Jumping to Conclusions

Loizos Michael
Open University of Cyprus
loizos@ouc.ac.cy

Abstract

Inspired by the profound effortlessness (but also the substantial carelessness) with which humans seem to draw inferences when given even only partial information, we consider a unified formal framework for computational cognition, placing our emphasis on the existence of naturalistic mechanisms for representing, manipulating, and acquiring knowledge. Through formal results and discussion, we suggest that such fast and loose mechanisms could provide a concrete basis for the design of cognitive systems.

1 Introduction

The founding statement of Artificial Intelligence [McCarthy *et al.*, 1955] proceeds on the basis that *"every [...] feature of intelligence can in principle be so precisely described that a machine can be made to simulate it"* and proposes to *"find how to make machines [...] solve kinds of problems now reserved for humans"*. Philosophical considerations aside, the exceptionally powerful machine learning algorithms and the ingeniously crafted reasoning algorithms readily testify that contemporary Artificial Intelligence research has placed more emphasis on the latter front — the effectiveness of algorithms in terms of their behavior — and less emphasis on the former front — their design to simulate features of the human intellect. With Artificial Intelligence research turning sixty years old, this choice on its emphasis has ultimately led to the recently popularized concerns on its future [Open Letter, 2015].

Refocusing on the former front would seem to necessitate the abandonment of rigid and convoluted algorithms, and a shift towards more robust and naturalistic solutions, guided by psychological evidence on human cognition. Further, and contra to the specialization of contemporary Artificial Intelligence research, a holistic view of cognition seems warranted, with perception, reasoning, and learning all being considered in a unified framework that facilitates their close interaction.

We undertake such an investigation herein, with emphasis on the fast nature of drawing inferences [Kahneman, 2011], and the interplay of cognition and perception [Clark, 2013].

2 Perception Semantics

We assume that the environment is at each moment in a ***state***. An agent cannot directly access such a state. Rather, it uses a pre-specified language to assign finite names to ***atoms***, which are used to represent concepts related to the environment. The set of all such atoms is not explicitly provided upfront. Atoms are encountered while the agent perceives its environment, or introduced through the agent's cognitive processing mechanism. At the neural level, each atom might be thought of as a set of neurons assigned to represent a concept [Valiant, 2006].

A ***scene*** s is a mapping from atoms to $\{\mathtt{0}, \mathtt{1}, *\}$. We write $s[\alpha]$ to mean the value associated with atom α, and call atom α ***specified*** in scene s if $s[\alpha] \in \{\mathtt{0}, \mathtt{1}\}$. Scenes s_1, s_2 ***agree*** on atom α if $s_1[\alpha] = s_2[\alpha]$. Scene s_1 is an ***expansion*** of scene s_2 if s_1, s_2 agree on every atom specified in s_2. Scene s_1 is a ***reduction*** of scene s_2 if s_2 is an expansion of s_1. A scene s is the ***greatest common reduct*** of a set S of scenes if s is the only scene among its expansions that is a reduction of each scene in S. A set S of scenes is ***compatible*** if there exists a particular scene that is an expansion of each scene in S.

In simple psychological terms, a scene can be thought of as the contents of an agent's working memory, where the agent's perception of the environment state, and any relevant thereto drawn inferences, are made concrete for further processing. Following psychological evidence, the maximum number, denoted by $\mathtt{w}$, of specified atoms in any scene used by the agent can be assumed to be a small constant [Miller, 1956].

A propositional formula ψ is ***true*** (resp., ***false***) and ***specified*** in s if ψ (resp., $\neg\psi$) is classically entailed by the conjunction of: atoms α such that $s[\alpha] = \mathtt{1}$, and the negation of atoms α such that $s[\alpha] = \mathtt{0}$; otherwise, ψ is ***unspecified*** in s.

When convenient, we represent fully and unambiguously a scene as the set of its true literals (atoms or their negations).

To formalize the agent's interaction with its environment, let $\mathbb{E}$ denote the set of environments of interest, and let a particular ***environment*** $\langle \mathtt{dist}, \mathtt{perc} \rangle \in \mathbb{E}$ determine: a ***probability distribution*** $\mathtt{dist}$ over states, capturing the possibly complex and unknown dynamics with which states are produced; a stochastic ***perception process*** $\mathtt{perc}$ determining for each state a probability distribution over a compatible subset of scenes. The agent has only oracle access to $\langle \mathtt{dist}, \mathtt{perc} \rangle$ such that: in unit time the agent senses its environment and obtains a ***percept*** s, resulting by an unknown state t being first drawn from $\mathtt{dist}$, and scene s then drawn from $\mathtt{perc}(t)$. A percept s represents, thus, what the agent senses from state t.

