Author's accepted manuscript. Forthcoming in *Philosophy of Science* as part of the *PSA 2026 Contributed Papers* proceedings. The published version may differ

A Topological Learning Theoretic Justification for Bounded Rational Analysis as a Methodological Strategy in Cognitive Science

Brendan Fleig-Goldstein

Department of Philosophy, Brown University, Providence, RI

brendan_fleig-goldstein@brown.edu

March 2026

Abstract

This paper presents a topological learning-theoretic analysis of an approach in cognitive science called bounded rational analysis. In this approach, modelers begin by deriving an optimal cognitive model, then use discrepancies between idealized calculations and observed human behavior to identify psychological constraints (e.g., memory limits), and incorporate such constraints into a newly derived resource-optimal model in an iterative scientific process. I show that this de-idealization process exploits an epistemic asymmetry: models positing greater rationality are more falsifiable. A methodological preference for more rational models can therefore be epistemically justified via learning theory.

1 Introduction

Irrationality provides a rich source of evidence for cognitive models. Errors often give insight into what is going on under the hood. Bounded rational analysis (BRA) is an approach in cognitive science that attempts to harness such evidence into a systematic methodology (Fleig-Goldstein, 2025; Icard, 2023).

BRA begins by establishing a task-environment (the goals and environmental structure relevant to a cognitive task) and deriving an optimal cognitive model. Discrepancies between behavior predicted by the optimal model and actual human behavior indicate sources of the discrepancy in the form of psychological constraints (e.g., memory limits). Such process details are incorporated into the derivation of a new model that is resource-optimal relative to bounds: that is, relative to the task-environment and a set of constraints, what is the best that can be done? This new resource-optimal model then provides a new baseline from which further discrepancies can be observed, indicating new psychological process details to be incorporated, and the methodology iterates.

The suggestion is to start with a minimal set of constraints and de-idealize by adding further constraints in response to data—constraints indicated via signatures in deviations between theoretical calculation and observation. But why start with the minimal set? Why not initially suppose that agents are only minimally rational, as opposed to minimally constrained, and proceed in the opposite direction? How can one justify the iterative de-idealization of optimality as the starting point?

In what follows, I attempt to establish the following: the de-idealization process, starting with fewer constraints, exploits an epistemic asymmetry: models positing fewer constraints are more falsifiable in the sense of Kelly (1996) and Schulte (2023). This means that the method can be justified as one guaranteed to stabilize to the truth.

2 Topological Analysis

The following topological framework comes from Genin (2018), Genin and Kelly (2019), and Kelly (2024). The topology analyzes degrees of falsifiability of propositions, and shows how methods that select hypotheses in ways that ensure progressive falsification converge to the truth.

Consider a standard propositional space. Let W be a set of worlds w, where each world is a complete assignment of truth values to all propositions of interest. The set W represents the relevant epistemic possibilities. Propositions correspond to the set of worlds in which they are true. Some parts of a world are observable; these determine the information states afforded by w. In this setting, empirical information is itself a proposition. Let I be the union of all information states afforded by worlds w, i.e., $\bigcup_{w \in W} I(w)$. A scientific question Q is a partition of W, with each cell representing a distinct answer. An empirical problem is the triple (W, I, Q): the epistemic possibilities, the available data, and the question to be answered.

In cognitive science, worlds or relevant epistemic possibilities correspond to different ways the mind might work. As a simplifying assumption, I take the possible information states to be human behavior (rather than physiological data). These observable behaviors include choices such as moves in games, answers to questions, elicited credences, and similar actions.

A method M is a function from information states to propositions. A method gives a relevant response to a question if it maps information states to disjunctions of answers to that question. A method verifies a proposition (e.g., a hypothesis) if and only if it converges infallibly to that proposition in the limit when it is true. A proposition is verifiable if there exists a method that verifies it. Genin and Kelly (2019) prove that the verifiable propositions correspond exactly to the open sets in the topology generated from I. This topology is the set I closed under arbitrary unions and finite intersections.

