Position: Beyond *Reasoning Zombies* — AI Reasoning Requires Process Validity

Rachel Lawrence [1 *] **Jacqueline Maasch** [2 *]

Abstract

Autonomous reasoning is among the most scientifically and economically motivating topics in AI today. Historically the purview of symbolic AI, recent advances have mainly emerged from deep probabilistic generative models. Despite immense interest and rapid progress, the generative AI community has not clearly converged on operational definitions for reasoning and often implicitly rejects the historical treatment of this topic in logic, verifiable automated reasoning, and symbolic methods in general. **We contend that this definitional ambiguity leaves the construct validity of reasoning evaluation unverifiable, and undermines quantifiable progress toward the collective goal of trustworthy autonomous reasoning.** We also contend that this ambiguity is addressable. To that end, we provide (1) general and extensible definitions for *valid* and *sound reasoning* based on a synthesis of the literature, which can serve as an accessible reference and a starting point for community discussion; and (2) a checklist for best practices in the communication of AI reasoning research.

REASONING — INFORMAL DEFINITION

The process of selecting and applying sequences of rules that act on prior beliefs and current evidence to obtain principled belief updates in evolving states.

CORE POSITIONS

Thesis 1 ***Define, then measure.***
- 1.1 Research concerned with AI reasoning should provide formal operational definitions for the reasoning phenomena under investigation.
- 1.2 The construct validity of reasoning evaluation should be explicitly justified with respect to the operational definitions provided.

Thesis 2 ***Reasoning is a (learnable) rule-based process.***
- 2.1 Reasoning is a process of *exact rule application*. Learnable rules unambiguously map reasoning inputs to outputs and can encompass theorems, functions, policies, etc., including rules pertaining to stochasticity, uncertainty, and approximation.

Thesis 3 ***Rule-based reasoning is valid.***
- 3.1 The *validity* of a reasoning process arises from exact rule application, independent of rule selection.

Figure 1. Core theses of this position.

1. Introduction

The prospect of AI reasoning is among the most scientifically and economically motivating advancements of the current era. Recent progress has been fueled by the remarkable empirical performance of large reasoning models (LRMs): large language models (LLMs) fine-tuned for *reasoning tasks* (Def. E.1; Huang & Chang 2023). A wave of benchmarking successes invites many questions: Is autonomous reasoning an emergent behavior that arises with scale (Wei et al., 2022a; González & Nori, 2024)? Is it a foregone conclusion that LRMs can be formally characterized as autonomous reasoners? The answers are contingent on how reasoning is defined. So then, *what is reasoning*? Though a universal definition may not exist, we argue that practical *operational definitions* (Def. E.2) are achievable but not yet in widespread use in generative AI. Building on Feynman's adage,[1] we are motivated by the following: *What I cannot define, I do not understand.*

Lack of Consensus Reasoning remains an elusive target, despite prolific study across the history of human thought. Although claims of emergent reasoning in generative AI are now commonplace, "there is not a clear definition of what it entails" (Huang & Chang, 2023). In the absence of community consensus on what formally constitutes reasoning in generative AI, we observe a normalization of research outputs that claim to study, improve, measure, or promote AI reasoning without rigorously defining the form of reasoning under investigation. This definitional void enables shifting goalposts and leaves the construct validity of reasoning evaluation unverifiable, obscuring clear progress

[*]Equal contribution [1]Microsoft Research, Cambridge, UK [2]Cornell Tech, Department of Computer Science, New York, NY. Correspondence to: Rachel Lawrence <rachel.lawrence@microsoft.com>, Jacqueline Maasch <maasch@cs.cornell.edu>.

[1]*What I cannot create, I do not understand.*

toward human-level reasoning. Avoidance of formal definitions may owe to an implicit assumption that reasoning is an intuitive concept requiring no explicit definition; evasion of the hard work of devising operational definitions; silent rejection of historical definitions from symbolic AI; or (un)intentional conflation of benchmark accuracy with reasoning itself. This position aims to make the danger of such avoidance evident, and to suggest an alternative path.

Reasoning Zombies & Other Hard Problems The black-box design and natural language interface of LRMs present a nontrivial challenge: differentiating true reasoning from reasoning-like speech. The latter represents superficial emulation: *talking like a reasoner* with no guarantees that conclusions arose from reasoning rather than memorization, guessing, or some other man-behind-the-curtain (Mitchell, 2025a). This challenge is not unique to AI reasoning: parallels can be drawn to the hard task of human cognitive testing and to distinguishing human-level intelligence from sophisticated emulation, as canonized by the Turing Test (Pinar Saygin et al., 2000). This problem evokes a rough analog of the philosophical zombie (*p*-zombie) thought experiment, which we term the *reasoning zombie* (*r-zombie*). In the classic thought experiment, *p*-zombies are systems that superficially behave like conscious beings, yet lack any conscious internal experience (Chalmers, 1997; 2020). Analogously, ***r*-zombies are systems that superficially behave as autonomous reasoners, but lack valid internal reasoning mechanisms.** While a *perfect* *r*-zombie (i.e., one which behaves identically to a true reasoner in all circumstances) remains purely theoretical, we argue that (1) *imperfect* AI *r*-zombies have already come into existence; (2) differentiating AI *r*-zombies from AI reasoners is theoretically and, often, empirically possible; and (3) we must responsibly determine when real-world use cases *require reasoners*, and when (im)perfect *r*-zombies suffice.

What This Position Is Our core theses (Fig. 1) follow from two main problems.

- **P1** Reasoning in generative AI has experienced unnecessary and addressable definitional ambiguity, where imprecise and overloaded definitions are often misaligned with historical treatments of this topic (when definitions are provided at all).
- **P2** This breeds mismeasurement, promotes an illusion of shared understanding among researchers, and subverts measurable progress toward trustworthy AI reasoning.

What This Position Is Not We do not claim that the AI community must converge on one universal definition for reasoning. We do not attempt to conclusively answer whether LRMs are autonomous reasoners, nor the kind or extent of reasoning they can perform. This position is not an endorsement for or against symbolic AI, purely data-driven approaches, nor neuro-symbolic AI. We do not make claims about reasoning in natural intelligences, nor do we argue that reasoning implies understanding or consciousness.

Contributions & Artifacts

1. **An operational definition for reasoning as a learnable, rule-governed process.** Based on a synthesis of the literature, we introduce an operational definition of AI reasoning for general use and community discussion (§2). Per Thesis 1, we express this definition in (1) natural language for intuition; (2) mathematical notation for concretization; and (3) pseudocode (Algorithm 1). Operationalization is confirmed by trivial Python implementations.[2] We illustrate the application of our definitions to special cases, including logical deduction, Bayesian inference, reinforcement learning (RL), and probabilistic next token prediction. In §3, we address rebuttals to our definitions and theses.
2. **Recommendations for scientific communication.** We propose a checklist of community guidelines that complies with Thesis 1–Thesis 3 (Appendix A).

1.1. Problem Significance: Why Do We Care?

The import of P1 and P2 lies primarily in the following: (1) reasoning is a necessary (but not sufficient) precondition for artificial general intelligence (AGI); (2) AI evaluation faces a construct validity crisis, which has spilled over into reasoning evaluation; and (3) the rate of user uptake for LRMs has outpaced evidence of trustworthy reasoning.

Reasoning is a Precondition for AGI Though contentious, AGI is widely considered an implicit or explicit north star for contemporary AI research (Blili-Hamelin et al., 2025). However, lack of community consensus on the definition and measurement of AGI hinders progress. A recent effort to operationalize AGI promotes a taxonomy of subcomponents and benchmark-based means of measuring these (Hendrycks et al., 2025). Based on human cognitive testing, this taxonomy emphasizes *on-the-fly reasoning* as an essential component of measurable AGI. We agree with Hendrycks et al. (2025) that the ability to reason is a necessary (but not sufficient) precondition for AGI. An excess of valuable use cases aside, this alone is sufficient to justify AI reasoning as a critical area of inquiry. However, like AGI, a shroud of ambiguity, confusion, and debate looms over the definition and measurement of AI reasoning. Because reasoning is a necessary precondition for AGI, measurable progress toward clearly defined reasoning will be necessary for measurable progress toward AGI. For example, we could pose the open question: can an *r*-zombie, perfect or imperfect, ever achieve AGI? Whether the answer is positive or negative, answering such questions will require theory and methods for differentiating AI reasoners from *r*-zombies.

[2] https://anonymous.4open.science/r/valid_reasoning/

Construct Validity is Underemphasized Recent waves of generative AI tend to emphasize exploratory research and empirical evaluations over hypothesis-driven confirmatory research, proof of theoretical guarantees, or formal verification (Def. E.3) (Herrmann et al., 2024). Historically, empirical fields have taken explicit precautions against mismeasurement via *construct validation* (Cronbach & Meehl, 1955): justifying that experimental measures capture the phenomena of interest by devising operational definitions that relate latent abstract constructs (e.g., intelligence, bias, ideology) to measurable real-world proxies. And yet, "validity and other quality criteria of empirical research have gained little attention in ML so far" (Herrmann et al., 2024), eliciting commentary that evaluation in natural language understanding is largely "broken" (Bowman & Dahl, 2021) and that AI evaluation must "mature into a proper 'science'" (Weidinger et al., 2025). Benchmarking with static datasets is the standard framework for generative AI evaluation, but it faces multiple crises, e.g.: poor construct validity (Wallach et al., 2025; Alaa et al., 2025), data contamination (White et al., 2025), overfitting, minimal quality control, gaming, SOTA hacking, and selective reporting (Cheng et al., 2025). We observe several points of risk for construct validity in current reasoning evaluation strategies, including:

1. *A process and its product should not be conflated.*[3] Reasoning benchmarks often treat question-answering (QA) as a proxy for reasoning (Clark et al. 2018, *inter alia*). However, final-answer accuracy does not guarantee the mechanism by which the answer was generated (Zhang et al., 2025). We contend that (i) reasoning is a *process* and not an *output* (Figure 1), (ii) accurate QA final-answers can be obtained by *r*-zombies via non-reasoning behaviors, and thus (iii) accurate QA is not sufficient for demonstrating that reasoning has taken place.
2. *Chain-of-thought (CoT) is not trustworthy.* If intermediate reasoning steps are evaluated, CoT "reasoning traces" often serve as a stand-in for the LRM's internal reasoning process. However, CoT is neither necessary nor sufficient for obtaining trustworthy explanations (Barez et al., 2025). Though attractively anthropomorphic, CoT is not guaranteed to be faithful to the model's internal decision-making (Turpin et al., 2023; Lyu et al., 2023; Lanham et al., 2023; Kambhampati et al., 2025; Zhang et al., 2025). Mid-CoT shifts (e.g., "aha!" moments) may be rarer and less impactful than previously thought, reflecting unstable inference rather than true self-corrective reasoning (d'Aliberti & Ribeiro, 2026). We contend that an imperfect *r*-zombie could produce convincing but untrustworthy (or adversarial) CoT by emulating reasoning structure rather than content (Li et al., 2025).
3. *Reasoning evaluation should control for priors and experience.* Echoing Chollet (2019) on the measure of intelligence (Def. E.4), many reasoning benchmarks are easily gamed by instilling near-unlimited priors and experience through large-scale pre- and post-training. This is a core challenge in differentiating reasoning from recall in LRMs (Hüyük et al., 2025; Xu et al., 2025; Maasch et al., 2025a), raising the potential for *r*-zombies that lack reasoning yet are SOTA on benchmarks.