Example 1. *At a pedestrian crossing, the "Don't Walk" state (t_1) is signaled by a light being red and no audible cue, while*

the "Walk" state (t_2) is signaled by the same light being green along with an audible cue. A person ($\texttt{perc}_1$) hears the audible cue or not irrespectively of whether the signal post happens to be obscured. A color-blind person ($\texttt{perc}_2$) sees both red and green lights as yellow, but still perceives the audible cue, unless the person ($\texttt{perc}_3$) also suffers from hearing loss.

Let $s_1 = \{\neg Cue\}$, $s_2 = \{\neg Cue, Red\}$, $s_3 = \{\neg Cue, Yellow\}$, $s_4 = \{Cue\}$, $s_5 = \{Cue, Green\}$, $s_6 = \{Cue, Yellow\}$, $s_7 = \{Yellow\}$.

Consider a probability distribution $\texttt{dist}$ *assigning probabilities* $0.9, 0.1$ *to states* t_1, t_2*, and three perception processes:*

$\texttt{perc}_1(t_1)$ *assigns probabilities* $0.3, 0.7$ *to scenes* $\{s_1, s_2\}$
$\texttt{perc}_1(t_2)$ *assigns probabilities* $0.6,\ 0.4$ *to scenes* $\{s_4, s_5\}$

$\texttt{perc}_2(t_1)$ *assigns probabilities* $0.3, 0.7$ *to scenes* $\{s_1, s_3\}$
$\texttt{perc}_2(t_2)$ *assigns probabilities* $0.6,\ 0.4$ *to scenes* $\{s_4, s_6\}$

$\texttt{perc}_3(t_1)$ *assigns probabilities* $0.3, 0.7$ *to scenes* $\{\emptyset, s_7\}$
$\texttt{perc}_3(t_2)$ *assigns probabilities* $0.6,\ 0.4$ *to scenes* $\{\emptyset, s_7\}$

Scenes in each set form a compatible subset, capturing the possible ways in which the underlying state can be perceived. Not assigning an *a priori* meaning to states obviates the need to commit to an objective representation of the environment, and accommodates cases where an agent's perception process determines not only what the agent does (or can possibly) perceive, but also its interpretation. For instance, no perception process above specifies the atom *Safe* of whether the agent itself *believes* it is safe to cross. The value of this atom could be inferred internally by the agent's reasoning process after perceiving the other signals, but cannot be meaningfully determined by the environment state. Furthermore, how an agent perceives the two states, or even whether it perceives them as being distinct, depends on the agent's perception abilities.

An agent's key task is to decide how to act optimally (given its perception process) in the current state of the environment. Decision making is often facilitated by having access to more information, and reasoning serves this role (amongst others): *it completes information not explicitly available in a percept*.[1] To do so, it utilizes knowledge that the agent has been given, or has acquired, on certain regularities in the environment.

3 Reasoning Semantics

Since reasoning serves to complete information, one naturally seeks representations and processes that determine efficiently what inferences follow, and allow inferences to follow often.

A ***rule*** is an expression of the form $\varphi \rightsquigarrow \lambda$, where formula φ is the ***body*** of the rule, and literal λ is the ***head*** of the rule, with φ and λ not sharing any atoms, and with φ being read-once (no atom appears more than once). The intuitive reading of a rule is that when the rule's body holds in a scene, an agent has certain *evidence* that the rule's head should also hold.

A collection of rules could happen to simultaneously provide evidence for conflicting conclusions. To resolve such conflicts, we let rules be qualified based on their priorities.

A ***knowledge base*** $\kappa = \langle \varrho, \succ \rangle$ over a set $\mathcal{R}$ of rules comprises a finite collection $\varrho \subseteq \mathcal{R}$ of rules, and an irreflexive antisymmetric ***priority relation*** $\succ$ that is a subset of $\varrho \times \varrho$. Although we may not make this always explicit, the rules in ϱ are named, and the priority relation $\succ$ is defined over their names. In general, then, duplicate rules can coexist in κ under different names, and have different priorities apply on them.