The intuition behind this result is as follows: (1) information states themselves are verifiable propositions (e.g., "the mercury level in the graduated cylinder is at this mark"), (2) arbitrary disjunctions of verifiable propositions are verifiable, since verifying a disjunction requires only that one disjunct be true (e.g., "life exists on another planet" can be verified even in an infinite universe), and (3) only finite conjunctions of verifiable propositions are verifiable, as only finitely many conjuncts can be checked (e.g., "the sun will rise every day" is not verifiable). Thus, given a set of possible information states, the topology yields exactly the set of verifiable propositions.

Refutable propositions are those for which a method exists that can verify their complement. They correspond to the closed sets in the information topology—that is, the set I closed under arbitrary intersections and finite unions.

For these results to hold, I must satisfy certain axioms to serve as a topological basis from which a topology can be generated. The basis axioms are as follows: (I.1) $I(w) \neq \varnothing$; (I.2) for each $E, F \in I(w)$, there exists $G \in I(w)$ such that $G \subseteq E \cap F$; (I.3) I is countable. For this framework to be applicable to cognitive science, the information states used must satisfy these axioms. Axioms I.1 and I.3 are easily satisfied across virtually all scientific contexts. I.1 holds because there is always some attainable information (in the worst case, trivial information such as W). I.3 holds because empirical data are always recorded in a language that is at most countably infinite (Genin, 2018, p. 30).

The difficult basis axiom to establish is I.2. This axiom states that information accumulates: acquiring some data should never preclude the future acquisition of other attainable data. A violation would occur, for instance, if one had only a single sample of a substance to use in an experiment—after performing one reaction, it may be impossible to restore the initial conditions and conduct a different experiment on the same sample. Unfortunately, this begins to sound a lot like psychology. If a scientist is interested in a particular subject's ability to memorize a list of items under different conditions, then once the subject has undergone training with the stimuli, it is no longer possible to observe how they would have performed had they learned the material in a different way or setting. The obvious solution is to conduct between-subject experiments and change the focus from the psychology of particular individuals to general features of human cognition.

To the extent that there is a common psychological structure, inquiry into this structure allows Axiom I.2 to be satisfied. Additional observations of behavior under new conditions remain possible, as initial conditions can be restored by drawing on new subjects. That the topological analysis cannot be applied to all questions in psychology due to violations of I.2 is, in my view, a virtue rather than a vice. Rather than justifying methods too broadly, the topology highlights specific domains where a distinctive kind of knowledge is attainable. In this way, it becomes prescriptive: researchers are encouraged to pose questions and structure inquiry so as to satisfy I.2.

Most scientific hypotheses of interest are neither verifiable nor refutable. However, connecting these concepts to topology allows one to define a rich landscape of new concepts within the formal framework and prove various properties about them. Kelly's insight is that while many scientific hypotheses are neither open (verifiable) nor closed (refutable), many important hypotheses are locally closed (verifutable). A locally closed proposition is a conjunction of an open and a closed proposition. Epistemically, this means that if it is true, it will eventually become refutable: the verifiable conjunct will be verified, leaving only the refutable conjunct.

To use a standard example from Kelly's work: suppose one is interested in the polynomial degree of a curve generated by some scientific phenomenon (e.g., the relationship between sunlight levels and the growth of a particular strain of fungus). The information states consist of open intervals along this curve—imperfect measurements of sunlight levels and colony size. The hypothesis that the governing law is polynomial of degree n is neither verifiable nor refutable, but it is verifutable. It is the conjunction of the hypothesis that the curve is at least degree n (verifiable) and at most degree n (refutable).

The hypothesis "at least n" (e.g., at least quadratic) is verifiable because, if true, data will eventually deductively entail it. It is not refutable, since no data can conclusively rule it out if it is false. Conversely, the hypothesis "at most n" is not verifiable, as no data can deductively establish it if true, but it is refutable, because if false, data will eventually show this. Therefore, the hypothesis that the polynomial degree is exactly n has the property that, if true, it will eventually become refutable.

