leaves the system's authority over the case intact while removing nothing that caused the harm (Section 6). Third, it takes seriously, rather than dismisses, the two objections most likely to be pressed against this position: that scope-monitoring is increasingly formalised in open-ended systems, and that a closure argument cannot be a principled limit if the human reasoners who allegedly transcend it are themselves physical systems (Sections 6 and 7). The remainder of the article proceeds as follows. Section 2 positions the claim against existing accounts of algorithmic failure, bounded rationality, and specification gaming, and states the gap each leaves open. Section 3 sets out the two vocabularies of correctness at stake. Section 4 traces what a formal specification keeps and discards at each stage of construction. Section 5 examines how these commitments behave under competitive and deployment pressure, where a specification's shelf life and its measured performance can come apart. Section 6 develops the central argument — the measurability claim and the distinction between execution failure and over-fidelity — together with its two strongest objections. Section 7 reconstructs what a defensible notion of rationality after formalism looks like, and states the conditions under which formal optimisation should be preferred regardless. Section 8 concludes.

2. Related Work and Positioning

Three literatures bear on the claim and each stops short of it in an identifiable way.

The dominant framing of algorithmic harm treats it as deviation from a correct target: a model is biased, and bias is corrected by better data, reweighted losses, or post-hoc calibration (Barocas, Hardt, and Narayanan 2023). This framing presupposes an undistorted quantity the estimate can be compared against, and it is well suited to cases where such a quantity exists. Where the disagreement concerns which target belongs in the specification at all — whether to predict re-arrest or re-offence, cost or need — deviation has nothing to grip, since there is no undistorted quantity from which the estimate deviates. Sociotechnical accounts of this abstraction problem come closer, arguing that the move which makes a social problem computationally tractable is also the move that strips out the context a full treatment would require (Selbst et al. 2019); the present article differs in giving that observation a formal rather than a sociological ground, via the measurability argument of Section 6, and in showing that the same structure recurs however far the abstraction is renegotiated.

Bounded rationality supplies the second literature. Real reasoners search until an alternative clears an aspiration level rather than optimising exhaustively, because computation and information are scarce (Simon 1955, 1997). Applied to machine systems, the framework is imported more often than it is examined: a learning system is treated as another boundedly rational agent, constrained like a human one, only faster. Section 4 argues this inverts the diagnosis. A large model is not short of search; its constraint is the narrowness of what it is permitted to represent, and the two conditions — bounded search and bounded representation with unbounded search — have opposite remedies, since more capacity improves the first and worsens the second by executing a narrow description with greater fidelity.

The third literature is closest in spirit and furthest in method: technical work on specification gaming and reward misspecification, which documents empirically that optimising a proxy reliably degrades the target once the proxy and the target diverge in some region of the search space (Amodei et al. 2016; Krakovna et al. 2020). This work is an ally rather than a rival, and Section 5 of the present article draws on it directly. What it does not supply is an account of why the divergence is structural rather than a correctable measurement error — why, that is, better specification pushes the problem to a new location rather than removing it. That account is the target of the measurability argument developed below, and it is what separates the present claim from a call for better proxies.

3. Two Vocabularies of Correctness

Validity, in its logical sense, is a relation between sentences that holds independently of what the sentences are about. An argument is valid when the truth of its premises guarantees the truth of its conclusion:

$$\Gamma \vDash \varphi \Leftrightarrow \forall \mathcal{M}\,(\mathcal{M} \vDash \Gamma \rightarrow \mathcal{M} \vDash \varphi) \qquad (1)$$

Topic-neutrality is the source of the concept's power. A valid schema transfers without loss between medicine, law, and astronomy; it can be checked by someone who understands nothing of the subject matter; it permits mechanisation, and mechanisation permits scale. The same neutrality carries a cost that is rarely counted as one. Truth-preservation says nothing about the standing of the premises, and a formal system that faithfully preserves truth will just as faithfully preserve the consequences of a premise set that misdescribes the situation. Nothing internal to (1) marks the difference.

A second property of $\vDash$ separates it from anything deliberation does. Validity is monotonic: enlarging the premise set never destroys a conclusion already secured,

$$\Gamma \vDash \varphi \;\Rightarrow\; \Gamma \cup \{\psi\} \vDash \varphi \qquad (2)$$

whereas a reason to keep a promise is cancelled outright by learning that keeping it will cost a life, and a diagnosis supported by six findings can be overturned by a seventh. Defeasibility is not a defect to be cleaned up by better formalisation; it is the shape practical inference has, since the considerations bearing on a decision are not fixed in advance and each new one can reorganise the standing of the others. Non-monotonic logics capture part of this and pay by fixing the admissible exceptions in advance, which reinstates the difficulty one level along. Probabilistic learning is defeasible in its credences — a posterior can reverse — and rigid in what the credences are about: fresh information enters only through a channel already carved for it, as a value of an existing feature, never as a consideration that redistributes the weight of everything else.

Machine learning systems are not deductive engines, and the transposition needs care. What they inherit is not the syntax of proof but the deeper feature that made proof mechanisable: correctness is defined relative to a specification that the system does not itself select. Supervised learning fixes a feature representation, a target variable, a hypothesis class, and a loss, and then searches for the element of that class minimising empirical risk:

$$\hat{h} \;=\; \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L\,(h(x_i), y_i) \;+\; \lambda\, \Omega(h) \qquad (3)$$

Whether $\widehat{h}$ is correct is a question with a determinate answer, and the answer is computable. Whether $L, \Omega, \mathcal{H}$, and the map from situation to (x, y) were the right choices is a question of a different order, and its answer is not a term in the optimisation. Specification-relative correctness is what both the syllogism and the gradient step possess; it is a genuine achievement, and it is not what people mean when they call a decision reasonable.

The rival vocabulary is older. Practical reason in the Aristotelian tradition is exercised in the perception of particulars: a general grasp of what health or justice requires does not, by itself, settle what to do with this patient in these circumstances, and the person of practical wisdom is distinguished by seeing which of the indefinitely many features of a case are the salient ones (Aristotle 2009). Perception here is not a decorative addition to a rule; it is the operation that makes the rule usable at all, and it cannot be replaced by a further rule without inviting the same question again. The point survives translation into the most rule-bound moral theory available. A criterion of obligation can be made entirely formal, resting on whether a maxim could be willed as universal law and deriving its authority from that form rather than from any expected outcome (Kant 1997), and the formality is what gives such a criterion its reach. Applying it to a case is another matter: subsumption requires a faculty that cannot be taught by rule, since a rule for applying rules would itself stand in need of application, and the regress terminates only in a capacity acquired through practice rather than instruction (Kant 1998).