Definition 1 (Exogenous and Endogenous Qualifications)**.** *Rule r_1 is **applicable** on scene s_i if r_1's body is true in s_i. Rule r_1 is **exogenously qualified** on scene s_i by percept s if r_1 is applicable on s_i and its head is false in s. Rules r_1, r_2 are **conflicting** if their heads are the negations of each other. Rule r_1 is **endogenously qualified** on scene s_i by rule r_2 if r_1, r_2 are applicable on s_i and conflicting, and $r_1 \not\succ r_2$.*

Based on qualification, we define the reasoning semantics.

Definition 2 (Step Operator)**.** *The **step operator** for a knowledge base κ and a percept s is a mapping $s_i \overset{\kappa, s}{\dashrightarrow} s_{i+1}$ from a scene s_i to the scene s_{i+1} that is an expansion of s and differs from s only in making true the head of each rule r in κ that: (i) is applicable on s_i, (ii) is not exogenously qualified on s_i by s, and (iii) is not endogenously qualified on s_i by a rule in κ; such a rule r is called **dominant** in the step.*

Intuitively: The truth-values of atoms specified in percept s remain as perceived, since they are not under dispute.[2] The truth-values of other atoms in s_i are updated to incorporate in s_{i+1} the inferences drawn by dominant rules, and also updated to drop any inferences that are no longer supported.[3]

The inferences of a knowledge base on a percept are determined by the set of scenes that one reaches, and from which one cannot escape, by repeatedly applying the step operator.

Definition 3 (Inference Trace and Inference Frontier)**.** *The **inference trace** of a knowledge base κ on a percept s is the infinite sequence $\texttt{trace}(\kappa, s) = s_0, s_1, s_2, \ldots$ of scenes, with $s_0 = s$ and $s_i \overset{\kappa, s}{\dashrightarrow} s_{i+1}$ for each integer $i \geq 0$. The **inference frontier** of a knowledge base κ on a percept s is the subset-minimal set $\texttt{front}(\kappa, s)$ of the scenes that appear in $\texttt{trace}(\kappa, s)$ after removing some finite prefix of $\texttt{trace}(\kappa, s)$.*

Theorem 1 (Properties of the Inference Frontier)**.** *Consider a knowledge base κ, and a percept s. Then, $\texttt{front}(\kappa, s)$ exists, is unique, is non-empty, and includes finitely-many scenes.*

Proof. An immediate consequence of Definition 3. □

Example 2. *Consider a knowledge base κ with the rules r_1 : Penguin $\rightsquigarrow \neg$Flying, r_2 : Bird $\rightsquigarrow$ Flying, r_3 : Penguin $\rightsquigarrow$ Bird, r_4 : Feathers $\rightsquigarrow$ Bird, r_5 : Antarctica $\wedge$ Bird $\wedge$ Funny $\rightsquigarrow$ Penguin, r_6 : Flying $\rightsquigarrow$ Wings, and the priority $r_1 \succ r_2$. For percept $s = \{Antarctica, Funny, Feathers\}$, $\texttt{trace}(\kappa, s) =$*

$\{Antarctica, Funny, Feathers\}$,
$\{Antarctica, Funny, Feathers, Bird\}$,
$\{Antarctica, Funny, Feathers, Bird, Flying, Penguin\}$,
$\{Antarctica, Funny, Feathers, Bird, \neg Flying, Penguin, Wings\}$,
$\{Antarctica, Funny, Feathers, Bird, \neg Flying, Penguin\}$,
$\{Antarctica, Funny, Feathers, Bird, \neg Flying, Penguin\}, \ldots$

[1] Mercier and Sperber [2011] call this process "inferencing", and reserve the term "reasoning" for a process that produces arguments.

[2] Overriding percepts can be accounted for by introducing rules that map each perceived atom to a duplicate version thereof, which is thereafter amenable to endogenous qualification by other rules.

[3] Such updates are accommodated by having scene s_{i+1} be an expansion of the percept s, but not necessarily of the current scene s_i.

and $\texttt{front}(\kappa, s)$ *is the singleton set whose only member is the scene* $\{Antarctica, Funny, Feathers, Bird, \neg Flying, Penguin\}$.