Usership Outpaces Trustworthiness Science is fundamentally a "collective epistemic enterprise," and as such, *epistemic trust* (Def. E.5) underpins scientific integrity (Wilholt, 2013). Epistemic trust in machine reasoning has been championed most in mathematical domains, as epitomized by the Lean language for automated theorem proving (De Moura et al., 2015). Lean addresses the "trust bottleneck" through formal verification, providing guarantees on correctness (Castelvecchi, 2023). However, the shift from deterministic systems and formal verification to probabilistic generative AI has raised new specters for epistemic trust, including evidence that hallucination is a feature and not a bug (Xu et al., 2024; Bastounis et al., 2024), accuracy collapse on tasks of scaling complexity (Shojaee et al., 2025), poor out-of-distribution generalization (Chollet et al., 2024; Mirzadeh et al., 2025; Xu et al., 2025), and low explainability. LLM-hallucinated citations (Shmatko et al., 2025; Sakai et al., 2026) and other sources of epistemic distrust in peer review at flagship AI conferences has elicited calls for reform (Kim et al., 2025). Rampant accusations of "AI hype" (Placani, 2024; MIT, 2025) coincide with broader linguistic trends: decreased hedging of uncertainty in scientific communication (Yao et al., 2023a) mirrors trends across diverse English text sources (Scheffer et al., 2021), reflecting a normalization of language that exaggerates confidence and obscures limitations. Meanwhile, 59% of AAAI survey respondents agreed that AI trustworthiness remains ill-defined, while 60% predicted that neither trustworthiness nor factuality would be solved in the near future (Rossi et al., 2025a). See Appendix D for further discussion.

2. Operationalizing Valid & Sound Reasoning

Defining reasoning is a nontrivial challenge spanning millenia (Appendix D.2). Broome (2013) admits that five years of iterative self-correction were required to reach an understanding of reasoning. Thus, it is unsurprising that re-

[3] We echo Chollet (2019) on the risks of "confusing the process of intelligence" (reasoning, in our case) "with the artifact produced by this process" (e.g., QA responses), ignoring the generating mechanism: "In the case of AI, the focus on achieving task-specific performance while placing no conditions on *how the system arrives at this performance* has led to systems that, despite performing the target tasks well, *largely do not feature the sort of human intelligence that the field of AI set out to build*" (original emphasis). Simon (2000) similarly argued that a theory of bounded rationality "will be as much concerned with... the quality of the processes of decision, as with... the quality of the outcome."

searchers can struggle to choose an authoritative, operational definition that is tailored for use in contemporary AI.

As a step toward addressing P1 and P2, we provide working definitions for reasoning that take a rule-centric perspective while remaining amenable to ML (Thesis 1, Thesis 2). Definitions are a synthesis of prior efforts from diverse domains. They are intended to be a starting point for discussion and a community reference for those that require practical, operational definitions. We first address *reasoning* and *reasoners* (§2.1) and then discuss *validity* and *soundness* (§2.3).

2.1. Working Definitions for Reasoning

2.1.1. Intuition in Natural Language

We begin with a definition in plain English to establish intuition. Colored terms denote core components.

Definition 2.1 (Reasoning, informal)**.** The *process of selecting and applying sequences of rules* that act on *prior beliefs* and *current evidence* to obtain *principled belief updates* in *evolving states*.

Definition 2.2 (Reasoner, informal)**.** A *goal-oriented* decision-maker that implements reasoning.

This conceptualization is closely related to arguments by Chollet (2019) that intelligence is a process and by Broome (2013) that reasoning is a process, "something a person *does*" (emphasis added), and a "rule-governed operation" (p. xii). Framing reasoning as a sequential process implies a notion of time t. We can conceptualize a time-dependent snapshot of the reasoner's internal world representation, which we refer to as the *state* at time t.[4]

Definition 2.3 (State, informal)**.** The set of all parameters that are pertinent to the reasoner at time t, including some subset of the historical record of beliefs, evidence, and rules.

We provide further intuition for each component of Def. 2.1.

Process Reasoning is a dynamic process, not an output. Thus, reasoning entails $T \geq 1$ hops, stages, time steps, or *reasoning steps*. This process implies a design component: sequences of rules or actions are chosen by the reasoner according to some justification. The process of *selection* is where agency, intelligence, or creativity may come into play, while the process of *execution* necessitates exactness and rigor. Note that it may be perfectly reasonable for the selection criterion to be random selection.

Goal The reasoner generally executes a reasoning process to achieve some outcome of interest. This outcome is the *goal* one is reasoning toward: the answer to a complex question, the solution to a puzzle, the shortest path through a maze, a mathematical proof, the optimal action to take under resource constraints, etc. In distinguishing the goal-directed reasoner from the reasoning process itself, we highlight that the *validity* of the reasoning process is not necessarily tied to successful attainment of a goal. In practice, we can encode the goal in a stopping rule, where reasoning terminates when the rule is satisfied. We do not restrict our notion of goals to the formal sense used in RL (Sutton & Barto, 1998), though it is compatible with this interpretation.

Rules Collectively, the rule set unambiguously maps the reasoning state at t to the state at $t+1$. In general, rules are selected with some justification prior to deployment. Rules can take the form of algorithms, formulae, theorems, axioms, laws, policies, premises, assumptions, decision boundaries, etc. Rules can be extrinsically imposed on the reasoner (i.e., hard-coded by another individual or collective agent, such as a human or government) or they can be learned autonomously from data on-the-fly. Rules can be fixed or continuously updated in light of new information.

Evidence We model evidence as a continuous stream of data that is updated at each step t or at intervals. *Current evidence* denotes information presented at t, along with the historical record: aggregated information up to $k \geq 0$ steps prior to t. Evidence may be gained directly through sequential interactions with an uncertain environment (as in online RL, field work in the natural sciences, etc.) or provided without direct collection (e.g., retrospective data collected by another agent). In trivial cases, external evidence is the empty set or is provided at $t = 0$ and never updated.

Prior Beliefs While evidence is a form of exogenous or extrinsically obtained knowledge, we model beliefs as a form of endogenous or intrinsically obtained knowledge. *Prior beliefs* are the outputs of previous reasoning steps, up to step $t - k$ for $t > k \geq 1$. They can be viewed as intermediate conclusions along the reasoning pathway that led to step t. Often, they are defeasible: they can be overwritten if proven false (e.g., in backtracking proof search), refined if insufficient, or maintained and aggregated with current beliefs at step t. They can also be provided at $t = 0$ (e.g., initializing Bayesian priors based on convention when supporting evidence is not yet available). See Pearl (1990) for another account of *belief* as informed by *evidence*.

Current Beliefs Current beliefs denote the conclusions drawn in the transition from $t-1$ to t. When $t = T$, current belief is equivalent to the *terminal conclusion* of the reasoning process. The nature of the terminal conclusion is a defining property of the type of reasoning performed, e.g.: the output of a function in mathematical reasoning,

[4] Note that the state is not necessarily a *world model* as commonly conceived in RL or structural causal modeling (Richens & Everitt, 2024; Richens et al., 2025; Maasch et al., 2025b): it is not necessarily predictive of the dynamics governing an evolving environment nor sufficient for causal identifiability. Further, it may be only partially observed or partially stored in memory.

an optimal action in practical reasoning, a moral verdict in moral reasoning, a judiciary decision in legal reasoning, etc.

EVOLVING STATES A reasoner will generally maintain an *internal representation* of its world state (Def. 2.3), which updates over time. The existence of an *external environment* is also implied by our choice to model evidence as a stream of extrinsic signals. However, we note that a well-defined concept of external environment is not relevant in all cases (e.g., in some mathematical reasoning domains). Thus, we place no requirements on the existence or direct observability of an external environment, physical world, etc., and only require an internal representation of the world (i.e., the state). We use the notion of an *evolving state* broadly to encode all of the above concepts: (1) dynamically updated internal state representations, (2) changing and/or uncertain external worlds, and (3) extrinsic sources of evidence.

2.1.2. A FORMAL OPERATIONAL DEFINITION

While natural language is useful for high-level intuition, it is too ambiguous to formally convey measurable definitions in the general case (Thesis 1). We offer a formalization of Def. 2.1 in mathematical notation and pseudocode. While Def. 2.4 is an operationalization of Def. 2.1, this is not the only operational definition that could follow from Def. 2.1.

Definition 2.4 (Reasoning, formal)**.** Let $\mathcal{S}_t \coloneqq \langle \mathcal{B}_t, \mathcal{E}_t, \mathcal{R}_t \rangle$ denote the reasoner's state at time step t, where $\mathcal{B}_t$ denotes current belief, $\mathcal{E}_t$ denotes aggregated evidence up to time t, and $\mathcal{R}_t$ denotes the current set of established rules. Then, *reasoning* is the iterated application over steps t of rules $r \in \mathcal{R}_{t-1}$ to prior beliefs $\mathcal{B}_{t-1}$ and current evidence $\mathcal{E}_t$, by which we obtain dynamically updated states $\mathcal{S}_t$, and where every output $\mathcal{B}_t$ for $t > 0$ is the result of a rule application $r(\mathcal{B}_{t-1}, \mathcal{E}_t)$ to the contents of state $\mathcal{S}_{t-1}$.

Thus, rules and extrinsic evidence updates are the mechanism by which $\mathcal{S}_t$ changes over time: each $r \in \mathcal{R}_t$ is a function acting on subsets of the current state $\mathcal{S}_t$ to generate some attribute of the next state $\mathcal{S}_{t+1}$.

Rule set $\mathcal{R}$, beliefs $\mathcal{B}$, and evidence $\mathcal{E}$ comprising state $\mathcal{S}$ are each elements of a corresponding space $\mathbf{R}$, $\mathbf{B}$, $\mathbf{E}$, and $\mathbf{S}$. $\mathcal{R}$ is a set of functions, with domains and ranges as defined below. Other implementation details, constraints, and type systems defining these spaces are problem-specific.

Definition 2.5 (Reasoning components)**.**

$t \in [0, ..., T]$	Reasoning step.
$\{\mathcal{B}_i\}_{i=0}^{T},\ \mathcal{B}_i \in \mathbf{B}$	Beliefs.
$\{\mathcal{E}_i\}_{i=0}^{T},\ \mathcal{E}_i \in \mathbf{E}$	Evidence.
$\{\mathcal{R}_i\}_{i=0}^{T},\ \mathcal{R}_i \in \mathbf{R}$	Rule set.
$\mathcal{S}_i \coloneqq \langle \mathcal{B}_i, \mathcal{E}_i, \mathcal{R}_i \rangle,\ \mathcal{S}_i \in \mathbf{S}$	States.

The rule set is partitioned into two sets of functions with distinct type signatures — local rules $\mathcal{R}^L$, which update beliefs, and meta rules $\mathcal{R}^M$, which update rules:

$$\mathcal{R}_t^L \coloneqq \{r \in \mathcal{R}_t \mid r : \mathbf{B} \times \mathbf{E} \to \mathbf{B}\}$$
$$\mathcal{R}_t^M \coloneqq \{r \in \mathcal{R}_t \mid r : \mathbf{R} \times \mathbf{B} \times \mathbf{E} \to \mathbf{R}\}$$

where $\mathcal{R}_t^L \cap \mathcal{R}_t^M = \emptyset$ and $\mathcal{R}_t^L \cup \mathcal{R}_t^M = \mathcal{R}_t$. The rule set may include *identity rules*, which trivially return the rules or beliefs from time t at time $t+1$:

$$I^M \in \mathcal{R}_1^M \text{ such that } I^M(\mathcal{R}, \mathcal{B}, \mathcal{E}) = \mathcal{R} \text{ and}$$
$$I^L \in \mathcal{R}_1^L \text{ such that } I^L(\mathcal{B}, \mathcal{E}) = \mathcal{B}$$

for any $(\mathcal{R}, \mathcal{B}, \mathcal{E}) \in \mathbf{S}$. State updates $\mathcal{S}_t \to \mathcal{S}_{t+1}$ are defined by the receipt of new evidence $\mathcal{E}_{t+1}$, if any, followed by a sequence of two[5] rule applications:

$$\mathcal{B}_{t+1} = r^L(\mathcal{B}_t, \mathcal{E}_{t+1}) \text{ for some } r^L \in \mathcal{R}_t^L$$
$$\mathcal{R}_{t+1} = r^M(\mathcal{R}_t, \mathcal{B}_{t+1}, \mathcal{E}_{t+1}) \text{ for some } r^M \in \mathcal{R}_t^M$$
$$\mathcal{S}_{t+1} \coloneqq \langle \mathcal{R}_{t+1}, \mathcal{B}_{t+1}, \mathcal{E}_{t+1} \rangle.$$

In order to specify a reasoning algorithm (Algorithm 1), we introduce the concept of a *rule selector function*. Because these functions do not impact whether or not a process constitutes reasoning, we define them separately in Def. 2.6, as part of the *reasoner's implementation* of a reasoning process. A full implementation may also involve additional components, such as a goal (or "stopping rule") and a trace recording historical reasoning steps, as specified in Def. 2.6.