Rule-following analysis makes the same structure sharper without any appeal to moral psychology. No rule contains its own criterion of application; any finite specification is compatible with divergent continuations, and the practice of going on in the same way is not itself secured by a further formulation of the rule (Wittgenstein 1953). Legal theory hits the identical wall from the direction of institutions: general terms have a settled core and a penumbra where the rule neither clearly applies nor clearly fails, and penumbral cases are decided by considerations about the rule's purpose rather than by the rule's content (Hart 1961). What these traditions converge on is not that formal criteria are useless but that their use presupposes an operation of a different type, and that this operation is systematically invisible from inside the criteria.

A tempting reconciliation says the difference is one of degree: practical reasoning is just formal reasoning with more premises, and a sufficiently rich specification would close the gap. Something in this is right and has to be conceded. Considerations once thought ineffable have been formalised, from risk aversion to fairness constraints, and the history of the exact sciences is largely a history of such conquests. What the concession cannot reach is the reflexive case. Enriching a specification is itself an act performed on the basis of a judgment that the current specification is deficient, and that judgment is made in the vocabulary that the enrichment has not yet produced. Each successful formalisation is evidence for the reconciliation's local claim and evidence against its global one, since each depended on an operation the formalism did not contain.

4. What Optimisation Keeps and What It Discards

Formalisation is a map from a situation to a tuple: a measurable space, a target, a loss, a hypothesis class, a decision rule. Figure 1 sets out the seven commitments this map makes and the residue each leaves behind. Reading the diagram from the top down gives the standard engineering pipeline; reading it from the right gives the paper's thesis in compressed form.

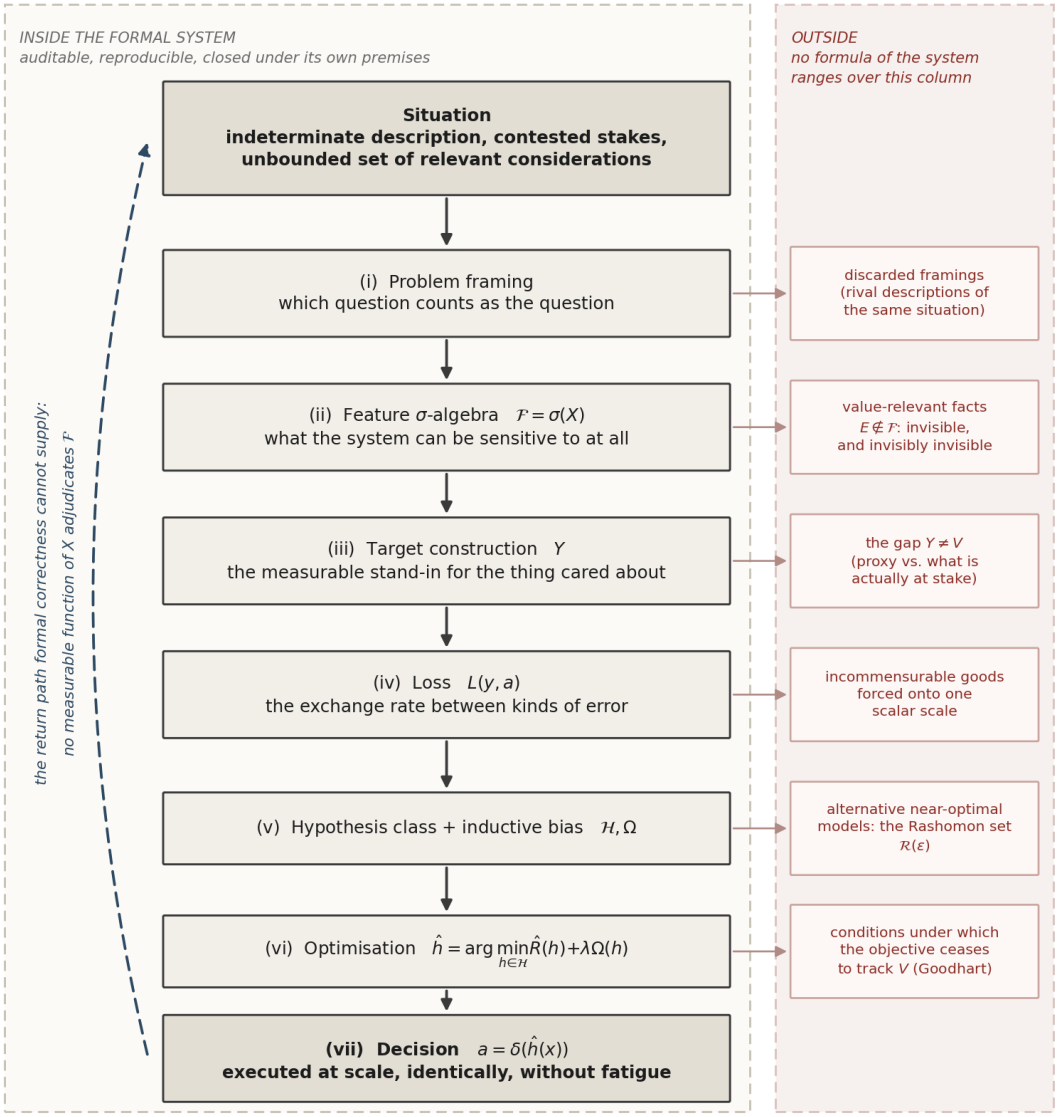

Figure 1. Seven commitments made before the first computation, and the residue each one leaves behind.