Observe the back and forth while computing $\texttt{trace}(\kappa, s)$. Initially *Bird* is inferred, giving rise to *Flying*, and then to *Wings*. When *Penguin* is later inferred, it leads rule r_1 to oppose the inference *Flying* from rule r_2, and in fact to override and negate it. As a result of this overriding of *Flying*, inference *Wings* is no longer supported through rule r_6, and is also dropped, even though no other rule directly opposes it.

Thus, the inference trace captures the evolving contents of an agent's working memory, while the inference frontier captures the memory's final (possibly fluctuating) contents. Related is a point by Harman [1974], who insists that we should not infer that intermediate steps do not occur simply because we do not notice them, and that our inability to notice them might be due to the sheer speed with which we go through them. Indeed, assuming that rule applicability is checked in parallel (as for neurons in the brain), and recalling that scene capacity is upper-bounded by a small constant $\texttt{w}$ (ensured, for instance, by keeping only the subpart of each scene that is coherent, as determined by an agent's knowledge base [Murphy and Medin, 1985]), one can see this sheer speed of reasoning.

Intuitively, each rule (i.e., its associated neuron) checks, in parallel, to see if it is applicable on the current scene s_i, and if its head is not specified in the input percept s. The bound $\texttt{w}$ on the size of scenes ensures the high efficiency of this check. All rules that pass the check proceed to attempt to write, in parallel, their head in a shared memory location (one for each atom), and the rule with the highest priority succeeds, giving rise to a scene s_{i+1} such that $s_i \xrightarrow{\kappa, s} s_{i+1}$. Going from here to computing the inference frontier requires checking for repeated scenes in the inference trace. If only singleton inference frontiers are of interest (as discussed next), such checking reduces to whether $s_i = s_{i+1}$, which, again, can be done efficiently. The nature of this computation is supported within the Priority CRCW PRAM model [Cormen *et al.*, 2009].

3.1 Entailment of Formulas

In general, the inference frontier may include multiple scenes, and one can define multiple natural notions for entailment.

Definition 4 (Entailment Notions)**.** *A knowledge base* κ *applied on a percept* s ***entails*** *a formula* ψ *if* ψ *is:*
(N_1*) true in a scene in* $\texttt{front}(\kappa, s)$*;*
(N_2*) true in a scene in* $\texttt{front}(\kappa, s)$ *and not false in others;*
(N_3*) true in every scene in* $\texttt{front}(\kappa, s)$*;*
(N_4*) true in the greatest common reduct of* $\texttt{front}(\kappa, s)$.

Going from the first to the last notion, entailment becomes more skeptical. Only the first notion of entailment captures what one would typically call credulous entailment, in that ψ is possible, but $\neg\psi$ might also be possible. The following result clarifies the relationships between these notions.

Theorem 2 (Relationships Between Entailment Notions)**.** *A knowledge base* κ *applied on a percept* s *entails* ψ *under* N_i *if it entails* ψ *under* N_j*, for every pair of entailment notions* N_i, N_j *with* $i < j$*. Furthermore, there exists a particular knowledge base* κ *applied on a particular percept* s *that entails a formula* ψ_i *under* N_i *but it does not entail* ψ_i *under* N_j*, for every pair of entailment notions* N_i, N_j *with* $i < j$.

Proof. The first claim follows easily. For the second claim, consider a knowledge base κ with the rules $r_1 : \top \rightsquigarrow a$, $r_2 : a \rightsquigarrow b$, $r_3 : a \wedge b \rightsquigarrow c$, $r_4 : c \rightsquigarrow \neg a$, $r_5 : c \rightsquigarrow b$, and the priority $r_4 \succ r_1$, and consider a percept $s = \emptyset$. $\texttt{trace}(\kappa, s)$ comprises the repetition of the five scenes $\{a\}$, $\{a, b\}$, $\{a, b, c\}$, $\{\neg a, b, c\}$, $\{\neg a, b\}$, which constitute $\texttt{front}(\kappa, s)$. The claim follows by letting $\psi_1 = a$, $\psi_2 = b$, $\psi_3 = a \vee b$, and observing that the greatest common reduct of $\texttt{front}(\kappa, s)$ is $\emptyset$. □

Note the subtle difference between the entailment notions N_3 and N_4: under N_4 an entailed formula needs to be not only true in every scene in $\texttt{front}(\kappa, s)$, but true *for the same reason*. This excludes reasoning by case analysis, where an inference can follow if it does in each of a set of collectively exhaustive cases. When $\texttt{front}(\kappa, s) = \{\{\alpha\}, \{\beta\}\}$, for instance, the formula $\alpha \vee \beta$ is true in every scene in $\texttt{front}(\kappa, s)$ by case analysis, and is entailed under N_3, but not under N_4.

