What is AI Reasoning?

Rachel Lawrence‡
Microsoft Research, Cambridge, UK
rachel.lawrence@microsoft.com

Jacqueline Maasch‡
Cornell Tech, New York, NY
maasch@cs.cornell.edu

‡*Equal contribution.*

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. This position contends that definitional ambiguity leaves the construct validity of reasoning evaluation unfalsifiable, undermining quantifiable progress toward trustworthy autonomous reasoning. We also contend that this ambiguity is addressable. Based on a synthesis of the literature, we provide general definitions for AI reasoning as a *learnable rule-based process*. Operational definitions are expressed in natural language, math, and pseudocode. We address rebuttals to our definitions and propose community guidelines for scientific communication in AI reasoning research.

Core Positions

Thesis 1. ***Define, then measure.*** Operational definitions should be stated for the reasoning phenomena under investigation, and the construct validity of reasoning evaluation should be explicitly justified with respect to these definitions.

Thesis 2. ***Reasoning is a learnable rule-based process.*** 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.*** The *validity* of a reasoning process arises from exact rule application, independent of rule selection.

Intuition: Informal Definitions

Definition 1 (Reasoning)**.** 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 (Reasoner)**.** A goal-oriented decision-maker that implements a reasoning process.

Operationalizing Valid & Sound Reasoning

Definition 3 (Reasoning (formal))**.** Let $\mathcal{S}_t := \langle \mathcal{B}_t, \mathcal{E}_t, \mathcal{R}_t \rangle$ denote the reasoner's state at time step t, where $\mathcal{B}_t$ denotes current internal beliefs, $\mathcal{E}_t$ denotes extrinsically obtained evidence aggregated 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 $\mathcal{S}_{t-1}$.

Rules are *learnable* and *revisable* operators whose operands are information: (1) exogenous evidence, (2) endogenous beliefs, and/or (3) other rules. The rule set is partitioned into two subsets with distinct type signatures — local rules $\mathcal{R}^L$, which update beliefs, and meta rules $\mathcal{R}^M$, which update rules:

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

where $\mathcal{R}_t^L \cap \mathcal{R}_t^M = \emptyset$ and $\mathcal{R}_t^L \cup \mathcal{R}_t^M = \mathcal{R}_t$. State updates $\mathcal{S}_{t-1} \to \mathcal{S}_t$ are defined by the receipt of new evidence $\mathcal{E}_t$, if any, followed by a sequence of two rule applications:

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

Definition 4 (Validity)**.** A transition from state $\mathcal{S}_{t-1}$ to $\mathcal{S}_t$ is *valid* if and only if it arises from the application of rules $r \in \mathcal{R}_{t-1}$ to components of state $\mathcal{S}_{t-1}$.
Definition 5 (Soundness)**.** A valid transition from state $\mathcal{S}_{t-1}$ to $\mathcal{S}_t$ 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.

Algorithm 1 Valid reasoning as exact rule application.

Input. Initial rules $\mathcal{R}_0$, beliefs $\mathcal{B}_0$, evidence stream $\{\mathcal{E}_i\}_{i=0}^T$, stopping rule $\mathtt{s_{stop}}$.

$\mathcal{R}, \mathcal{B}, \mathcal{E} \leftarrow \mathcal{R}_0, \mathcal{B}_0, \mathcal{E}_0$
$\mathcal{S} \leftarrow (\mathcal{R}, \mathcal{B}, \mathcal{E})$
$t \leftarrow 0$
while not $\mathtt{s_{stop}}(\mathcal{S})$ **do**
  $\mathcal{E}' \leftarrow \mathcal{E}_{t+1}$
  $r^L \leftarrow \mathtt{s_L}(\mathcal{R}, \mathcal{B}, \mathcal{E}')$ {Select local rule.}
  $\mathcal{B}' \leftarrow r^L(\mathcal{B}, \mathcal{E}')$ {Apply local rule, update beliefs.}
  $r^M \leftarrow \mathtt{s_M}(\mathcal{R}, \mathcal{B}', \mathcal{E}')$ {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 += 1
end while
Return $\mathcal{B}, \mathcal{T}$ =0

Abstract based on the paper *Position: Reasoning is a Learnable Rule-Based Process* (ICML 2026).