Once one recognizes that the hypotheses under consideration are all locally closed (i.e., that there is a locally closed cover of W), an ordering among them naturally emerges. One can think of the verifiable conjunct of a verifutable proposition as the trigger and the refutable conjunct as the defeater. In the polynomial degree case, the trigger for the hypothesis "polynomial degree n" (data showing "at least n") is the defeater for the hypothesis "polynomial degree $n-1$". The reverse does not hold. Kelly shows that for a locally closed cover to exist on the hypothesis space, defeaters must serve as triggers for other locally closed hypotheses, thereby chaining the hypotheses together. This structure induces an ordering that arises from epistemic asymmetries between verifutable propositions.

This ordering captures the intuitive idea that lower-degree polynomial hypotheses are "simpler" than higher ones (e.g., a quadratic law is simpler than a cubic law). Genin and Kelly (2019) formalize this with a simplicity relation defined as follows: $A \lhd B \iff A \subseteq \text{frnt} B$, where $\text{frnt} B$ is the topological frontier of B. If a world lies in the frontier of B, it means that B is false but will never be refuted. For example, if the true law is quadratic, the cubic hypothesis will never be refuted. The polynomial degree hypotheses are nested accordingly, with lower-degree hypotheses lying in the frontier of higher-degree ones (see Figure 1).

With a simplicity relation in place, one can show that it is transitive and asymmetric, and thus induces an ordering over a set of hypotheses. A method is called *Ockham* if it always outputs the simplest hypothesis consistent with the data.

Additionally, Genin (2018, p. 38) define a solution to a question Q as follows: "A method M is a solution to Q if it converges, on increasing information, to the true answer in Q; that is,

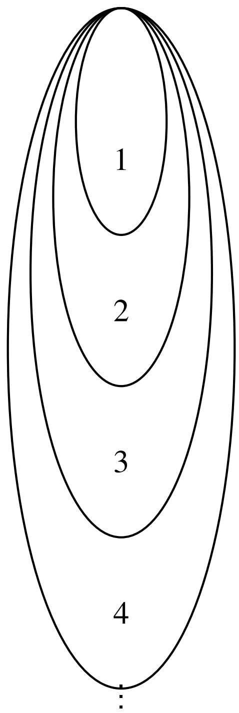

Figure 1: Topological relations among hypotheses of increasing polynomial degree. If the true hypothesis is of degree n, then a degree $n+1$ hypothesis is false but cannot be refuted; the reverse does not hold. Lower-degree hypotheses are nested within the frontiers of higher-degree ones. Information states consistent with degree n are also consistent with higher degrees.

for every $w \in W$, there exists $E \in I(w)$ such that $M(F) \subseteq Q(w)$ for all $F \in I(w)$ entailing E. A problem is *solvable* if it has a solution."

A method is *progressive* if, once it outputs the true answer, it never outputs a different one—that is, it never abandons the truth once it finds it. The key result is that if the answers to a question Q can be enumerated in an order that agrees with the simplicity order, then there exists a progressive solution, and this solution will necessarily be an Ockham method (Genin, 2018).

The upshot is that, in science, if a question can be formulated so as to induce a simplicity order, then one can justifiably prefer simpler answers and be assured that the data will drive one to converge progressively to the truth. However, in many scientific contexts, no such simplicity order is available. For example, in cognitive science, one central question is whether certain concepts are innate or learned through domain-general mechanisms. Neither hypothesis is topologically simpler than the other.

It would be a reductio of the topological analysis if it implied that every scientific question admits a simple, truth-converging recipe. The rarity of simplicity orderings in nature comports with the fact that many empirical questions have not yielded to progress, and those areas that have, often did so only after very clever ways of asking the right questions were devised.

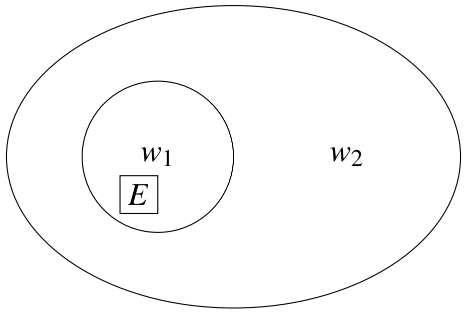

Figure 2: Topological analysis of two hypotheses: w_1: Humans are ideally rational. w_2: Humans are not ideally rational. Information states E compatible with w_1 are compatible with w_2 but not vice versa.