The decision-theoretic optimum makes the dependence explicit. For a fixed loss, the best action at x is

$$\delta^*(x) \;=\; \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}[L(Y, a) \mid X = x] \qquad (4)$$

and in the binary case with asymmetric error costs this collapses to a threshold on the posterior:

$$\text{choose } a = 1 \Leftrightarrow \eta(x) \;=\; \mathbb{P}(Y = 1 \mid X = x) \;>\; \frac{c_{FP}}{c_{FP} + c_{FN}} \qquad (5)$$

Nothing in the data determines the right-hand side of (5). The ratio c_{FN}/c_{FP} encodes how much worse it is to miss a case than to raise a false alarm, and answering that requires knowing whose interests are at stake, how reversible each error is, and what obligations the deciding institution has already incurred. Figure 3 traces what happens when the ratio moves while the predictive model is held fixed: one model, three thresholds, three sets of people admitted or refused, and no empirical fact distinguishing among them. The value parameter enters the calculation as a number and

thereby acquires the appearance of a technical quantity, which is the exact mechanism by which contested commitments become invisible.

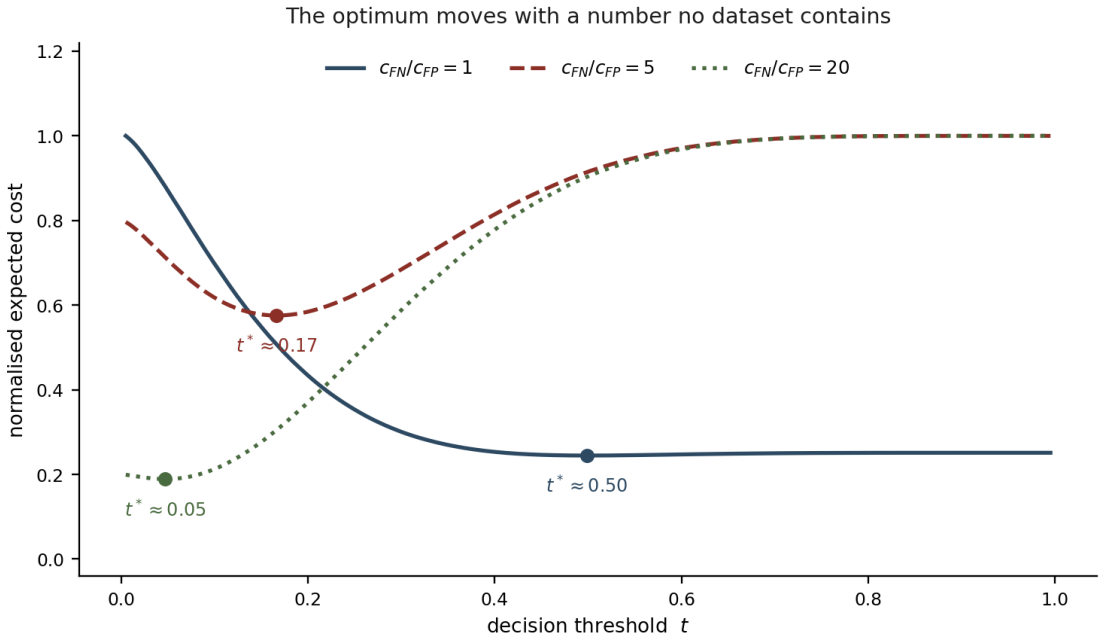

Figure 3. One predictive model, three defensible decision rules: the ratio c_{FN}/c_{FP} is a value commitment, not an empirical quantity, and it alone relocates the optimum.

Underdetermination runs deeper than the threshold. Define the Rashomon set of a learning problem as the collection of hypotheses whose risk is within ϵ of the minimum:

$$\mathcal{R}(\epsilon) \;=\; \left\{ h \in \mathcal{H} \;:\; \hat{R}(h) \;\leq\; \hat{R}\left(\hat{h}\right) + \epsilon \right\} \qquad (6)$$

For realistic tabular problems this set is large, and its members can differ sharply in which individuals they classify how, a fact known in the statistical literature well before it became an ethical concern (Breiman 2001). Figure 4 renders the situation: models statistically indistinguishable in loss, separated by a wide gap in subgroup outcomes. Selecting one member of $\mathcal{R}(\epsilon)$ is a real choice with real victims, and empirical risk does not make it. Whatever does make it — convenience, defaults, the ordering of a random seed — is a decision that formal criteria have not merely failed to justify but have concealed, which is worse.

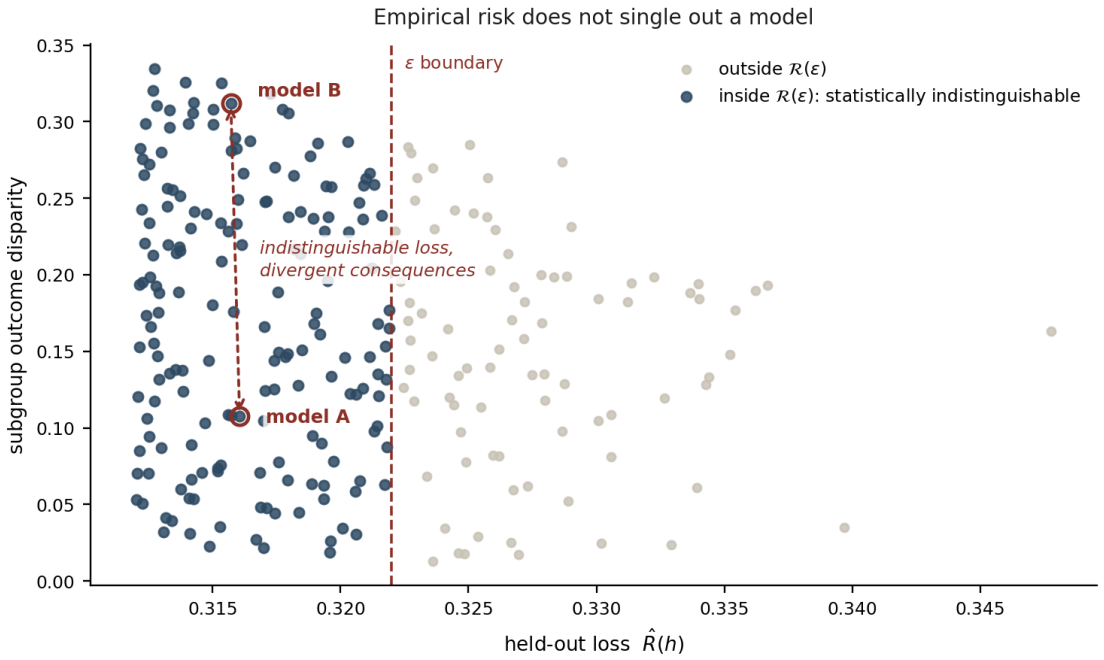

Figure 4. Schematic Rashomon set. Where many hypotheses share the minimum, the remaining choice is real, consequential, and settled by something other than the loss.