Definition 2.6 (Reasoner components)**.** A reasoner can contain or generate the following elements (among others), which are extrinsic to the reasoning process itself.

$\mathtt{s}_{\mathtt{L}} : \mathbf{R} \times \mathbf{B} \times \mathbf{E} \to \mathcal{R}^L$	Local rule selector.
$\mathtt{s}_{\mathtt{M}} : \mathbf{R} \times \mathbf{B} \times \mathbf{E} \to \mathcal{R}^M$	Meta rule selector.
$\mathtt{s}_{\mathtt{stop}} : \mathbf{S} \to \{0, 1\}$	Stopping rule.
$\mathtt{tr} : \mathbf{S} \times \mathcal{R}^L \times \mathcal{R}^M \times \mathbf{S} \to \Sigma^*$	Trace writer.
$\mathcal{T} \coloneqq \left\{\mathtt{tr}\left(\mathcal{S}_{i-1}, r_i^L, r_i^M, \mathcal{S}_i\right)\right\}_{i=1}^{T}$	Reasoning trace.

where $r_i^L \coloneqq \mathtt{s}_{\mathtt{L}}(\mathcal{R}_t, \mathcal{B}_t, \mathcal{E}_{t+1})$ and $r_i^M \coloneqq \mathtt{s}_{\mathtt{M}}(\mathcal{R}_t, \mathcal{B}_t, \mathcal{E}_{t+1})$.

Rule selectors use the current state's rules and beliefs along with any new evidence, and output a single rule. The black-box nature of the rule selectors in Def. 2.6 is powerful: the freedom to implement selectors in any way (hard-coding, learning from data, or hybrid) is the bridge between symbolic and ML interpretations of reasoning. The *stopping rule* uses the current state to output a boolean expressing whether or not to end the reasoning process. Often, the stopping rule will encode the end-goal of reasoning, evoking the goal-directed nature of a reasoner under Def. 2.2. The

[5] Multiple belief updates in immediate sequence can be implemented by setting the corresponding rule updates to the identity.

Algorithm 1 Valid reasoning as exact rule application.

Input. Initial rules $\mathcal{R}_0$, beliefs $\mathcal{B}_0$, evidence stream $\{\mathcal{E}_i\}_{i=1}^{T}$.

$\mathcal{R}, \mathcal{B}, \mathcal{E} \leftarrow \mathcal{R}_0, \mathcal{B}_0, \mathcal{E}_0$
$\mathcal{S} \leftarrow (\mathcal{R}, \mathcal{B}, \mathcal{E})$
$t \leftarrow 0$
while not $\mathtt{s}_{\mathtt{stop}}(\mathcal{S})$ **do**
  $\mathcal{E}' \leftarrow \mathcal{E}_{t+1}$
  $r^L \leftarrow \mathtt{s}_{\mathtt{L}}(\mathcal{R}, \mathcal{B}, \mathcal{E}')$ {Select local rule.}
  $\mathcal{B}' \leftarrow r^L(\mathcal{B}, \mathcal{E}')$ {Apply local rule, update beliefs.}
  $r^M \leftarrow \mathtt{s}_{\mathtt{M}}(\mathcal{R}, \mathcal{B}', \mathcal{E}')$ {Select meta rule.}
  $\mathcal{R}' \leftarrow r^M(\mathcal{R}, \mathcal{B}', \mathcal{E}')$ {Apply meta rule, update rules.}
  $\mathcal{S}' \leftarrow (\mathcal{R}', \mathcal{B}', \mathcal{E}')$
  $\mathcal{T}.\mathtt{append}(\mathtt{tr}(\mathcal{S}, r^L, r^M, \mathcal{S}'))$ {Update trace.}
  $\mathcal{R}, \mathcal{B}, \mathcal{E}, \mathcal{S} \leftarrow \mathcal{R}', \mathcal{B}', \mathcal{E}', \mathcal{S}'$
  $t \mathrel{+}= 1$
end while
Return $\mathcal{B}, \mathcal{T}$

trace writer considers the selected rules and resulting state change, and optionally outputs a string (using alphabet Σ) to include in the *reasoning trace*.

With these definitions in place, we describe a generalized *reasoning algorithm* in Algorithm 1.

Remark 2.1 (How is Definition 2.4 operational?)**.** Operationalization of Def 2.4 and Algorithm 1 is confirmed by Python implementation.[6] In general, the onus is on the researcher to map the core components of Def 2.4 to the unique problem setting, justify the absence of any components, and confirm that process validity is present. We provide a checklist of best practices in Appendix A.

2.2. Examples from Domain-Specific Reasoning

Defs. 2.1 and 2.4 are intentionally broad, and indeed a large number of phenomena could be said to satisfy them. To illustrate their flexibility, we map them to specific forms of reasoning that are commonly encountered in mathematics, computer science, and AI. We consider these specific forms of reasoning to be special cases of Def. 2.4 that vary in how rules, beliefs, and evidence are defined or obtained. See Appendix B for additional examples. Table B.1 compares all examples by the nature of rules, beliefs, and evidence. For strong examples of operational definitions for reasoning in mathematics, see Defs. 1 and 2 in Zhang et al. (2025).

Example 2.1 (Logical deduction)**.** Our framework is heavily inspired by deductive systems (e.g., Hilbert systems, sequent calculi, natural deduction, or resolution calculi) over classical first-order logic, although it is not limited to these settings.[7] Concretely, a *natural deductive system* over a formal language is initialized with a set of premises Γ, and a static set of inference rules (e.g., *modus ponens* or *modus tollens*) acting on premises. A derivation (*deduction*) of a conclusion φ is a finite sequence of premises where each is either in Γ, or obtained from earlier formulas in the sequence by application of an inference rule. If such a derivation exists, φ satisfies the consequence relation $\Gamma \vdash \varphi$. Derivations yield a monotonically increasing belief set in the closure of Γ under the logical consequence relation.

We note that natural deductive systems are a highly restricted instantiation of Def. 2.4, such that no new evidence is provided ($\mathcal{E}_i = \emptyset \; \forall \; i$), and the set of inference rules is fixed ($\mathcal{R}_i^M = \{I^L\} \; \forall \; i$). Logical systems other than classical first-order logic can also be expressed under Def. 2.4; see, for example, nonmonotonic logic in Appendix B.1, which allows for principled belief retraction.

Example 2.2 (Bayesian inference)**.** Bayesian inference provides principled means of revising beliefs in hypotheses as new evidence emerges. We iteratively refine posterior estimate $p(\theta \mid \mathcal{D})$ for unknown parameters θ by repeatedly applying Bayes' rule (Equation 1) as our prior over θ and observed data $\mathcal{D}$ update across time t:

$$r_{bayes} := \left\{ p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})} \right\}. \tag{1}$$

Conclusion $p(\theta \mid \mathcal{D})$ is always valid when r_{bayes} is applied, though it might be biased with respect to ground truth.

Example 2.3 (Reinforcement learning)**.** RL is the ML paradigm concerned with training optimal goal-directed decision-makers (i.e., *agents*) through sequential interactions with an uncertain environment. The agent learns a *policy* that maps states to actions. Thus, the RL agent meets our informal definition of a goal-oriented reasoner (Def. 2.2). Update rules in RL often take the following form (Sutton & Barto 1998, p. 37):

$$r_{update} := \varphi_{new} \leftarrow \varphi_{old} + \alpha(\tau - \varphi_{old}) \tag{2}$$

where φ is some estimate, α is step size, τ is the target or a desirable (yet perhaps noisy) direction (e.g., the reward), and $(\tau - \varphi_{old})$ is an estimation error. For example, we can estimate the agent's reward for some action at step $t+1$ as

$$Q_{t+1} = \frac{1}{t} \sum_{i=1}^{t} R_i = Q_t + \frac{1}{t}[R_t - Q_t] \tag{3}$$

where Q_t is the estimated t^{th} reward (prior belief) and R_t is the observed t^{th} reward (evidence).

2.3. Validity & Soundness

We now elaborate on *valid* and *sound* reasoning. Validity is meant to replace our heuristic use of *true* reasoning with

[6] https://anonymous.4open.science/r/valid_reasoning/

[7] Deductive systems encompass proof systems and formal semantics for zeroth, first, and higher-order logics, and additionally form the basis for automated theorem provers, SMT solvers, and proof assistants; each of which satisfy Def. 2.4.

a more concrete concept: any superficially reasoning-like behavior that does not satisfy Def. 2.7 *is not reasoning*, though it may be useful reasoning emulation.

Definition 2.7 (Validity)**.** A transition from state $\mathcal{S}_t$ to $\mathcal{S}_{t+1}$ is *valid* if and only if it arises from the application of a rule $r \in \mathcal{R}_t$ to components of state $\mathcal{S}_t$.

Following from Def. 2.7, we claim the following.

Claim 2.1 (Valid reasoning arises from exact rule application)**.** Validity requires that each rule is always executed exactly: not partially, not approximately, not sometimes. This does not preclude rule-based means of handling stochasticity, uncertainty, and approximate inference.

We can use Def. 2.7 to further clarify our definition of an r-zombie: a system that generates reasoning-like output but lacks the mechanisms necessary for *validity*. Claim 2.1 is in line with treatments in symbolic AI, as well as recent work toward LRM reasoning and AGI: Zhang et al. (2025) claim that "operations must be **exact**" in reasoning (original emphasis), while György et al. (2025) argue for a shift away from statistical learning toward *exact learning* to address LRM "jagged intelligence" (Grand et al., 2025). See Objection 3.1 and Appendix D.2 for further discussion of the history and revival of rule-based reasoning.

Unlike valid reasoning, *soundness* requires a notion of correctness or alignment with respect to external assessments.

Definition 2.8 (Soundness)**.** A valid transition from state $\mathcal{S}_t$ to $\mathcal{S}_{t+1}$ is *sound* if and only if all premises (as encoded by $\mathcal{B}$, $\mathcal{R}$, and $\mathcal{E}$) are true with respect to *external evaluation*.

Validity and soundness are classically used to evaluate logical arguments. While sound reasoning is always valid, valid reasoning need not be sound. This gives way to Claim 2.2.

Claim 2.2 (Validity is independent of rule selection)**.** Implementing a reasoning process requires selecting which specific rule to apply at each step. Because validity is independent of soundness, and any properly-typed rule application creates a valid output, the validity of a reasoning process is independent of the algorithm used to select the rule sequence, regardless of external ground truth.

Claim 2.2 echoes Broome's (2013) *correctness-by-permissibility*: " Correct reasoning is not reasoning you are *required* to do by rationality, but reasoning you are *permitted* to do by rationality" (p. xii; emphasis added). Emphasizing validity over soundness allows for *bounded rationality* in reasoning, where incomplete information and uncertainty can lead the reasoner's conclusions to be "as much determined by the 'inner environment' [our notion of *state*]... as by the 'outer environment'" (Simon, 2000). Claim 2.2 highlights that validity says nothing of the optimality, usefulness, nor external correctness of the reasoning process, and indeed, ground truth may not exist.[8] For example, the rule selector could select rules at random, act adversarially, or always return the identity function, and yet the process would still be valid.

2.4. Implications of Definition 2.4

Claim 2.3 (Rules are learnable)**.** We contend that rule-governed reasoning and data-driven models are not mutually exclusive, and we reject the false dichotomy between rule-based symbolic AI and contemporary probabilistic deep learning (as summarized in Appendix D.2). *Learnable rules* are essential for tying Def. 2.4 to modern AI and the bitter lesson (Sutton, 2019): rules do not need to be hard-coded by human domain experts, and the future of autonomous reasoning will likely include systems that learn rules and defeasible beliefs on-the-fly. See Oh et al. (2025), in which an artificial agent autonomously discovered a SOTA RL rule that outperformed human-designed rules.