The topological analysis is prescriptive: ask questions in cognitive science whose answers will result in a simplicity order. I claim that the virtue of BRA as a methodology in cognitive science is that it asks the right question: one whose answers, when properly understood, reveal facts about psychological processes that can be enumerated in a simplicity order. This, in turn, justifies the method as one that is guaranteed to converge progressively to the truth in the limit.

The question BRA asks is: "How rational are human minds?" With a proper understanding of rationality, a topological analysis can show that hypotheses positing greater rationality are topologically simpler. This, in turn, justifies methods that preferentially select such hypotheses.

Consider first a simplified binary question: are humans ideally rational or not? There are two epistemic possibilities corresponding to the answers: $W = \{w_1, w_2\}$, where w_1 is the hypothesis that humans are ideally rational, and w_2 is the hypothesis that they are not.

For ease of exposition, suppose that by "ideally rational" we mean that humans conform to the norms of logic and probability. The information states are the possible observable behaviors. For example, experimentalists might present subjects with tasks requiring them to assign credences to hypotheses as new information is introduced, in order to assess whether belief updates conform to Bayesian norms.

Suppose, more specifically, that the task is the categorization task discussed above, where subjects infer the probability that a new object possesses certain properties based on previous examples. The information states are the credences subjects assign to new objects having those properties.

Consider an arbitrary information state E that is compatible with w_1. E must be a case of a subject reporting a credence in accordance with Bayes' theorem. However, E is also compatible with w_2: it is always possible that, in the future, an information state F will be observed in which a subject's credence disagrees with Bayesian norms and deductively rules out w_1.

Suppose you continue to test the subject on a categorization task using increasingly complex and numerous stimuli, and the subject continues to assign credences in line with Bayesian norms (Anderson, 1990). You can never rule out that at some point the stimuli will become so complex that the subject can no longer perform the correct computation. Thus all information compatible with w_1 is also compatible with w_2, but not all information compatible with w_2 is compatible with w_1. Thus, w_1 is refutable but not verifiable, while w_2 is verifiable but not refutable. If w_1 is true, then w_2 is false but will never be refuted. Topologically, this corresponds to the standard Sierpiński space (Figure 2).

The situation is epistemically analogous to the "all ravens are black" case, as analyzed within the same learning-theoretic framework (Kelly, 1996). If the question is "Are all ravens black?", a method that answers "no" forever is always correct if the true answer is no, but it will never find the truth if the true answer is yes. In contrast, a method that answers "yes" until it encounters a counterexample—and then answers "no" forever after—will always eventually stabilize to the correct answer. This epistemic asymmetry is a reason to prefer the latter method.

If the question is "Are humans ideally rational?", a method that answers "no" forever is always correct if the true answer is no, but it never stabilizes to the truth if the answer is yes. In contrast, a method that answers "yes" until it encounters a counterexample—and then answers "no" thereafter—will always eventually stabilize to the correct answer. This epistemic asymmetry thus provides a reason to adopt the method that assumes humans are ideally rational until proven otherwise.

It is easy enough to elicit irrational behavior from humans, thereby falsifying the hypothesis that they are ideally rational. What is needed now is a way to generalize this method so that it can be applied iteratively. The way to do this is to ask a different question: "How rational are humans?"—formulated in such a way that hypotheses positing greater rationality are topologically simpler.

Standard methods for comparing the rationality of non-ideal agents do not produce a notion of "more" rational that is topologically simpler. For example, evaluating agents by quantitative performance measures—such as accuracy, coherence, or monetary gain—does not result in hypotheses positing greater rationality being topologically simpler. Consider the question, "How rational is this agent?", where answers specify how calibrated the agent's credences are. If the question concerns performance on a particular finite test, it is decidable, and there is no inductive problem or generalizable method. However, if the question instead concerns the agent's general calibration—beyond a single test—then a demon argument can be made against any method. No information will ever refute a particular hypothesis about an agent's calibration. Incoming data can make an agent appear calibrated to any value for an arbitrarily long time, forcing a method to converge on that value, only for information to change this value indefinitely afterward. Similar arguments can be constructed against other similar notions of rationality.