A related commitment hides inside the choice of $\mathcal{H}$ and Ω. Any finite sample is compatible with infinitely many extensions to unseen cases, and a learner that assumes nothing about which extensions are more plausible has no basis for preferring one to another; generalisation is not something data accomplishes but something assumptions accomplish with the help of data. Smoothness, sparsity, hierarchical structure, the convolutional assumption that a pattern means the same thing wherever it appears — each is a substantive claim about how the world is put together, imported before the first observation is read and rarely stated as a claim at all. Describing such systems as letting the data speak inverts the actual order of authority. The data are permitted to adjudicate among hypotheses that a prior commitment has already selected as candidates, and the commitment does its work invisibly, since it appears in the code as an architectural choice rather than as an assertion about the world that might be false.

Bounded rationality is the standard resource for saying that real reasoners cannot optimise, and it fits human institutions well: organisms with limited computation and limited information search until an alternative clears an aspiration level and then stop, rather than ranking all alternatives (Simon 1955). Satisficing can be written down as plainly as optimisation,

$$a^{*} \;=\; \text{first } a \in \mathcal{A} \text{ encountered with } V(a) \;\geq\; \tau \qquad (7)$$

and the aspiration level τ adjusts with experience, which makes the procedure adaptive rather than merely lazy (Simon 1997). Transposed to machines, however, the diagnosis inverts. A large model is not short of computation in the relevant sense; it can search a space no human could enumerate, and it does not tire, forget, or lose interest at hour nine. The constraint that binds it is not the bound on its search but the narrowness of the description it searches within. Bounded rationality names a limitation of the reasoner; what afflicts algorithmic judgment is a limitation of the representation, executed by a reasoner with almost no limitations at all. The machine is not boundedly rational. It is unboundedly faithful to a bounded description, and the two conditions have opposite remedies: the first improves with more capacity, the second worsens.

Critiques of the optimising model from within economics reach an adjacent conclusion by a different route. Treating an agent as a consistent preference ordering makes commitment unintelligible, since acting against one's own welfare on principle registers only as inconsistency (Sen 1977). What is lost is not accuracy but a distinction: the

difference between wanting something and endorsing wanting it. Rationality as internal consistency also leaves untouched the question of whether the ends are worth having, which is why coherence is compatible with absurdity, and why a formal account of reasons has to include the capacity to stand back from one's own desires and ask whether they merit satisfaction (Elster 2009). Incommensurability is the sharpest version: some goods stand in no determinate ranking, and forcing them onto a common scale does not discover a truth but manufactures one (Raz 1986). A loss function is exactly such a manufacture. It does not measure the exchange rate between a wrongful detention and a missed diagnosis; it stipulates one, and the stipulation is then carried through every subsequent computation with perfect consistency and no memory of its own contingency.

5. Formal Criteria Under Competitive and Deployment Pressure

Competitive machine learning is an unusually clean specimen of a formal system with a public criterion, which is what makes it philosophically useful rather than merely biographical. A Kaggle competition fixes a dataset, a train–test split, and a scalar metric; participants submit predictions; a leaderboard ranks them. Every question that would ordinarily be contested has been settled in advance, and what remains is the purest form of the specification-relative correctness described above. The design is a small epistemology: within it, to know more is to score higher, and nothing else counts as knowing.

The competitive record bears this out at a level of detail practitioners rarely make explicit. In tabular competitions, the margin between a strong and a mediocre entry typically owes less to the choice of algorithm than to feature construction — to decisions about what the model is permitted to see as is widely reported in post-competition write-ups. Two entries separated by a fraction of a percent in log loss can embody substantively different theories of which historical facts about a case bear on its future outcome, and the leaderboard registers only the fraction of a percent. The disagreement between the two theories is real and consequential — one entry's feature construction may encode a defensible causal story, the other a spurious correlation that happens to hold in this sample — and the metric has no vocabulary in which that disagreement could be raised, let alone adjudicated.

Overfitting to a public leaderboard sharpens the same point in a form that is fully formal. Repeated evaluation against a held-out set turns that set into training data through the analyst's own adaptive choices, and validity guarantees degrade in a way that is invisible to any participant checking only their score; the phenomenon has a rigorous treatment and even a partial technical remedy (Dwork et al. 2015). What the remedy cannot supply is the recognition that a number has stopped meaning what it meant. The mechanism is general: a measure adopted as a target ceases to be a good measure, because optimisation pressure discovers the region where the proxy and the goal diverge. Writing the proxy as the goal plus an error term,

$$U = V + \varepsilon, \qquad \varepsilon \text{ dominates } V \text{ in the upper tail of } U \;\Rightarrow\; \lim_{u \uparrow \bar{u}} \mathbb{E}[\, V \mid U \geq u \,] < \mathbb{E}[V] \tag{8}$$

makes the antecedent explicit rather than leaving it implicit in the decomposition: the decomposition $U = V + \varepsilon$ alone entails nothing about where the maximum of U sits relative to the maximum of V, and the degenerate result only follows once ε is stipulated to dominate V in the region being selected on. That stipulation is not exotic — it is what "optimising a proxy" means once the proxy and the goal start to separate — but stating it as a hypothesis rather than a free consequence is what keeps the formalism honest. Under that hypothesis, conditioning on high U selects for large error rather than for large value, and the expected value of the selected alternatives falls below the unconditional mean. Weak optimisation never reaches that tail and therefore never encounters the pathology; strong optimisation searches exactly there. Improved formal performance and degraded substantive performance are not independent events but causally linked ones, which is why capability gains can worsen outcomes without anything in the pipeline going wrong (Manheim and Garrabrant 2019).

Deployment history supplies the blunter lesson. The winning ensemble of the most celebrated prediction contest ever run achieved the full ten percent accuracy gain required and was never put into production, the engineering cost

exceeding the value of the gain while the business shifted from posting discs to streaming, which changed what a recommendation was for (Amatriain and Basilico 2012). No one erred. The metric answered its question correctly; the question had a shelf life the metric could not see. The asthma case is the same structure with the stakes raised, and the model was caught only because it had been rendered in a form that let clinicians read what it had learned (Caruana et al. 2015). A black-box model of equal accuracy would have passed every test in (3) and killed people — an argument for interpretability owing nothing to claims about trust or user experience (Rudin 2019).