Claim 2.4 (Rules are explanations)**.** The explainability of a reasoning process lies in the rule set, as rules are the justifications by which each intermediate reasoning step is executed. In this conceptualization, *rules themselves are explanations* for how the reasoner reached conclusions. By extension, the absence or unobservability of a rule set results in poor explainability. We contrast this notion of rules-as-explanations with CoT, which is neither necessary nor sufficient for explainability (see §1.1).

Claim 2.5 (Operationalization facilitates trust)**.** A central aspect of trust is the accurate representation of the capabilities or expected behavior of a system (Kaur et al., 2022). Validity formalizes an expectation found in many common definitions of reasoning (see Appendix C for further discussion). Claims of "reasoning" applied to models which fail to meet a minimal bar of validity thus endanger trust. Similarly, claims about "reasoning" without a clear operationalization of the term leave validity and soundness unfalsifiable.

Claim 2.6 (Reasoning requires memory)**.** Notions of prior beliefs, evidence, and rules imply the existence of *memory*, as this body of information must be stored and recalled. This does not preclude special cases of *memoryless* or *Markovian* reasoning processes where all information needed at step t is contained in $\mathcal{S}_{t-1}$, as these rely on a persistent representation of the immediately preceding state. Several proposals for autonomous machine intelligence (LeCun,

[8] The import of Claim 2.2 is clear in settings where there is no singular objective truth, as it permits disagreement and subjectivity. Consider pluralism in moral reasoning (Snoswell et al., 2026): two moral actors with conflicting moral frameworks could both be said to validly reason even if their moral conclusions differ, as long as both exactly apply their respective moral rules. Plurality can also arise in sound reasoning: a single problem often admits multiple sound reasoning paths (Wang et al., 2023), though some paths may be more useful; see González & Nori (2024) and Maasch et al. (2025a), which use commutative diagrams to model this case.

2022), AGI (Hendrycks et al., 2025), and transformer-based LRMs (Cheng et al., 2026) explicitly emphasize memory or persistent state as a core component of intelligent behavior.

Claim 2.7 (Natural language is not necessary for reasoning)**.** Defs. 2.1 and 2.4 do not imply a necessary role of natural language in AI reasoning. Similarly, Broome (2013) does not assume that natural language is necessary for human reasoning. Evidence from neuroscience suggests that language may not be required for complex symbolic thought (Fedorenko et al., 2024), deductive reasoning (Coetzee et al., 2022), nor mathematical and logical reasoning (Fedorenko & Varley, 2016). Increasing, neural methods explore reasoning in latent space rather than language space (Hao et al. 2025; Zhu et al. 2025; Wang et al. 2025; *inter alia*).

3. Alternative Views

See Appendix C for an extended discussion of alternative definitions for reasoning from diverse domains. Here, we comment on mainstream objections to our core theses. While Objection 3.1 argues that *this perspective won't work*, Objection 3.2 argues that *this perspective isn't needed*.

Objection 3.1. *(1) Def. 2.4 violates Sutton's bitter lesson (Sutton, 2019), (2) symbolic AI has already failed, and (3) scaling is all you need.* We observe several variations of these arguments on rule-based systems, which rightfully highlight the knowledge acquisition bottlenecks and lack of generalization in classical expert systems. **Rebuttal: Points (1) and (2) are false, and (3) is speculative.** We acknowledge the historical context of an "AI winter" following the "first wave" of AI, in contrast to the groundbreaking successes of AI's "second wave" (Fouse et al. 2020; Appendix D.2). We understand that this context may raise skepticism about the feasibility of designing systems that meet our standard of process validity. However, we contend that our theses are compatible with symbolic methods, ML, and neuro-symbolic AI. Per Claim 2.3, the learnability of rules makes Def. 2.4 amenable to contemporary ML. Because Def. 2.4 does not require hardcoding nor injection of human domain expertise, it is not incompatible with the bitter lesson. Though recent advances in generative AI are compelling, outright rejection of symbolic methods is *throwing the baby out with the bathwater*, while claiming that symbolic methods have failed is *forgetting history*. See Lean, a symbolic system for gold-standard automated theorem proving (De Moura et al., 2015); the neuro-symbolic AlphaGeometry 2 (Chervonyi et al., 2025) and AlphaProof (Hubert et al., 2025), which can solve Olympiad-level math; *inter alia*. While scaling model parameters, data size, and inference-time compute has resulted in profound performance gains (Kaplan et al., 2020; Bi et al., 2024; Muennighoff et al., 2025), it remains pure speculation whether scaling is sufficient to reach various goals. Scale has not yet resolved hallucination, explainability, out-of-distribution generalization, or other factors that undermine trustworthy reasoning. A AAAI survey found that 76% of respondents believed "scaling up current AI approaches" was "unlikely" to "very unlikely" to produce AGI (Rossi et al., 2025a).

Objection 3.2. *Empirical performance matters more than theoretical guarantees, so rule-based validity is a waste of time.* We observe a common argument that (1) benchmark accuracy is a sufficient proxy for reasoning and (2) if empirical evaluations yield consistently high scores, then the underlying process is of lesser concern. **Rebuttal: Sometimes yes, sometimes no.** Circling back to the discussion in §1, an essential task in contemporary AI will be thoughtfully delineating where reasoners are required and where (im)perfect r-zombies are sufficient. Relatedly, "How deep and how reliable does the reasoning have to be in order to do certain important things?" (Holger Hoos in Rossi et al. 2025b). Indeed, sometimes "the best is the enemy of the good" and "optimizing is the enemy of satisficing" (Simon, 2000). However, the fallibility of empirical evaluation becomes especially problematic under distribution shift, rare events, adversarial attack, and safety-critical or high-stakes domains. As discussed in §1.1, benchmarks have finite coverage and are prone to design flaws (Wallach et al., 2025; Alaa et al., 2025; White et al., 2025; Cheng et al., 2025), including the conflation of process (reasoning) with the product of that process (QA accuracy, etc.) (Chollet, 2019). As commonly held in formal logic, we hold validity as a prerequisite for soundness (Def. 2.8), so any domain requiring external correctness will require validity guarantees. We contend that many scientifically and economically important use cases require validity, including many decision-making systems with safety or fairness implications (e.g., in medicine, policing, etc.).

4. Conclusion: A Call to Action

Based on a synthesis of the literature, we propose an operational definition for rule-based reasoning that remains amenable to modern ML. However, this is not the only operational definition that can provide research value. **We urge the community to engage with our definitions and claims, identify shortcomings, and propose alternatives.** We encourage applying our checklist of best practices (Appendix A) to any reasoning-related research or product communication, with a particular focus on domain-specific operationalization. We recommend that researchers and marketing professionals alike **refrain from using "reasoning" to describe processes which fall short of the standard of process validity** and instead reference formalizations of *reasoning emulation* (e.g., *r*-zombies). We further call AI researchers and engineers to prioritize *trustworthy reasoning* in future research and product releases. Echoing calls for "interpretability by design" in mechanistic interpretability (Sharkey et al., 2025), **we strongly encourage researchers**

to build AI reasoning systems with *validity by design*, particularly in domain-specific settings where process validity is legally, ethically, practically, or mathematically mandated.

Acknowledgments

The authors thank Dr. Helen Nissenbaum for insightful discussions on epistemic trust in AI. Author J. Maasch acknowledges the Cornell Tech Digital Life Initiative Fellowship and the US National Science Foundation Graduate Research Fellowship under Grant No. DGE–2139899.

References

Alaa, A., Hartvigsen, T., Golchini, N., Dutta, S., Dean, F., Raji, I. D., and Zack, T. Position: Medical large language model benchmarks should prioritize construct validity. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

American Psychological Association. APA Dictionary of Psychology, "operational definition". https://dictionary.apa.org/operational-definition. Accessed: 2026-01-22.

Barez, F., Wu, T.-Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., et al. Chain-of-thought is not explainability. *Preprint*, 2025.

Bastounis, A., Campodonico, P., van der Schaar, M., Adcock, B., and Hansen, A. C. On the consistent reasoning paradox of intelligence and optimal trust in ai: The power of'i don't know'. *arXiv preprint arXiv:2408.02357*, 2024.

Baumann, J., Urman, A., Leicht-Deobald, U., Roman, Z. J., Hannák, A., and Christen, M. Reduced ai acceptance after the generative ai boom: evidence from a two-wave survey study. *arXiv preprint arXiv:2510.23578*, 2025.

Belle, V. and Marcus, G. The future is neuro-symbolic: Where has it been, and where is it going? In *The 40th Annual AAAI Conference on Artificial Intelligence*, 2025.

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Blanchette, J. C., Kaliszyk, C., Paulson, L. C., and Urban, J. Hammering towards qed. *Journal of Formalized Reasoning*, 9(1):101–148, 2016.

Blili-Hamelin, B., Graziul, C., Hancox-Li, L., Hazan, H., El-Mhamdi, E.-M., Ghosh, A., Heller, K., Metcalf, J., Murai, F., Salvaggio, E., et al. Position: Stop treating agi as the north-star goal of ai research. *International Conference on Machine Learning*, 2025.

Bobzien, S. Ancient Logic. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020.

Bowman, S. and Dahl, G. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 2021.

Boyer, R. S. and Moore, J. S. Proving theorems about lisp functions. *Journal of the ACM (JACM)*, 22(1):129–144, 1975.

Britannica, E. Nyaya. In *Encyclopedia Britannica*. 2017. Accessed 27 January 2026.

Broome, J. The unity of reasoning? *Spheres of reason*, 1 (9):62–93, 2009.

Broome, J. *Rationality Through Reasoning*. John Wiley & Sons, 2013.

Castelvecchi, D. How will ai change mathematics? *Nature*, 615:15–16, 2023.

CDEI. Public attitudes to data and ai: Tracker survey (wave 3). https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-3/public-attitudes-to-data-and-ai-tracker-survey-wave-3, December 2023.

Challapally, A., Pease, C., Raskar, R., and Chari, P. The genai divide: State of ai in business 2025. 2025. URL https://www.media.mit.edu/projects/mit-nanda/overview/.

Chalmers, D. Spatiotemporal functionalism v. the conceivability of zombies. *Noûs*, 54(2):488–497, 2020.

Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford Paperbacks, 1997.

Cheng, X., Zeng, W., Dai, D., Chen, Q., Wang, B., Xie, Z., Huang, K., Yu, X., Hao, Z., Li, Y., et al. Conditional memory via scalable lookup: A new axis of sparsity for large language models. *arXiv preprint arXiv:2601.07372*, 2026.

Cheng, Z., Wohnig, S., Gupta, R., Alam, S., Abdullahi, T., Ribeiro, J. A., Nielsen-Garcia, C., Mir, S., Li, S., Orender, J., et al. Benchmarking is broken–don't let ai be its own judge. *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Chervonyi, Y., Trinh, T. H., Olšák, M., Yang, X., Nguyen, H. H., Menegali, M., Jung, J., Kim, J., Verma, V., Le, Q. V., et al. Gold-medalist performance in solving

olympiad geometry with alphageometry2. *Journal of Machine Learning Research*, 26(241):1–39, 2025.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Chollet, F., Knoop, M., Kamradt, G., and Landers, B. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Coetzee, J. P., Johnson, M. A., Lee, Y., Wu, A. D., Iacoboni, M., and Monti, M. M. Dissociating language and thought in human reasoning. *Brain Sciences*, 13(1):67, 2022.

Coquand, T. and Huet, G. *The calculus of constructions*. PhD thesis, INRIA, 1986.

Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

d'Aliberti, L. G. and Ribeiro, M. H. The illusion of insight in reasoning models, 2026. URL https://arxiv.org/abs/2601.00514.