Consider, however, the account of resource-rationality proposed by Fleig-Goldstein (2025). On this view, all agents are analytically resource rational: a maximally general notion of constraint includes any psychological factor that impairs task performance in a given environment, leading to the conclusion that all agents are doing their best relative to their cognitive limitations (Figure 3).

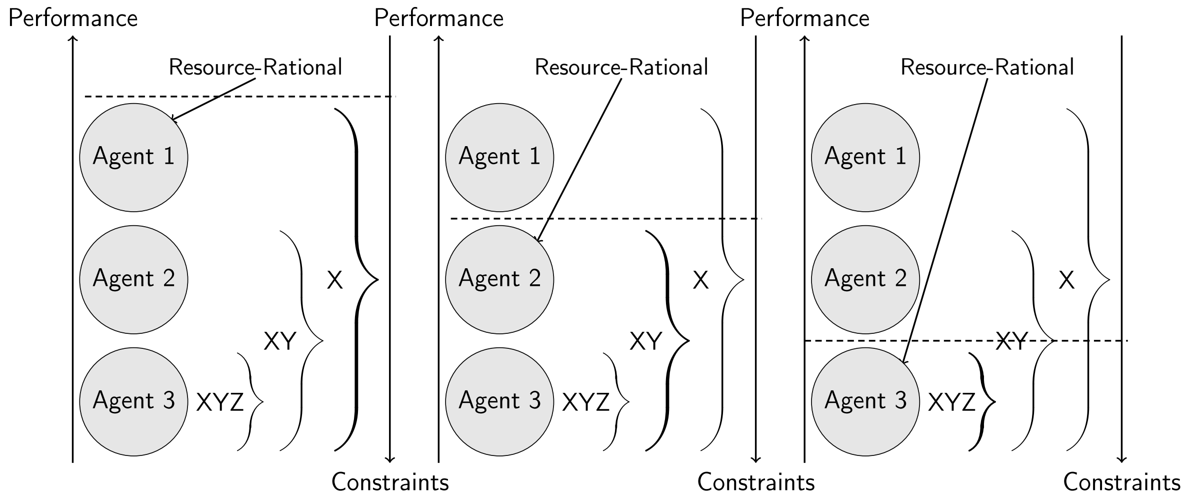

Figure 3: With a maximally broad notion of psychological constraint, one can always lower the bar enough to bring into view the resource-rationality of an agent. The question then changes from "Is this agent ideally rational, resource-rational, or irrational?" to "Relative to what constraints is this agent optimal?"

One advantage of this account for BRA is that it makes the resource-rationality assumption secure (indeed, unfalsifiable), therefore when one derives a cognitive model using the resource-rationality assumption as a premise, all inductive risk is isolated to claims about what constraints an agent is subject to.

This notion of resource-rationality also permits the following comparison relation $>_R$: an agent is more resource-rational than another if the former's constraints are a proper subset of the latter's.

Being subject to qualitatively different constraints is, on this view, not comparable. If one agent has bounded memory and another has bounded metabolic resources for computation, it makes little sense to ask which one is more constrained. Even if one constraint seems to hamper performance more than another, if they affect different performance areas, tasks, or environments, it is not clear how to compare them. Perhaps a principled, systematic theory for comparing qualitatively distinct constraints can be established, and then its implications for the topological analysis can be examined. For now, I stick to the idea that the only time agents can be clearly compared in terms of their rationality is if one agent's constraints are a subset of another's. If an agent is subject to both memory and metabolic constraints, then that agent is clearly more constrained than an agent subject to only one of those constraints.

A consequence of this view is that the "more rational" relation induces a partial order. This partial order is a virtue of this account and, I suspect, necessary to allow for a topological simplicity order to be induced from rationality comparisons. If the rationality of all agents were fully comparable, then a simplicity order would not be induced. There are no epistemic asymmetries to exploit between the hypothesis that an agent has bounded memory and the hypothesis that they are metabolically bounded. The fact that interval scales of rationality, such as calibration scores, allow for total orders is plausibly part of the reason they do not induce simplicity orders. Simplicity relations are rare, so a rationality comparison relation should be selectively applicable if hypotheses that posit greater rationality are to be simpler.