Scale converts these local properties into political ones. A scoring system opaque to those it ranks, applied to millions, and never corrected by the outcomes it produces entrenches whatever its specification encoded, and those worst placed to contest it are those with most at stake (O'Neil 2016). Deployment is also not a passive observation of a fixed distribution. A model that directs patrols to a district generates arrests there, which enter the next training set as evidence that the district is dangerous; a model that denies credit prevents the repayment history that would have refuted its estimate. Writing this as

$$\mathcal{D}_{t+1} \;=\; g(\mathcal{D}_t,\, \delta_t), \qquad \delta_t \;=\; \delta\big(\hat{h}_t\big) \tag{9}$$

makes the loop explicit: the quantity being estimated is partly the system's own product, so accuracy can rise while the estimate detaches further from what it was supposed to track. Validation against held-out data cannot register the drift, since the held-out data are drawn from the same contaminated process. A formalism that is silent about the difference between measuring the world and making it will report success throughout.

Formal methods can even prove that formal methods will not settle a normative dispute. Where base rates differ across groups, calibration within groups, equal false negative rates, and equal false positive rates cannot jointly hold except in degenerate cases:

$$\begin{aligned}&\text{calibration} \;\wedge\; \text{balance for the positive class} \;\wedge\; \text{balance for the negative class}\\ &\quad\Rightarrow\; \text{perfect prediction or equal base rates}\end{aligned} \tag{10}$$

(Kleinberg, Mullainathan, and Raghavan 2017). Each criterion is a reasonable formalisation of treating people fairly; the theorem establishes that one must be surrendered, and it is silent on which. A choice therefore has to be made in a currency the formalism does not stock, and the mathematics is what proves the necessity of the extra-mathematical step. Attempts to escape by defining fairness more carefully tend to inherit the same problem, since the abstraction that makes a criterion tractable is what strips out the social context that would tell you which criterion applies (Selbst et al. 2019).

6. Over-Fidelity as a Failure Mode

Two failures look alike from outside and are not alike at all. In the first, a system violates its own specification: a bug, a leak, a mis-specified likelihood, an assumption relied upon and not met. In the second, the system honours its specification exactly, and the specification was not adequate to the situation. Repairs for the first are internal and convergent; repairs for the second require an operation the system cannot perform on itself. Figure 2 sets out the diagnostic sequence, including what happens when the second remedy is attempted.

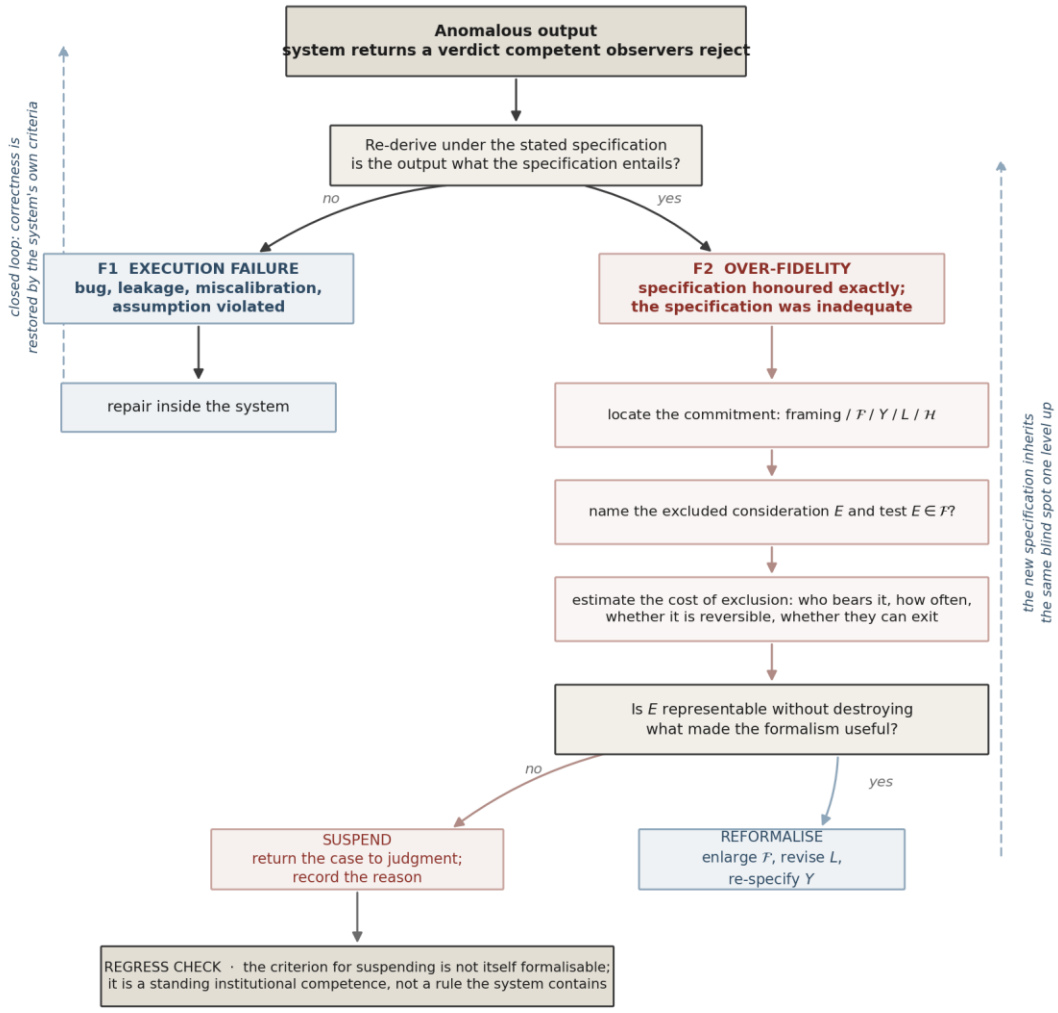

Figure 2. Diagnostic separation of the two failure modes, and the asymmetry of their remedies.