Davis, R. and King, J. J. The origin of rule-based systems in ai. *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*, 1984.

de Bruijn, N. G. Automath, a language for mathematics. In *Automation of Reasoning: 2: Classical Papers on Computational Logic 1967–1970*, pp. 159–200. Springer, 1983.

De Moura, L., Kong, S., Avigad, J., Van Doorn, F., and von Raumer, J. The lean theorem prover (system description). In *International Conference on Automated Deduction*, pp. 378–388. Springer, 2015.

Evans, J. S. B. and Stanovich, K. E. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

Fagin, R. and Halpern, J. Y. Belief, awareness, and limited reasoning. *Artificial intelligence*, 34(1):39–76, 1987.

Fagin, R. and Halpern, J. Y. Reasoning about knowledge and probability. *Journal of the ACM (JACM)*, 41(2):340–367, 1994.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. *Reasoning about knowledge*. MIT press, 2004.

Fedorenko, E. and Varley, R. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369(1):132–153, 2016.

Fedorenko, E., Piantadosi, S. T., and Gibson, E. A. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024.

Fonagy, P. and Allison, E. *The role of mentalizing and epistemic trust in the therapeutic relationship.*, volume 51. Educational Publishing Foundation, 2014.

Fouse, S., Cross, S., and Lapin, Z. Darpa's impact on artificial intelligence. *AI Magazine*, 41(2):3–8, Jun. 2020. doi: 10.1609/aimag.v41i2.5294. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/5294.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Garcez, A. d. and Lamb, L. C. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023.

Gärdenfors, P. *Knowledge in flux: Modeling the dynamics of epistemic states*. The MIT press, 1988.

Gillon, B. Logic in Classical Indian Philosophy. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2024 edition, 2024.

González, J. and Nori, A. Does reasoning emerge? examining the probabilities of causation in large language models. *Advances in Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Gordon, M. Hol: A machine oriented formulation of higher order logic. Technical report, University of Cambridge, Computer Laboratory, 1985.

Grand, G., Tenenbaum, J. B., Mansinghka, V. K., Lew, A. K., and Andreas, J. Self-steering language models. *Conference on Language Modeling (COLM 2025)*, 2025.

Guan, L., Valmeekam, K., Sreedharan, S., and Kambhampati, S. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094, 2023.

György, A., Lattimore, T., Lazić, N., and Szepesvári, C. Beyond statistical learning: Exact learning is essential for general intelligence, 2025. URL https://arxiv.org/abs/2506.23908.

Halpern, J. Y. *Reasoning about uncertainty*. 2017.

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J. E., and Tian, Y. Training large language models to reason in a continuous latent space. In *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models*, 2025.

Hayes-Roth, F. Rule-based systems. *Communications of the ACM*, 28(9):921–932, 1985.

Hendrycks, D., Song, D., Szegedy, C., Lee, H., Gal, Y., Brynjolfsson, E., Li, S., Zou, A., Levine, L., Han, B., et al. A definition of agi. *arXiv preprint arXiv:2510.18212*, 2025.

Herrmann, M., Lange, F. J. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., Rügamer, D., Hüllermeier, E., Boulesteix, A.-L., and Bischl, B. Position: Why we must rethink empirical research in machine learning. In *International Conference on Machine Learning*, pp. 18228–18247. PMLR, 2024.

Hieronymi, P. The use of reasons in thought (and the use of earmarks in arguments). *Ethics*, 124(1):114–127, 2013.

Howard, W. A. et al. The formulae-as-types notion of construction. *To HB Curry: essays on combinatory logic, lambda calculus and formalism*, 44:479–490, 1980.

Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, 2023.

Hubert, T., Mehta, R., Sartran, L., Horváth, M. Z., Žužić, G., Wieser, E., Huang, A., Schrittwieser, J., Schroecker, Y., Masoom, H., et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, pp. 1–3, 2025.

Humphrey, R. *Stream of Sonsciousness in the Modern Novel*, volume 3. University of California Press, 1954.

Hüyük, A., Xu, X., Maasch, J., Nori, A. V., and González, J. Reasoning elicitation in language models via counterfactual feedback. *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.03767.

Irzik, G. and Kurtulmus, F. What is epistemic public trust in science? *The British Journal for the Philosophy of Science*, 2019.

James, W. The principles of psychology. *Henry Holt*, 1890.

Jurafsky, D. and Martin, J. H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models (third edition draft), 2025.

Kahneman, D. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

Kambhampati, S., Stechly, K., Valmeekam, K., Saldyt, L., Bhambri, S., Palod, V., Gundawar, A., Samineni, S. R., Kalwar, D., and Biswas, U. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Bridging Language, Agent, and World Models (LAW)*, 2025. URL https://arxiv.org/abs/2504.09762.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.

Kim, J., Lee, Y., and Lee, S. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Kolodny, N. Why be rational? *Mind*, 114(455):509–563, 2005.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, 2023.

LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Preprint*, 2022.

Li, D., Cao, S., Griggs, T., Liu, S., Mo, X., Tang, E., Hegde, S., Hakhamaneshi, K., Patil, S. G., Zaharia, M., et al. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*, 2025.

Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., and Su, H. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 2023.

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

Lloyd, J. W. *Foundations of logic programming*. Springer Science & Business Media, 2012.

Lukyanenko, R., Maass, W., and Storey, V. C. Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. *Electronic Markets*, 32(4):1993–2020, 2022.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329, 2023.

Maasch, J., Hüyük, A., Xu, X., Nori, A. V., and Gonzalez, J. Compositional causal reasoning evaluation in language models. *International Conference on Machine Learning*, 2025a.

Maasch, J., Kalantari, J., and Khezeli, K. Causalarc: Abstract reasoning with causal world models. *39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Bridging Language, Agent, and World Models (LAW)*, 2025b. URL https://arxiv.org/abs/2509.03636.

Macfarlane, M. V. and Bonnet, C. Searching latent program spaces. *39th Conference on Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2411.08706.

Manuvinakurike, R., Moss, E., Watkins, E. A., Sahay, S., Raffa, G., and Nachman, L. Thoughts without thinking: Reconsidering the explanatory value of chain-of-thought reasoning in llms through agentic pipelines. *arXiv preprint arXiv:2505.00875*, 2025.

Markie, P. and Folescu, M. Rationalism vs. Empiricism. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition, 2023.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., et al. In *Artificial Intelligence Index Report 2025*. Stanford University Human-Centered Artificial Intelligence, 2025.

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E., and Morimoto, J. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275, 2022.

McKinsey. The state of ai in 2025: Agents, innovation, and transformation. 2025. URL https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/.

Merriam-Webster. Reasoning. In *Merriam-Webster.com Dictionary*. 2026. Accessed 27 January 2026.

Merrill, W. and Sabharwal, A. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.

Merrill, W., Sabharwal, A., and Smith, N. A. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022.

Mirzadeh, S. I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, 2025.

MIT. The great ai hype correction of 2025. *MIT Technology Review*, 2025.

Mitchell, M. Artificial intelligence learns to reason. *Science*, 387(6740):eadw5211, 2025a.

Mitchell, M. On the science of "alien intelligences": Evaluating cognitive capabilities in babies, animals, and ai, 2025b. URL https://neurips.cc/virtual/2025/loc/san-diego/invited-talk/109607. Invited talk at The Thirty-Ninth Annual Conference on Neural Information Processing Systems. San Diego, California.

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. B. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Neelakantan, A., Le, Q. V., and Sutskever, I. Neural programmer: Inducing latent programs with gradient descent. *arXiv preprint arXiv:1511.04834*, 2015.

Negri, S. and Von Plato, J. *Structural proof theory*. Cambridge university press, 2008.

Oh, J., Farquhar, G., Kemaev, I., Calian, D. A., Hessel, M., Zintgraf, L., Singh, S., Van Hasselt, H., and Silver, D. Discovering state-of-the-art reinforcement learning algorithms. *Nature*, pp. 1–2, 2025.

OpenAI. How people are using chatgpt, 2025. URL https://openai.com/index/how-people-are-using-chatgpt/. Published on September 15, 2025; accessed on January 6, 2026.

Paulson, T. N. L. C. and Wenzel, M. A proof assistant for higher-order logic, 2013.

Pearl, J. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 4(5-6):363–389, 1990.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Pinar Saygin, A., Cicekli, I., and Akman, V. Turing test: 50 years later. *Minds and Machines*, 10(4):463–518, 2000.

Placani, A. Anthropomorphism in ai: hype and fallacy. *AI and Ethics*, 4(3):691–698, 2024.

Portoraro, F. Automated Reasoning. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2025 edition, 2025.

Richardson, H. S. Moral Reasoning. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2018 edition, 2018.

Richens, J. and Everitt, T. Robust agents learn causal world models. *International Conference on Learning Representations*, 2024.

Richens, J., Everitt, T., and Abel, D. General agents need world models. In *Forty-second International Conference on Machine Learning*, 2025.

Rossi, F., Bessiere, C., Biswas, J., Conitzer, R. B. V., Dietterich, T. G., Dignum, V., Etzioni, O., Forbus, K. D., Freuder, E., Gil, Y., et al. Aaai 2025 presidential panel on the future of ai research. *Association for the Advancement of Artificial Intelligence, Washington, DC*, 2025a.

Rossi, F., Hoos, H., and Kambhampati, S. Aaai presidential panel on ai reasoning. Association for the Advancement of Artificial Intelligence, 2025b. URL https://www.youtube.com/watch?v=yXU7lABVIWE.

Sakai, Y., Kamigaito, H., and Watanabe, T. Hallucitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in acl conferences, 2026. URL https://arxiv.org/abs/2601.18724.

Scheffer, M., van de Leemput, I., Weinans, E., and Bollen, J. The rise and fall of rationality in language. *Proceedings of the National Academy of Sciences*, 118(51):e2107848118, 2021.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Shmatko, N., Adam, A., and Esau, P. Gptzero finds 100 new hallucinations in neurips 2025 accepted papers, 2025. URL https://gptzero.me/news/neurips/.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., and Farajtabar, M. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

Simon, H. A. Search and reasoning in problem solving. *Artif. Intell.;(Netherlands)*, 1, 1983.

Simon, H. A. Bounded rationality in social science: Today and tomorrow. *Mind & Society*, 1(1):25–39, 2000.

Sjøberg, D. I. and Bergersen, G. R. Construct validity in software engineering. *IEEE Transactions on Software Engineering*, 49(3):1374–1396, 2022.

Sloman, S. A. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3, 1996.

Smith, R. Aristotle's Logic. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations (ICLR)*, 2025.

Snoswell, A. J., Kilov, D., and Lazar, S. Beyond verdicts: Evaluating language model moral competence. *AAAI Conference on Artificial Intelligence*, 2026.

Stechly, K., Valmeekam, K., Gundawar, A., Palod, V., and Kambhampati, S. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. *arXiv preprint arXiv:2505.13775*, 2025.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., and Smyth, P. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Strasser, C. and Antonelli, G. A. Non-monotonic logic. 2001.

Sutton, R. The bitter lesson. 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*, volume 1. MIT press Cambridge, 1998.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36: 74952–74965, 2023.

Van Ditmarsch, H., van Der Hoek, W., and Kooi, B. *Dynamic epistemic logic*. Springer, 2008.

Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., et al. Position: Evaluating generative ai systems is a social science measurement challenge. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Wang, G., Li, J., Sun, Y., Chen, X., Liu, C., Wu, Y., Lu, M., Song, S., and Yadkori, Y. A. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

Wang, X. and Zhou, D. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Bommasani, R., Ganguli, D., Koyejo, S., and Isaac, W. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*, 2025.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S., et al. Livebench: A challenging, contamination-limited llm benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wilholt, T. Epistemic trust in science. *The British Journal for the Philosophy of Science*, 2013.

Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, J. X., Kan, M.-Y., He, J., and Xie, M. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 2023.