Constraints in this context, by definition, hamper performance. Constraints are always individuated relative to a task and environment. The same finite memory is a constraint relative to a task and environment where it hampers performance, but not relative to a different context where it does not.

Since constraints are indexed to particular tasks and environments, they are always additive; once a constraint is uncovered, it will never be found that it was not, in fact, a constraint. Thus, even though there are almost certainly non-linear interactions among disparate psychological facts' contributions to the performance of a cognitive system, it will always be the case that a psychological detail was a detriment to performance relative to the context in which it was indicated. While its status as a constraint is indexed to the task-environment, as a descriptive psychological fact it will persist beyond such contexts, and the collection of such facts is indeed the objective of the methodology.

In general, psychological constraints can be characterized with an internal handle (physical-biological specification) and an external handle (the material conditions under which a system exhibits behavioral deficit signatures). For example, memory limits can be characterized in terms of internal memory elements in a cognitive architecture or else can be characterized as the external ability to handle certain informational loads from the task-environment (recall of lists of size x but not x+1). Note it is not the behavioral deficit itself that is the constraint–that would make resource-rationality empty. It is rather the task-environment conditions under which the behavioral deficit occurs. The goal is not unique mechanistic identification but a reliable external handle that, however coarse-grained, reflects internal facts about the system that affect its performance. Thus, while behavioral data may underdetermine the internal handle, the data allow for verification of the externally defined constraint.

In sum, indexing constraints to particular task-environments ensures that constraints are additive and behaviorally verifiable, while still reflecting important internal psychological facts.

Constraint	Description	resource-rational strategy
c_1: Bounded perception	Inability to perceive facts in the environment; creates need to infer	Bayesian inference
c_2: Finite memory	Inability to calculate Bayesian inference for complex stimuli; creates need to approximate Bayes	Approximate Bayesian inference as well as possible with finite memory
c_3: Finite sampling resources	Sampling is metabolically costly; limits number of samples relative to memory	Approximate Bayesian inference with finite memory and sampling resources
c_4: Inappropriate sampling resource allocation	Too few samples are used given available metabolic resources	Approximate Bayesian inference with finite memory and available resources

Table 1: Constraints on inference and corresponding resource-rational strategies

Now, let $W = \{w_1, w_2, w_3, w_4\}$, where (see Table 1):

$w_1 = c_1$	(Bayesian agent)
$w_2 = c_1 + c_2$	(Approximate Bayesian agent)
$w_3 = c_1 + c_2 + c_3$	(Weaker Approximate Bayesian agent)
$w_4 = c_1 + c_2 + c_3 + c_4$	(Even Weaker Approximate Bayesian agent)

Given the definition of resource-rationality, we have the ordering $w_1 \succ_R w_2 \succ_R w_3 \succ_R w_4$. The proper subset notion of "more rational" results in the relation holding over nested hypotheses, just like in the case of polynomial degrees:

$$w_1 : Y = \alpha X + \beta X^2$$
$$w_2 : Y = \alpha X + \beta X^2 + \gamma X^3$$
$$w_3 : Y = \alpha X + \beta X^2 + \gamma X^3 + \delta X^4$$

Claim: $w_m \succ_R w_n \implies w_m \lhd w_n$ (more rational hypotheses are simpler). Just as the hypothesis that the true law is a polynomial of degree exactly n is a conjunction of the hypotheses that the degree is at least n and at most n, the hypothesis that an agent is resource-rational relative to certain constraints amounts to the claim that the agent is "subject to exactly these constraints." This hypothesis is a conjunction of "subject to at least this many constraints" and "subject to no more than this many constraints."