The core structural claim can be stated with more precision than the surrounding literature usually allows. Let $\mathcal{F} = \sigma(X)$ be the σ-algebra generated by the features available to the system, and let E be an event that bears on what should be done — that this applicant's arrears followed a hospitalisation, that this defendant's neighbourhood is policed at four times the rate of the next one. If $E \notin \mathcal{F}$, then

$$\nexists\, g\ \mathcal{F}\text{-measurable such that } \mathbf{1}_E = g(X) \text{ a.s.} \qquad (11)$$

As mathematics this restates a definition rather than proving a theorem, and pretending otherwise would be the borrowed rigour the argument is against. Its value lies in what it forces one to say next, and two qualifications come first. Completion matters: if E differs from an element of $\mathcal{F}$ by a null set, an almost-sure reconstruction exists, so the exclusion must be non-trivial rather than notational. And the non-existence holds for a fixed X alone — adding a feature, an override flag, an audit channel, or a free-text field enlarges the σ-algebra, and E becomes representable at once.

The second qualification is where the argument actually lives. Enlarging $\mathcal{F}$ is an act performed from outside, on the basis of a judgment that something is missing, and that judgment cannot be a computation on X, because the fact that $\mathcal{F}$ omits E is not a fact expressible in the completed algebra generated by X. What follows is narrower than

a claim about the limits of computation as such, and it is exactly this: no $\mathcal{F}$-measurable self-assessment functional takes the adequacy of $\mathcal{F}$ as an argument. Every quantity a model reports about its own reliability — calibration curves, predictive entropy, conformal intervals, ensemble disagreement — is computed inside $\mathcal{F}$ and quantifies uncertainty about Y given the world as described, never uncertainty about the description. A model can be maximally uncertain and still blind in the way that matters, and it will report high confidence in the asthma case, because within $\mathcal{F}$ the evidence really is overwhelming. Representational adequacy is not an unsolvable problem; it is a problem whose solutions arrive from outside the representation, which is a different and more precise complaint.

Gödelian analogies are available here and mostly do damage: incompleteness concerns the provability of arithmetical sentences, not the adequacy of empirical representations, and the two are related by metaphor rather than by theorem. Only the shape is worth keeping — a system's expressive resources fix what it can raise as a question, and questions about its own frame tend to fall outside them.

Contrast this with the dominant framing of algorithmic harm. Bias talk presupposes a correct target from which the output deviates: the model is skewed, and the task is to unskew it. That framing is right often enough to be worth keeping, and wrong in the cases that matter most. Where the disagreement is about which target belongs in (3) at all — whether to predict re-arrest or re-offence, whether to predict cost or need, whether a proxy that encodes access to care should stand in for illness — the language of deviation has nothing to grip, because there is no undistorted quantity for the estimate to be compared against. Debiasing a well-fitted model of the wrong quantity produces a fairer instrument for doing the wrong thing. The abstraction traps that arise when social context is stripped out are not artefacts of careless engineering but the price of the representational move that made computation possible in the first place (Selbst et al. 2019). Symbolic AI reached the same barrier from the opposite direction decades earlier: attempts to specify relevance in advance failed not for want of computing power but because the relevance of a fact depends on a situation that is not itself given as a list of facts (Dreyfus 1992).

The strongest objection at this point is that scope monitoring is already being formalised, and rather well. Out-of-distribution detection flags inputs unlike the training data; conformal methods supply coverage guarantees under weak assumptions; abstention mechanisms let a model decline rather than guess; distribution-shift tests raise alarms when the input stream changes character. Each is genuine progress, and the boundary of the formal is plainly movable. What none of them escapes is that novelty is assessed in the metric of the representation whose adequacy is in question, so an input can be perfectly typical in $\mathcal{F}$ while being atypical in every respect the modeller failed to record. The asthma patients were not outliers: numerous, well represented, unremarkable along every recorded axis, and invisible to any detector operating on those axes. Formal scope monitoring catches the cases where the world has moved away from the description and misses the cases where the description was never right, which is the more dangerous of the two.

Generality supplies a second and more contemporary objection. Systems trained on open-ended text and images take input that no one enumerated in advance; there is no feature table, no fixed schema, and apparently no $\mathcal{F}$ of the kind (11) presupposes, so the whole analysis may look like a description of an earlier technology. What has happened is a relocation rather than a dissolution. Tokenisation fixes what counts as an elementary symbol; the context window fixes how much of a situation can be present at once; the training distribution fixes which regions of language the model has any purchase on; and preference tuning fixes, through a finite set of comparisons made by particular annotators under particular instructions, which continuations are treated as better. Each of these is a commitment of the same logical type as a column in a table, and their combination defines a σ-algebra that is enormous, irregular, and unwritten — which makes it harder to audit, not less binding. Generality also introduces a failure the tabular case does not have: a model that can produce a sentence about any consideration whatever, including a sentence acknowledging its own limitations, gives the appearance of scope-awareness while the appearance is generated by the same distribution that generated everything else. Fluent self-doubt is not scope monitoring, and a system able to articulate the objection to itself is not thereby able to act on it. The problem becomes less visible as the interface widens, which is the opposite of the usual expectation and a reason to state the underlying structure now rather than after the interface stops looking like a table.

The dialectic cannot stop here, because the natural conclusion — trust human judgment, which is context-sensitive by nature — is empirically indefensible. Simple actuarial rules outperform expert clinical prediction across an enormous range of tasks, a finding that has survived seventy years of attempts to overturn it (Meehl 1954). Human judgment is not merely biased in the systematic ways catalogued in the heuristics literature (Tversky and Kahneman 1974; Kahneman 2011); it is also noisy, in the sense that professionals disagree wildly with one another and with themselves on identical cases, and noise is invisible from the inside in a way bias is not (Kahneman, Sibony, and Sunstein 2021). The context-sensitivity celebrated in the previous section is the same faculty that lets irrelevant context in. Whatever is right about practical judgment cannot be that it produces more accurate answers, because usually it does not.

What distinguishes the two is better described along a different axis: not accuracy, but the structure of failure. Human errors are noisy — correlated, certainly, by institutional bias, professional norms, ideological frames, and bureaucratic routine, but correlated through imitation, training, and institutional pressure that stay lossy and locally perturbed rather than mechanically identical — locally visible, and expensive to scale; a bad clinician harms the patients in one clinic, and the next clinician's mistakes are different mistakes at the margin even where both err in the profession's direction. Formal errors are quiet, mechanically correlated, and free to replicate; a mis-specified model applies the same defect to every case in the same way, indefinitely, and its consistency guarantees that no disagreement will surface to flag the problem. Uniformity is normally an argument for automation, and it should be counted here as a hazard: the very property that makes formal systems auditable also removes the friction by which mis-specification is ordinarily detected. A million identical judgments are not a million pieces of evidence that the judgment is right.