Xin, H., Guo, D., Shao, Z., Ren, Z., Zhu, Q., Liu, B., Ruan, C., Li, W., and Liang, X. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

Xu, X., Lawrence, R., Dubey, K., Pandey, A., Ueno, R., Falck, F., Nori, A. V., Sharma, R., Sharma, A., and Gonzalez, J. Re-imagine: Symbolic benchmark synthesis for reasoning evaluation. In *International Conference on Machine Learning*, 2025.

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Yao, M., Wei, Y., and Wang, H. Promoting research by reducing uncertainty in academic writing: a large-scale diachronic case study on hedging in science research articles across 25 years. *Scientometrics*, 128(8):4541–4558, 2023a.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 2023b.

Ye, G., Pham, K. D., Zhang, X., Gopi, S., Peng, B., Li, B., Kulkarni, J., and Inan, H. A. On the emergence of thinking in llms i: Searching for the right intuition. *arXiv preprint arXiv:2502.06773*, 2025.

Zhang, Y., Kuzborskij, I., Lee, J. D., Leng, C., and Liu, F. Dag-math: Graph-guided mathematical reasoning in llms, 2025. URL https://arxiv.org/abs/2510.19842.

Zhu, H., Hao, S., Hu, Z., Jiao, J., Russell, S., and Tian, Y. Reasoning by superposition: A theoretical perspective on chain of continuous thought. *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Appendix

A. Checklist: Community Guidelines for Scientific Communication in AI Reasoning Research

REASONING RESEARCH CHECKLIST

1. ***Definition: Reasoning, Reasoners & Their Components***

- ☐ **1.1** Reasoning is framed as a process, distinct from any artifact produced by that process.
- ☐ **1.2** A formal, operational, and domain-specific definition of reasoning is provided.
- ☐ **1.3** Each essential component in Defs. 2.1 and 2.4 is explicitly defined for the problem setting, where applicable: process, rules, beliefs, evidence, and state. Absence of a component or presence of alternative components is explicitly justified.
- ☐ **1.4** Sources of extrinsic evidence are reported.
- ☐ **1.5** Research clearly defines the state, if and how it is recorded in memory, and how it is retrieved.
- ☐ **1.6** Research clearly reports how reasoning steps are selected, searched for, trialed, etc. If rules are selected by search, the search space and search procedure are defined.
- ☐ **1.7** Implementation details for all mechanisms of exact rule application are provided.
- ☐ **1.8** Can the system be formally characterized as a goal-directed decision-maker that implements a reasoning procedure (a *reasoner*, Def. 2.2), or is the system limited to the reasoning procedure itself?
- ☐ **1.9** When a distinct *reasoner* entity is present, its components are clearly described and its goal is operationally defined.

---

2. ***Reasoning Process Validity***

- ☐ **2.1** Validity is defined w.r.t exact rule application, per Def. 2.7. Alternative definitions of validity are rigorously justified.
- ☐ **2.2** Conditions for valid transitions $\mathcal{S}_t \to \mathcal{S}_{t+1}$ and exact versus approximate execution are stated.
- ☐ **2.3** Is each new belief *provably* obtained by exact rule application, or by some other mechanism? In the absence of proof, hypotheses should be provided.
- ☐ **2.4** Research clearly reports the provenance of rules and rule updates.
  - Are rules learned, or axiomatic?
  - Are rules defined in collaboration with domain experts?
  - Are rules continuously updatable? When meta rules exist, how exactly are rule updates obtained?
- ☐ **2.5** Potential sources of error are explained, along with means for identifying and preventing invalid reasoning steps.
- ☐ **2.6** All theoretical guarantees on validity are formally proven, including formal bounds on performance. Absence of guarantees is clearly stated and justified, and supported by rigorous empirics.

---

3. ***Evaluation & Construct Validity***

- ☐ **3.1** The construct validity of all evaluation methods is explicitly justified w.r.t. the operational definitions under use.
  - If evaluation relies on "reasoning tasks," what exactly constitutes a task? How does it capture reasoning behaviors?
  - Is soundness w.r.t. some external ground truth or preference relevant in this setting? How is it measured?
  - Are the validity, soundness, etc., of intermediate reasoning steps verified, and if so, how?
- ☐ **3.2** Evaluations must clearly address the distinction between the reasoning *process* (relative to internal generating mechanisms) versus the *artifacts* of that process (e.g., QA outputs). Tasks and metrics must measure **both** process quality and output quality.

---

4. ***Utility, Explainability & Trustworthiness***

- ☐ **4.1** Uses and limitations of the system are clearly defined.
- ☐ **4.2** Potential harms from use or misuse of the system are addressed.
- ☐ **4.3** All sources of explainability are described, and any absence of explainability is directly justified.
  - What role do reasoning traces play, what form do they take, and how observable are they to the user?
  - Is the reasoning trace theoretically guaranteed to accurately reflect the model's internal process? If not, what reasonable expectations of faithfulness are possible?
- ☐ **4.4** The operational definition of reasoning met by the system matches the intended use case. If it falls short, the alignment discrepancy is clearly and thoroughly communicated.
- ☐ **4.5** The system implements a useful, nontrivial reasoning process beyond standard ML inference, where usefulness is contextually defined w.r.t. the deployment setting.
- ☐ **4.6** Reported findings refrain from excessive or misleading claims, especially in title, abstract, public reporting to lay audiences, and marketing.

B. Domain-Specific Reasoning, Continued

Example B.1 (Nonmonotonic reasoning)**.** In this example, we consider *nonmonotonic* (or *defeasible*) logic, which adds a mechanism for principled belief retraction and revision to classical logic (Strasser & Antonelli, 2001).

Nonmonotonic reasoning describes a process of deduction and revision which admits both *strict* (static) and *defeasible* (modifiable) rules, including rules governing belief retraction, priority, and conflict-resolution. New, non-defeasible information available at time t, as well as observed contradictions within the current belief state may trigger retractions or updates (via rule application) to prior beliefs and existing rules. This results in a nonmononically updating belief state, in which a conclusion φ derived at time t may fail to hold at time $t' > t$.

Next, we consider a trivial case of reasoning where rules are hard-coded and evidence is the empty set.

Example B.2 (Hard-coded algorithm)**.** A hard-coded algorithm, as represented by a finite, deterministic Turing machine $\mathcal{M}$, can be considered as a reasoning process with a static rule set, such that exactly one applicable local rule (and no meta rules) exist for any given state. At every time step t following an initial instantiation, $\mathcal{M}$ executes a fixed procedure: read the tape symbol under the head, consult a transition table to determine the single applicable rule given the current state, then write a symbol, move left or right, and change state or halt.

As in deductive reasoning, new evidence is not provided during the reasoning process. Furthermore, all rule selectors are trivial, as only a single state transition is valid at any time step, and the conclusion is always the belief state if and when $\mathcal{M}$ halts.

Example B.2 illustrates that the ability of a system to map to the formal definition of reasoning is not necessarily meaningful in itself. *Useful* reasoning will usually require some kind of *alignment* with user preferences, resource constraints, requirements on soundness, or other details of the unique problem setting. For example, a hard-coded algorithm that responds to every input query with the answer "4" vacuously meets the standard of rigorous rule application, but fails to align with a domain-specific setting where soundness requires accurate answers to arithmetic queries.

Example B.3 (Probabilistic next token prediction)**.** This example presents a form of autoregressive probabilistic reasoning over natural language. Consider the n-gram language model that maximizes the probability $p(w \mid h)$ of token w given the history h of tokens preceding w (Jurafsky & Martin, 2025). Rule set $\mathcal{R}_t$ encodes assumptions over the number of relevant preceding tokens in h, along with formulae for valid estimation. For example, we can define $\mathcal{R}_t$ as the set containing

$$p(w_{1:n}) = \prod_{t=1}^{n} p(w_t \mid w_{1:t-1}) \tag{4}$$

$$p(w_t \mid w_{1:t-1}) \approx p(w_t \mid w_{t-1}) \tag{5}$$

$$p(w_t \mid w_{t-1}) = \frac{\mathcal{C}(w_{t-1}w_t)}{\sum_{w'} \mathcal{C}(w_{t-1}w')} \approx \frac{\mathcal{C}(w_{t-1}, w_t)}{\mathcal{C}(w_{t-1})} \tag{6}$$

$$\widehat{w}_t = \arg\max_{w_t} p(w_t \mid w_{t-1}) \tag{7}$$

where Equation 4 is the chain rule of probability, Equation 5 is the Markov assumption, Equation 6 is maximum likelihood estimation and its simplification per the Markov assumption, and Equation 7 predicts the most likely next token. Thus, we can compute the maximum likelihood for $p(w \mid h)$ by taking the count $\mathcal{C}$ of n-grams beginning with h and terminating with w in the training corpus, normalized by the sum of counts for any n-gram beginning with h. Iterating the prediction procedure (Equation 7), we can extend the length of the output text one token at a time. Current belief at step t is $\widehat{w}_t$ and prior beliefs (intermediate conclusions) are $\widehat{w}_{t-1}$, as these are generated by the predictor. The final string $\widehat{w}_{1:n}$ can be framed as the terminal conclusion. The initial token(s) (or *context*, as in LLMs) can be framed as evidence, as these are extrinsically provided to the predictor.

Reasoning validity in Example B.3 arises from the exact application of $\mathcal{R}_t$, which says nothing of soundness (e.g., the factuality of $\widehat{w}_{1:n}$). It is clear that $\mathcal{R}_t$ is agnostic to factuality, as *truth* is not necessarily high probability (e.g., some factual statements describe extremely rare events, such that their constituent tokens are unlikely to coincide frequently in a text corpus). Even if the final string contains misinformation (as often occurs with hallucinations in LLMs, a more complex instantiation of next token prediction), the probabilistic reasoning expressed in Example B.3 would be valid under Def. 2.7. The important question is whether this validity rule is *sound* for the desired application. If soundness through factuality were necessary for the end user, additional constraints would need to be encoded in $\mathcal{R}_t$.

Table B.1 provides a summary of all domain-specific examples presented in this paper.

C. Alternative Definitions for AI Reasoning

Given the expansive range of phenomena that could satisfy our working definitions, what phenomena *do not* satisfy Defs. 2.1 and 2.4?

We review popular alternative viewpoints on what constitutes reasoning here. We discuss whether these alternative

Example	Beliefs $\mathcal{B}_t$	Evidence $\mathcal{E}_t$	Rules $\mathcal{R}_t$	Goal / Conclusion
Logical deduction	Derived formulas in the proof state Γ.	Premises E_0 are given at $t = 0$ and not updated thereafter.	Fixed inference rules (e.g., modus ponens, introduction/elimination rules)	Derive a target formula φ such that $\Gamma \vdash \varphi$.
Bayesian inference	Current posterior $p(\theta \mid D_{1:t})$	Newly observed data D_t (possibly aggregated with prior observations).	Bayes' rule and any auxiliary update rules (e.g., conjugate prior updates, approximation schemes).	Obtain updated posterior beliefs $p(\theta \mid D_{1:t})$.
Reinforcement learning	Current value function estimates, policy parameters, and internal state representations.	Observed environment states, state transitions, and rewards obtained by interaction with the environment.	Update rules such as temporal-difference or policy-gradient updates, plus any meta-rules adapting learning rates or architectures.	Learn a policy that maximizes expected return (i.e., select approximately optimal actions over time).
Nonmonotonic logic	Current set of accepted conclusions, including defeasible ones.	New information that may conflict with existing conclusions (e.g., exceptions, defaults).	Nonmonotonic inference rules that support belief revision and retraction, plus meta-rules for revising the rule set itself.	Maintain a coherent, defeasible belief set that updates appropriately under new, possibly contradictory evidence.
Turing machine	Current configuration of the machine: tape contents $\sigma_t \in \Sigma^*$, head position h_t, and control state $q_t \in Q$, collectively encoded as $\mathcal{B}_t = \langle \sigma_t, h_t, q_t \rangle$.	Input word $w \in \Sigma^*$ (typically fixed at $t = 0$).	Transition relation or function $\delta_t : Q \times \Sigma \to Q \times \Sigma \times \{L, R\}$.	Compute the value of a (partial) function $f : \Sigma^* \rightharpoonup \Sigma^*$ on input w, i.e., reach a halting configuration with output tape σ_T such that σ_T encodes $f(w)$.
Probabilistic next-token prediction	Current token given prior tokens.	Token(s) provided at initialization (e.g., the first word of a sentence, the prompt to an instruction-tuned LLM, etc.).	Probability rules (e.g., chain rule), structural assumptions (e.g., Markovianity), maximum likelihood formulae.	Iterate procedure until query is answered (e.g., string is of desired length, etc.).