"At least this many constraints" is verifiable: for example, if an agent is subject to at least c_1 and c_2, eventually the data will show that the agent is incapable of calculating Bayesian inference perfectly (i.e., once the difficulty of the task is cranked up enough) (See Table 1 and Icard (2014) and Sanborn et al. (2006, 2010)). "At most this many constraints" is refutable: for example, if the claim is that an agent is subject to at most c_1 and c_2, then eventually, if the agent is subject to c_3, data will show that the approximation is suboptimal relative to c_1 and c_2, indicating that c_3 is true and that "at most this many constraints" is false. Hence, "subject to exactly these constraints" is verifutable.[1]

[1]This makes BRA a "paradigm" in Kelly (2024)'s sense: paradigms are sigma-constructible propositions, which are countable disjunctions of locally closed propositions.

The trigger for being subject to n constraints is a defeater for being subject to $n-1$ constraints. If w_2 is the true world, w_3 is false but will never be refuted. If w_3 is the true world, w_2 is false and will be refuted, and w_3 will become refutable. Thus, w_2 is in the frontier of w_3, and in general, if $w_m >_R w_n$ then $w_m \lhd w_n$.

So now consider a method that violates the preference for more rational hypotheses. Such a method jumps the gun and supposes that an agent is subject to a constraint n before the trigger for w_n has been seen. Either this method will never converge to the truth, or it can be forced to drop the truth once the truth has been hypothesized (Kelly, 2024). That is, either it is not a solution method or it is not a progressive solution method. Suppose on the one hand, that after seeing data for an arbitrarily long time that suggests there is no constraint n, this method always sticks to its guns and maintains n. It will never conjecture the truth in any world w_m where $m < n$. Thus it would not be a solution. Suppose, on the other hand, that the method is such that after seeing data for an arbitrarily long time consistent with there being no constraint n, it changes its answer to w_m for some $m < n$. It is then always possible for nature to force such a method to always drop the truth: suppose the true world is w_n and the method happens to answer w_n before seeing data that triggers this hypothesis. Nature then shows data that is consistent with w_m where $m < n$ until the method changes its answer to w_m, thereby dropping the truth. Then nature eventually shows data showing that w_n is triggered after all, forcing the method into a cycle of belief change.

Thus, a simplicity order can be induced from the partial order formed by the "more resource-rational" relation. Given Kelly's Ockham necessity theorems, assuming it is possible to enumerate resource-rational hypotheses from more rational to less rational (e.g., enumerate further constraints in response to deviations from irrationalities), then a method that prefers more rational hypotheses will be a progressive solution to the question "how resource-rational" is this agent? By framing inquiry into cognition as a question about how constrained agents are, BRA ensures progressive falsification of hypotheses, such that the data will eventually drive the cognitive scientist to the truth.

3 Conclusion

Facts about constraints are, therefore, learnable in the learning-theoretic sense. Constraints reflect facts about agents' psychologies. The bet that resource-rationality makes is that most psychological process details can be discovered in the form of constraints. That is, the fixed structure of cognition—although conducive to rational behavior in most contexts—will result in mistakes in other contexts. If humans were "ideally" rational, there would be no constraints to learn, and this methodology would reveal little about human psychology. Nevertheless, the upshot of this paper is that preferring more rational models is a justifiable scientific strategy.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Fleig-Goldstein, B. (2025). Alternatives to hypothesis testing: Demonstrative induction in cognitive science.

Genin, K. (2018, September). *The topology of statistical inquiry* [Doctoral dissertation, Carnegie Mellon University, Department of Philosophy] [Email: konstantin.genin@gmail.com].

Genin, K., & Kelly, K. (2019). Theory choice, theory change, and inductive truth-conduciveness. *Studia Logica*, *107*, 949–989.

Icard, T. F. (2014). The algorithmic mind: A study of inference in action.

Icard, T. F. (2023). Resource rationality. https://philarchive.org/rec/ICARRT

Kelly, K. (1996). *The logic of reliable inquiry*. Oxford University Press.

Kelly, K. (2024, February). *The topology of scientific inquiry* [Unpublished manuscript], Carnegie Mellon University.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 1–6.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167. https://doi.org/10.1037/a0020511

Schulte, O. (2023). Formal Learning Theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University.