Nor can the argument stop at a division of labour in which machines handle the routine and humans the exceptional, since the exception is not marked in advance. Recognising a case as one the rule was not made for is the competence at issue, and it cannot be delegated to whichever party receives the output. A human kept in the loop without the standing to reject the system's frame produces a formal system with a signature attached: accountability transferred without capacity, and a mechanism for laundering the specification's commitments into someone else's professional responsibility.

7. Rationality After Formalism

A vocabulary for this contrast already exists in the sociological tradition, which distinguishes rationality that consists in calculability and procedural exactness from rationality that appeals to substantive values external to the calculation, and treats the historical dominance of the former as an institutional fate rather than an intellectual advance (Weber 1978). The diagnosis is powerful and, taken alone, misleading. Framing the opposition as calculation versus values invites the reply that values are subjective and calculation is not, which is a debate the substantive side cannot win and which misstates the issue. What the algorithmic case exposes is not that formal systems lack values but that they contain them in a specific and dangerous manner: a loss function is a value commitment already made, expressed in a notation that makes it look like a measurement, and thereby exempted from the kind of challenge that an openly stated commitment would attract.

The reconstruction that fits the evidence therefore has to be second-order. Judging well, at the level that distinguishes it from calculating well, is not a matter of getting more answers right; it is the competence to hold a formalism accountable to what it excludes, and it decomposes into the three operations the argument has been tracking. Identifying the commitments a representation has silently made was the work of section 4; monitoring the conditions under which those commitments stop holding was the work of section 6; revising or suspending the representation for reasons rather than results is what remains. Each is a comparative claim — human beings perform all three badly, and the machinery of (3) performs them not at all.

Two objections press hard. The first is regress. If the criterion for suspending a formalism cannot itself be formalised, either it is a further rule and the problem recurs one level up, or it is not a rule and looks like licence to override inconvenient outputs whenever intuition objects. The second is the observation, from the previous section,

that intuition is exactly what should not be trusted. Both objections are correct as stated, and the response cannot be to dissolve them.

What can be said is that the regress is not vicious in the same way at every level. Practical wisdom is uncodifiable in the sense that no finite specification of it will yield correct application without a residual act of judgment, and this is a structural feature of ethical concepts rather than a temporary deficiency in our theorising (McDowell 1979). Uncodifiable does not mean unaccountable. A judgment that a rule does not fit a case can be defended with reasons, contested by others, revised in light of the contest, and recorded so that a pattern of such judgments becomes visible and criticisable over time. Reflective equilibrium describes precisely this two-directional movement, in which principles are adjusted to considered judgments and judgments to principles until the pair stabilises, with neither given final authority (Rawls 1999). The procedure has no algorithm and terminates in no proof, and it is nonetheless the most disciplined available model of what revising a specification for good reasons looks like. Its limitation is equally clear: it presupposes a community of reasoners with time, standing, and the willingness to be moved, which is not what a deployed decision system has, and rarely what the people subject to one are granted.

A third objection is sharper than either and cuts closest. If human beings perform this second-order operation, it is performed by physical systems, and whatever a physical system does admits in principle of description; refusing that inference requires a commitment about mind that the argument has not earned and should not want. The concession is unavoidable and the position survives it, because the claim was never that scope-monitoring is uncomputable. It concerns closure: any system operating under a fixed representation cannot, within that representation, raise the question of the representation's adequacy, and this holds regardless of substrate, holds of human reasoners while they are inside a settled frame, and explains why professionals also fail to notice what their categories omit. What differs is not access to a view from nowhere but susceptibility: people can be moved from outside a frame — by an anomaly they cannot assimilate, by someone whose interests the frame discounted, by a case that will not sit still — through channels an optimiser has no analogue for. A machine could in principle be built to be disturbable in this way, and building it would consist in constructing those channels rather than in improving performance within the frame. Nothing in a model's accuracy contributes to that construction, which is why the two capacities need separate names.

Insisting on the second-order reading also disarms the standard complaint that substantive rationality is a licence for preference dressed as principle. What is being asked of a judgment is not that it express the right values but that it remain answerable to considerations it has not yet represented, and answerability is a discipline with recognisable failures: refusing to state which consideration was omitted, declining to say what would count as evidence that the omission mattered, invoking context whenever a conclusion is unwelcome. A second-order competence is also not in competition with prediction on prediction's own ground, which is why the empirical superiority of statistical rules leaves it untouched. Noticing that a question has been mis-posed is not a more accurate answer to that question; it is a refusal to treat it as the question, and no accuracy figure can be evidence for or against a refusal of that kind. Comparisons that place judgment and computation on a single scale of performance have already assumed what is at issue, since the scale is itself one of the commitments under examination.

The most sophisticated technical response to all of this takes the objective itself as unknown, so that the machine treats human preference as a latent quantity to be inferred and retains uncertainty about what it is supposed to be maximising, which yields deference and a willingness to be switched off as consequences rather than as patches (Russell 2019). The design instinct is exactly right — build the doubt into the formalism instead of stationing it outside. The limit is structural and is the argument of section 6 recurring one level up. Uncertainty about the objective is represented as a distribution over a space of candidate objectives, and that space is a representational commitment of the same kind as $\mathcal{F}$; a consideration absent from it is not assigned low probability but is unrepresentable, and the system's humility will be perfectly calibrated over the wrong domain. Alignment work that grows out of the same tradition tends to confirm this in practice, since the recurring difficulty is less that objectives are hard to optimise than that they are hard to state without residue (Christian 2020). Formalising the meta-level is progress, and it relocates the problem rather than removing it.

Older objections to computational accounts of mind bear on this obliquely and should not be leaned on. Whether symbol manipulation that satisfies every behavioural criterion could amount to understanding remains contested (Searle 1980), and the behavioural criterion was itself offered as a way around an unanswerable question rather than as an answer (Turing 1950). The present claim survives whichever way those disputes go, since it concerns the closure of a representation and not the presence or absence of anything inside it.