Table B.1. Example instantiations of $\mathcal{S}_t = \langle \mathcal{B}_t, \mathcal{E}_t, \mathcal{R}_t \rangle$ for the domain-specific reasoning examples in §2.2 and Appendix B.

definitions are operational and whether they satisfy Defs. 2.1 and 2.4. Per P1, it is not unusual for papers on reasoning to avoid defining reasoning at all. Thus, some of the alternative definitions discussed here are those that we deem to be *implied* by a subset of the literature, if not explicitly stated. While some alternative definitions provided here partially overlap with Defs. 2.1 and 2.4 and may provide research value in some settings, none feature every core component of our operational definitions (per colored highlighting).

We begin with the dictionary. Testament to the hardness of defining latent constructs like reasoning, even dictionaries can be ambiguous. Consider the self-referential definitions found in Merriam Webster (the oldest and most authoritative American English dictionary), which also conflate reason and another latent construct: intelligence.

Alternative Definition C.1 (Reasoning, Merriam-Webster 2026)**.** *Reasoning, noun.* The use of reason; the drawing of inferences or conclusions through the use of reason.
Reason, verb. To use the faculty of reason so as to arrive at conclusions; to discover, formulate, or conclude by the use of reason; to persuade or influence by the use of reason.
Reason, noun. The power of comprehending, inferring, or thinking especially in orderly rational ways; intelligence; proper exercise of the mind; the sum of intellectual powers.

This definition is not operational in multiple senses: how would one measure the "power of comprehending," the "sum of intellectual powers," or "orderly rational ways"? While this definition frames reasoning as a form of inference (which we do not disagree with), it does not address the substrates on which inference is performed (extrinsic evidence, prior beliefs, etc.) nor any concrete mechanisms by which inference is executed (exact rule application, etc.).

The Stanford Encyclopedia of Philosophy (SEP) does not provide a single authoritative definition, with definitions varying across articles.

Alternative Definition C.2 (Reasoning, Richardson 2018 in SEP)**.** "Active or explicit thinking, in which the reasoner, responsibly guided by her assessments of her reasons (Kolodny, 2005) and of any applicable requirements of rationality (Broome, 2009; 2013), attempts to reach a well-

supported answer to a well-defined question (Hieronymi, 2013)."

Alternative Definition C.3 (Automated reasoning, Portoraro 2025 in SEP)**.** "Reasoning is the ability to make inferences... [by] proving the conclusion from the given assumptions by the systematic application of rules of deduction embedded within the reasoning program."

Alternative Def. C.2 contains too many ambiguities to be easily operationalized ("responsibly guided by her assessments", "requirements of rationality", "well-supported", etc.). Further, Alternative Def. C.2 invokes Broome's notion of rational *requirement*, which Broome (2013) replaced with rational *permissibility* (a stance that we also take in this position; §2.3). Alternative Def. C.3 is clearer, and contains some ingredients from operational Def. 2.4: conclusions drawn by the systematic application of rules as embedded in the reasoning program implies (1) a sequential process of exact rule application and (2) process validity, correctness-by-permissibility, etc. Assumptions might encompass evidence, prior beliefs, and/or some forms of rules, though this is unclear. Sources of extrinsic evidence are not directly addressed. This definition also departs from ours in casting reasoning as an *ability* rather than a process.

Our informal Def. 2.1 closely resembles the definition proposed by Wang et al. (2025):

Alternative Definition C.4 (Reasoning, Wang et al. 2025)**.** The process of devising and executing complex goal-oriented action sequences.

Like Defs. 2.1 and 2.4, Alternative Def. C.4 frames reasoning as a sequential process. We can map "action sequences" to our concept of rule sequences: both act on evolving streams of intrinsic and/or extrinsic information and result in updates to the state. Defs. 2.1 and 2.4 make this even more explicit: rules *act on* prior beliefs and current evidence, and *output* updated beliefs about the state. We can then map the concept of *devising* action sequences to *selecting* rule sequences. However, Alternative Def. C.4 does not explicitly delineate sources of extrinsic information (evidence), nor define concepts comparable to belief and state. Wang et al. (2025) depart from Defs. 2.1 and 2.4 by placing goal-orientedness *within* the definition of reasoning. We present an alternative view where reasoning itself has no goal, but may be executed by a goal-directed decision-maker (the *reasoner*, Def. 2.2). This distinction might or might not have consequences for research. Additionally, Wang et al. (2025) explicitly invoke complexity (without a concrete threshold for what constitutes *complex*), while our definitions intentionally admit trivial cases.

The following two definitions are also similar to Defs. 2.1 and 2.4, but (1) are not clearly operational and (2) are overly specialized to human cognition (using anthropocentric language like "mental process," "thinking," etc.), which is of unclear value for designing automated systems.

Alternative Definition C.5 (Reasoning, Broome 2013)**.** Reasoning is a mental process in which you operate on the contents of your attitudes, following a rule.

Alternative Definition C.6 (Reasoning, Huang & Chang 2023)**.** Reasoning is the process of thinking about something in a logical and systematic way, using evidence and past experiences to reach a conclusion or make a decision.

The "contents of your attitudes" in Alternative Def. C.5 could encompass intrinsic beliefs and/or extrinsic evidence, but this is unclear. Alternative Def. C.6 notes evidence, but it is unclear how "past experiences" differ from evidence (where the latter can, in our conceptualization, be derived from interactions with the environment – i.e., *experiences*). Unlike Alternative Def. C.5, Alternative Def. C.6 does not explicitly invoke rules (though logical rules may be ambiguously implied by "a logical and systematic way"). We view these definitions as non-operational, as they leave many questions open: What qualifies as "logical" and "systematic"? Which logic system is being used? And what does it mean to "operate" on the "contents of your attitudes"? Our formal definition makes these notions more concrete.

Note that no components of Defs. 2.1 and 2.4 are explicitly present in the remaining alternative definitions discussed below (per colored highlighting).

Alternative Definition C.7. Reasoning is guided search.

Contemporary LRMs frequently employ search heuristics that enable exploration or deliberation over the solution space, often with self-evaluation (Yao et al., 2023b; Xie et al., 2023; Grand et al., 2025). We observe that the performance gains conferred by these heuristics may contribute toward the conflation of search and reasoning itself. Indeed, the relationship between search and reasoning is significant. Relatedly, a recent AAAI survey found that 44.7% of respondents agreed that "reasoning involves a search process" (Rossi et al., 2025a). While we agree that search can be an effective means of facilitating reasoning, it is not necessary for reasoning and is not reasoning in and of itself. To clarify, we quote Simon (1983):

> The same problem-solving algorithm can be viewed, now as search, now as reasoning [...] Consider, for example, a simple theorem-proving program that works forward from a set of axioms, applying its rules of inference to these to obtain new expressions that can be added to the axiom set. When it finishes tracing a path to a desired theorem, it has succeeded. Clearly it is a search algorithm. At the same time, the theorem prover is adding, at each step of its search, new propositions that follow logically from its axioms. It

> is gradually accumulating a larger and larger collection of deduced propositions. Clearly it is reasoning. [...] The search and constraint metaphors focus upon the process of finding the problem solution, while the reasoning metaphor focuses upon the logical validity of the linkage between initial problem state and solution. Search is centrally concerned with discovery, reasoning with proof (Simon, 1983).

Many principled search procedures can be framed as special cases of Def. 2.4, and discovery-via-search can be an effective strategy or subroutine for implementing proof-by-reasoning in some settings. However, many counterexamples exist where reasoning does not entail any search process (e.g., Example B.2). Thus, we conclude that (1) search is not necessary for reasoning; (2) the definition of search does not equate to a general operational definition for reasoning, as accomplished with Def. 2.4; and (3) we caution against conflating the two in the general case.

Alternative Definition C.8. Reasoning is chain-of-thought.

Following from Claim 2.7, we contend that CoT is a mode of expression and a strategy for eliciting higher quality outputs from LLMs (Wei et al., 2022b), but is not reasoning in itself. Consider the many counterexamples of reasoning that do not require natural language CoT, neither for execution nor for the communication of results (e.g., Example 2.1 and the other domain-specific examples presented in this work). This is echoed by evidence from neuroscience that many forms of human reasoning do not require natural language (Claim 2.7). Nevertheless, the phrase "CoT reasoning" remains popular (Wang et al. 2023; Ling et al. 2023; Wang & Zhou 2024; *inter alia*), fueling the conflation of this useful mode of expression and the reasoning process itself.

Serial CoT text naturally lends itself to human-interpretable explanations of multi-step processes. However, there is no innate requirement that CoT verbalizes the exact application of rules, nor that prior beliefs and current evidence can be incorporated in a principled fashion. Further, it is well-established that LLM CoT provides no guarantees that the latent reasoning process of the AI is faithfully conveyed (Turpin et al., 2023; Lyu et al., 2023; Li et al., 2025; Manuvinakurike et al., 2025; Barez et al., 2025; Stechly et al., 2025; Kambhampati et al., 2025). Thus, we contend that r-zombies could feasibly produce persuasive yet unsound or adversarial CoT with no faithful relation to internal data generating mechanisms.

As an example, consider *stream of consciousness* (SoC) as a form of CoT that does not represent reasoning (James, 1890; Humphrey, 1954). SoC is a continuous flow of serial, unfiltered, and unstructured mental states, similar to concepts like wandering thoughts or daydreams. While these inner monologue-esque narrative streams can be expressed as CoT, there is no requirement that SoC represents any logical progression, and indeed it may be completely irrational (Humphrey, 1954). We observe that unfaithful CoT, as can be produced by current LLMs, at time resembles SoC more than reasoning. The researcher bears the onus of proving that LLM CoT is more than SoC.

Alternative Definition C.9. Reasoning is test-time scaling.

Scaling test-time compute (Snell et al., 2025) is a dominant strategy for improving reasoning benchmark performance (Bi et al., 2024; Chollet et al., 2024; Muennighoff et al., 2025). In this vein, Ye et al. (2025) take thinking and reasoning synonymously, defining these as "the ability to take more time and compute during inference with the goal of producing a higher quality output to a given input." This evokes System 2 thinking: the slower, more deliberative, intentional, and logical mode of reflection modeled by Kahneman (2011). Like search and CoT, test-time scaling is a *means of facilitating reasoning* that is nevertheless not necessary for reasoning. See Def. 2.4 (which says nothing of the scale of computational resources and admits trivial implementations) and Example B.2 as a counterexample. Additionally, we can imagine an r-zombie that adversarially extends its processing time to emulate deliberation or System 2 thinking, without actually engaging in the rule-based mechanisms of valid reasoning. Thus, test-time scaling is not reasoning in itself.

The following two definitions share a common shortcoming.

Alternative Definition C.10. Reasoning is correct output.

Alternative Definition C.11. Reasoning is strong performance on *reasoning tasks* (Def. E.1): benchmark tasks that would require a human test-taker to perform reasoning.

Generative AI papers that target "strong reasoning performance" often do not define reasoning (Muennighoff et al., 2025), inadvertently contributing to the conflation of task accuracy and the reasoning process itself. Our main disagreements with Alternative Defs. C.10 and C.11 are described in §1.1. In short, the output-based view (conflating *process* and *product*) and success on benchmarks is *not sufficient* for proving that a system can reason. Relying on empirical task evaluation alone is especially fraught when the form of reasoning under study does not feature known or unique ground truth outputs, as in moral reasoning (Snoswell et al., 2026) or exploratory problem settings (e.g., scientific discovery). We direct the reader to Bowman & Dahl (2021); Cheng et al. (2025); Alaa et al. (2025); Weidinger et al. (2025); Wallach et al. (2025); Mitchell (2025b) for further reference on the problems associated with benchmarking.

D. Extended Discussions

D.1. Epistemic Trust in Generative AI & AI Reasoning

In psychology, trust can be framed as a mechanism for mitigating uncertainty, reducing resource costs when engaging with external entities, and increasing the probability of successful outcomes (Lukyanenko et al., 2022). Science is fundamentally a "collective epistemic enterprise," and as such *epistemic trust* (Def. E.5) underpins scientific integrity through two main social contracts: (1) successful collaboration requires that scientists trust the information provided by each other, and (2) societal investment requires that the lay public trusts the information provided by scientists.

Currently, epistemic trust in AI faces challenges both within the scientific community and with respect to public perception. Reports of public trust vary heavily: 39% of American respondents predicted that AI will be more beneficial than harmful, versus 83% of Chinese respondents (Maslej et al., 2025); 30% of Swiss respondents believed AI to be completely unacceptable (up from 23%), while 26% supported human-only decision-making (up from 18%) (Baumann et al., 2025); only 14% of UK respondents predicted that AI will have a positive impact on society, with negative perception increasing (CDEI, 2023). At the same time, pervasive mistrust coincides with conflicting phenomena: escalating capital investment and surging user uptake. OpenAI reports 700 million weekly active users for ChatGPT alone (OpenAI, 2025), while 88% of survey respondents regularly used AI in at least one business function (McKinsey, 2025). Despite widespread consumer usage, payoffs have not yet fully materialized. With an estimated $\$30-40$ billion in enterprise investment in generative AI, 95% of surveyed organizations have seen zero return on investment (ROI) (Challapally et al., 2025). Even with low returns and evidence of untrustworthiness (e.g., hallucinations), phenomena such as anthropomorphism and sycophancy often exaggerate model confidence, encouraging users to overestimate LLM accuracy (Steyvers et al., 2025).

These contradictory phenomena – high-variance public trust rates coupled with increasing user uptake, escalating financial investment coupled with low ROI, and propensities toward exaggerated confidence coupled with evidence of non-trivial error rates – point to an urgent need to resolve questions of trustworthiness in AI.

D.2. Historical Perspectives on Reasoning

Philosophy, Cognitive Science & the Social Sciences

The study of reasoning spans millenia of qualitative and quantitative inquiry. We provide a brief and nonexhaustive summary of historical contributions in the humanities, social sciences, and studies of cognition and the brain.

The history of reasoning is, in many ways, the history of logic and epistemology. Major contributions in ancient logic emanated from early Greek, Indian, Chinese, and Arab cultures, among others. The ancient Greek polymath Aristotle (384–322 BC) provided an early systematic study of logic, establishing deductive reasoning via syllogisms (Smith, 2022). Aristotle categorized reasoning into *prior analytics* (formal structural argumentation via syllogisms, analogous to notions of validity discussed in this position) and *posterior analytics* (focused on demonstration, definition, scientific knowledge, and inductive reasoning, where premises must be true, primary, immediate, and necessary; this notion maps roughly to soundness and operationalization). We refer the reader to Bobzien (2020) for further discussion of ancient traditions of the West.

In India, orthodox schools of Hindu philosophy such as Nyāya developed rigorous theories of logic and epistemology (Britannica, 2017). Various schemas of inference were proposed for the evaluation of knowledge and arguments (e.g., *premise, reason, example, application, and conclusion*). In the thirteenth century, the Navya-Nyāya or Neo-Logical school of Indian philosophy further systematized these logical systems, anticipating aspects of modern set theory and influencing later logicians such as Babbage, Boole, and DeMorgan. See Gillon (2024) for further discussion of logic in classical Indian philosophy.

The classic epistemological debate over *rationalism versus empiricism* concerns the sources by which we obtain knowledge about our external world (Markie & Folescu, 2023). While the rationalists emphasized deduction and mathematical certainty (as represented by French polymath René Descartes (1596–1650), German polymath Gottfried Wilhelm Leibniz (1646–1716), et al.), the empiricists emphasized sensory experience, causation, and probability (as represented by the English philosopher John Locke (1632–1704), Scottish philosopher David Hume (1711–1776), et al.). German philosopher Immanuel Kant (1724–1804) presented a critique of pure reason that attempted to bridge rationalism and empiricism. Modern formal logic (as represented by Gottlob Frege (1848–1925), Bertrand Russell (1872–1970), et al.) overhauled Aristotelian logic into symbolic mathematical logic. For more recent treatments of reasoning vis-à-vis logic and epistemology in the computer science community (with emphases on probabilistic reasoning and uncertainty), we refer the reader to Fagin & Halpern (1987); Pearl (1990); Fagin & Halpern (1994); Fagin et al. (2004); Pearl (2014); Halpern (2017).

Cognitive science, neuroscience, and psychology have contributed a brain- or mind-centric account of reasoning. Dual-process theories of reasoning have been explored (and challenged) for centuries (Evans & Stanovich, 2013), perhaps most famously with Daniel Kahneman's theory of *System*

1 and System 2 thinking in psychology and behavioral economics (Sloman, 1996; Kahneman, 2011). While System 1 is associated with fast, automatic, frequent, and intuitive forms of cognition (e.g., performing basic arithmetic, catching a ball), System 2 entails slow, deliberative, effortful, and logical cognition (e.g., proving a theorem). System 2 thinking is sometimes referenced as a metaphor for inference-time scaling in generative AI. Herbert Simon's theories on *bounded rationality*, reasoning, and decision-making under uncertainty had a significant impact on computer science, economics, and cognitive psychology. As in this position, Simon (2000) argues that interrogating the nature and quality of the *process* of reasoning, and not only its products, clarifies a reasoner's limitations (original emphasis):

> A theory of bounded rationality, then, will be as much concerned with procedural rationality, the quality of the processes of decision, as with substantive rationality, the quality of the outcome. To understand the former, one must have a theory of the psychology of the decision maker; to understand the latter, one needs have only a theory of the goal (the utility function) and the external environment. [...] When rationality is associated with reasoning *processes*, and not just with its *products*, limits on the abilities of Homo sapiens [sic] to reason cannot be ignored. So the reasoning we find in the classics sounds very different from the calculus of maximization of expected utility in modern neoclassical economics. Taking account of process as well as product is compatible, as neoclassical thinking is not, with the idea that, while human beings usually have reasons for what they do, these are seldom the best reasons, and are seldom consistent over the whole range of their choices.

Automated Reasoning Across "Three Waves" of AI

Foundational work on automated reasoning included production systems (Davis & King, 1984; Hayes-Roth, 1985), logic programming (Lloyd, 2012), belief revision (Gärdenfors, 1988; Van Ditmarsch et al., 2008) and early proof assistants (Boyer & Moore, 1975; de Bruijn, 1983; Gordon, 1985; Coquand & Huet, 1986), as well as theoretical work on the typed lambda calculus and structural proof theory (Howard et al., 1980; Negri & Von Plato, 2008). The symbolic, rule-based perspective of these approaches dominated early AI research but fell out of favor by the late 1980s, following the collapse of the specialized AI hardware market, unresolved scalability issues in expert systems, and DARPA funding cuts (Fouse et al., 2020). This period is popularly considered the end of the "First Wave of AI" and the beginning of an "AI Winter" of reduced global funding and interest in AI.

In contrast, statistical learning and neural networks drove the fast-paced "Second Wave of AI," as the dominance of deep learning overshadowed rule-based AI through the 2010s (Fouse et al., 2020). Prototypical AI systems from this wave prioritized data-driven approaches, viewed models primarily as black boxes, and provided limited explicit reasoning and transparency. Sutton's "Bitter Lesson" (Sutton, 2019) was particularly influential in expressing disillusionment with domain-specific understanding in AI, as contrasted with the superior performance of systems relying primarily on scaling laws of increasing compute and training data.

In recent years, a broad push toward the goal of *language reasoning models*, as well as renewed interest in formal methods and neuro-symbolic architectures, have challenged the perspective that symbolic AI is of mere historical interest (Huang & Chang, 2023; Belle & Marcus, 2025). Interactive theorem provers such as Lean and Isabelle/HOL (Paulson & Wenzel, 2013; De Moura et al., 2015; Blanchette et al., 2016) have demonstrated substantial progress toward scalable mathematical formalization and verification. In parallel, the rapid rise of neuro-symbolic architectures in an emerging "Third Wave of AI" (Garcez & Lamb, 2023) has enabled capabilities such as latent program induction (Neelakantan et al., 2015; Macfarlane & Bonnet, 2025) and theorem-proving systems that tightly integrate symbolic solvers with neural components (Xin et al., 2024; Chervonyi et al., 2025). Further advances in large-scale ML, such as retrieval-augmented generation, model-based planning, and world modeling, have strengthened the case for revisiting classical ideas under modern computational regimes (Matsuo et al., 2022; Gao et al., 2023; Guan et al., 2023).

This shift has been reinforced by growing awareness of the intrinsic limitations of current LLMs (see §1.1), including hallucination (Xu et al., 2024; Bastounis et al., 2024), reliance on heuristics or non-generalizing "shortcut solutions" (Liu et al., 2022; Chollet et al., 2024; Xu et al., 2025; Mirzadeh et al., 2025), and formal complexity-theoretic boundaries (Merrill et al., 2022; Merrill & Sabharwal, 2023). We echo Belle & Marcus (2025) in hypothesizing that these trends may collectively signal a timely re-evaluation of rule-based AI, not as an abandoned "First Wave" idea, but as a potential component in next-generation architectures and in the pursuit of more reliable, transparent, and generalizable reasoning systems.

E. Glossary

Definition E.1 (Reasoning task)**.** In the AI evaluation setting, we consider a *reasoning task* to be a task that, when previously unseen, would require the average human solver to perform reasoning. Thus, the notion of a reasoning task is tied to expectations on human problem solving and is therefore anthropocentric.

Definition E.2 (Operational definition, American Psychological Association)**.** "A description of something in terms of the operations (procedures, actions, or processes) by which it could be observed and measured. For example, the operational definition of anxiety could be in terms of a test score, withdrawal from a situation, or activation of the sympathetic nervous system. The process of creating an operational definition is known as *operationalization*."

Definition E.3 (Formal verification, De Moura et al. 2015)**.** "Formal verification involves the use of logical and computational methods to establish claims that are expressed in precise mathematical terms. These can include ordinary mathematical theorems, as well as claims that pieces of hardware or software, network protocols, and mechanical and hybrid systems meet their specifications. In practice, there is not a sharp distinction between verifying a piece of mathematics and verifying the correctness of a system: formal verification requires describing hardware and software systems in mathematical terms, at which point establishing claims as to their correctness becomes a form of theorem proving. Conversely, the proof of a mathematical theorem may require a lengthy computation, in which case verifying the truth of the theorem requires verifying that the computation does what it is supposed to do."

Definition E.4 (Intelligence, Chollet 2019)**.** Skill-acquisition efficiency over a range of tasks, controlling for priors, experience, and generalization difficulty.

Definition E.5 (Epistemic trust)**.** Per Wilholt (2013), "To invest epistemic trust in someone is to trust her in her capacity as provider of information." Fonagy & Allison (2014) consider epistemic trust to be "an individual's willingness to consider new knowledge from another person as trustworthy, generalizable, and relevant to the self." Similarly, Irzik & Kurtulmus (2019) argue that "Epistemic trust is about taking someone's testimony that P as a reason to believe that P on the assumption that she is in a position to know whether P and will express her belief truthfully... In the case of scientists, the requirement of good will for epistemic trust amounts to their commitment to the ethical norms of their trade and their sense of obligation to truthfully and accurately share significant knowledge with the public."

Definition E.6 (Construct validity, Sjøberg & Bergersen 2022)**.** A *construct* is a concept that is not directly measurable, but is represented by indicators at the operational level to make it measurable. The validity of a construct (i.e., *construct validity*) is defined by how adequate a concept definition is and how well the indicators represent the concept.