From a defined angle, then, substantive rationality is the more adequate notion — and the angle has to be stated rather than assumed, since the opposite verdict is correct elsewhere. Where a task is stable, its boundaries well understood, its errors reversible, its stakes distributed across many small decisions, and its subjects able to exit, formal optimisation is superior on every dimension that matters, including fairness, since consistency is a genuine virtue and human noise is a genuine harm. Where a decision is repeated at scale across people who cannot exit, where errors are correlated by construction, where the target is a proxy for something no one can measure directly, and where being wrong is not recoverable, the ordering reverses: the value of the formal system's consistency is exceeded by the cost of its silence about its own scope. Predictive models that decide who is treated, detained, hired, or refused credit fall on the second side, and the reason has nothing to do with accuracy figures, which are often excellent, and everything to do with what (11) says about what those figures cannot include.

Design consequences follow directly and are more modest than the usual calls for human oversight. Interpretability ceases to be a user-interface preference and becomes the precondition for detecting F2 failures at all, since a specification's inadequacy can only be seen through what the model has actually learned (Rudin 2019). Contestability matters more than explanation: the relevant capacity is not that an affected person receives an account of the decision but that they can put a consideration into the record which the representation omitted, and that the omission can force a revision. Institutional memory of suspensions matters, because the pattern of cases in which a formalism was set aside is the empirical trace of its scope conditions, and the closest thing to a measurement of specification adequacy available. Each of these is a way of building the return path in Figure 1 out of institutions, given that it cannot be built out of mathematics.

8. Conclusion: The Competence to Suspend a Formalism

Formal validity guarantees that conclusions inherit the standing of premises; it can guarantee nothing about the standing of premises, and a system whose criterion of success is defined by a specification cannot represent the question of whether that specification fits. Algorithmic decision-making is not irrational in the ordinary sense of the word. It is rational in one sense with a completeness no human achieves, and the completeness is what removes the friction, hesitation, and disagreement through which mis-specification is ordinarily caught. The pneumonia model was not confused. It was right about everything it was in a position to be right about, and the situation extended past that position in a direction it had no means to face.

Redefining rationality is therefore too weak a description of what is required; what is required is recognising that two capacities have been travelling under one word, that they can be maximised independently, and that maximising the first while neglecting the second produces systems whose errors are quiet, uniform, and immune to the ordinary channels of correction. The second capacity is not intuition, whose empirical record is poor. It is the disciplined, contestable, institutionally supported practice of asking what a formal description has left out and what would have to be true for it to keep holding — an activity with no algorithm, in which the ability to give reasons for suspending a rule is the substance of the competence rather than a failure of rigour.

Several questions remain open and are, from a computational standpoint, the interesting ones. Scope conditions might be given a partial formal treatment, not as a decision procedure but as a discipline of stating in advance the circumstances under which a model's outputs should not be used, and then testing whether those statements were honoured. Specification adequacy might be measured indirectly through the frequency and clustering of suspensions, which would turn Figure 2's regress check into a data-generating process. Rashomon multiplicity might be exploited rather than hidden, presenting decision-makers with the divergent behaviour of statistically equivalent models so that

the choice concealed inside (6) becomes an explicit one. None of these dissolves the structural point. They are ways of arranging institutions so that the boundary of a formalism is patrolled by something other than the formalism, which is what an adequate conception of rationality has always required, and what the accuracy of a model has never been able to supply.

REFERENCES

1. Amatriain, Xavier, and Justin Basilico. "Netflix Recommendations: Beyond the 5 Stars." *Netflix Technology Blog*, 2012.
2. Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." arXiv:1606.06565, 2016.
3. Aristotle. *Nicomachean Ethics*. Translated by David Ross, revised by Lesley Brown, Oxford University Press, 2009.
4. Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
5. Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science*, vol. 16, no. 3, 2001, pp. 199–231.
6. Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730.
7. Christian, Brian. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, 2020.
8. Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press, 1992.
9. Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. "The Reusable Holdout: Preserving Validity in Adaptive Data Analysis." *Science*, vol. 349, no. 6248, 2015, pp. 636–638.
10. Elster, Jon. *Reason and Rationality*. Translated by Steven Rendall, Princeton University Press, 2009.
11. Hart, H. L. A. *The Concept of Law*. Oxford University Press, 1961.
12. Kahneman, Daniel. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
13. Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. *Noise: A Flaw in Human Judgment*. Little, Brown Spark, 2021.
14. Kant, Immanuel. *Critique of Practical Reason*. Translated by Mary Gregor, Cambridge University Press, 1997.
15. Kant, Immanuel. *Critique of Pure Reason*. Translated by Paul Guyer and Allen W. Wood, Cambridge University Press, 1998.
16. Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *8th Innovations in Theoretical Computer Science Conference (ITCS)*, 2017, pp. 43:1–43:23.
17. Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. "Specification Gaming: The Flip Side of AI Ingenuity." *DeepMind Blog*, 2020.
18. Manheim, David, and Scott Garrabrant. "Categorizing Variants of Goodhart's Law." arXiv:1803.04585, 2019.
19. McDowell, John. "Virtue and Reason." *The Monist*, vol. 62, no. 3, 1979, pp. 331–350.
20. Meehl, Paul E. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, 1954.
21. O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.
22. Rawls, John. *A Theory of Justice*. Revised ed., Harvard University Press, 1999.
23. Raz, Joseph. *The Morality of Freedom*. Oxford University Press, 1986.
24. Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence*, vol. 1, 2019, pp. 206–215.
25. Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.
26. Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 417–424.
27. Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. "Fairness and Abstraction in Sociotechnical Systems." *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 59–68.
28. Sen, Amartya. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs*, vol. 6, no. 4, 1977, pp. 317–344.
29. Simon, Herbert A. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics*, vol. 69, no. 1, 1955, pp. 99–118.
30. Simon, Herbert A. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*. 4th ed., Free Press, 1997.
31. Turing, Alan M. "Computing Machinery and Intelligence." *Mind*, vol. 59, no. 236, 1950, pp. 433–460.
32. Tversky, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases." *Science*, vol. 185, no. 4157, 1974, pp. 1124–1131.
33. Weber, Max. *Economy and Society: An Outline of Interpretive Sociology*. Edited by Guenther Roth and Claus Wittich, University of California Press, 1978.
34. Wittgenstein, Ludwig. *Philosophical Investigations*. Translated by G. E. M. Anscombe, Blackwell, 1953.