When the inference frontier comprises only a single scene (which is, therefore, a fixed-point of the step operator), all entailment notions coincide. In the sequel we restrict our focus, and define our entailment notion only under this special case, remaining oblivious as to what entailment means in general.

Definition 5 (Resolute Entailment)**.** *A knowledge base* κ *is* ***resolute*** *on a percept* s *if* $\texttt{front}(\kappa, s)$ *is a singleton set; then, the unique scene in* $\texttt{front}(\kappa, s)$ *is the* ***resolute conclusion*** *of* κ *on* s*. A knowledge base* κ *applied on a percept* s *on which* κ *is resolute* ***entails*** *a formula* ψ*, denoted* $(\kappa, s) \mid\!\!= \psi$*, if* ψ *is true in the resolute conclusion of* κ *on* s.

Although the entailment semantics itself is skeptical in nature, the *mechanism* that computes entailment is distinctively credulous. It jumps to inferences as long as there is sufficient evidence to do so, and no immediate / local reason to qualify them. If reasons emerge later that oppose an inference drawn earlier, those are considered as they become available.

This fast and loose mechanism follows Bach [1984], who argues for approaching default reasoning as *"inference to the first unchallenged alternative"*. It is also reminiscent of the spreading-activation theory [Collins and Loftus, 1975], which can inform further extensions to make the framework even more psychologically-valid (e.g., reducing the inference trace length by including a decreasing gradient in rule activations).

3.2 Why Not Equivalences?

Are prioritized implications no more than syntactic sugar to conceal the fact that one is simply expressing a single equivalence / definition for each atom? We dismiss this possibility.

Consider a knowledge base κ. Let $\texttt{body}(r_0)$ and $\texttt{head}(r_0)$ mean, respectively, the body and head of rule r_0 in κ. Let $\texttt{str}(r_0)$ mean the set of rules r_i in κ such that r_0, r_i are conflicting, and $r_0 \not\succ r_i$; i.e., the rules that are stronger (or, more precisely, not less preferred) than r_0. Let $\texttt{exc}(r_0) \triangleq \bigvee_{r_i \in \texttt{str}(r_0)} \texttt{body}(r_i)$; i.e., the condition for exceptions to r_0.

Let $\texttt{cond}(\lambda) \triangleq \bigvee_{r_i : \texttt{head}(r_i) = \lambda} (\texttt{body}(r_i) \wedge \neg\texttt{exc}(r_i))$; i.e., the conditions under which literal λ is inferred. For each atom α, let $\texttt{def}(\alpha) \triangleq (\texttt{U} \vee \texttt{cond}(\alpha)) \wedge \neg\texttt{cond}(\neg\alpha)$, where $\texttt{U}$ is an atom that does not appear in κ and is unspecified in every percept of interest. Let $T[\kappa]$ be the theory comprising an equivalence $\texttt{def}(\alpha) \equiv \alpha$ for each atom α appearing in κ.

We show, next, a precise sense in which this set of equivalences captures the reasoning via the prioritized rules in κ.

Theorem 3 (Prioritized Rules as Equivalences)**.** *Consider a knowledge base κ, a percept $s = \emptyset$, and a scene s_i specifying every atom in rule bodies in κ. Then: $s_i \overset{\kappa,s}{\dashrightarrow} s_{i+1}$ if and only if $s_{i+1} = \{\alpha \mid \mathtt{def}(\alpha) \equiv \alpha \in T[\kappa], \mathtt{def}(\alpha)$ is true in $s_i\} \cup \{\neg\alpha \mid \mathtt{def}(\alpha) \equiv \alpha \in T[\kappa], \mathtt{def}(\alpha)$ is false in $s_i\}$.*

Proof. Each dominant rule in $s_i \overset{\kappa,s}{\dashrightarrow} s_{i+1}$ leads the associated equivalence to infer the rule's head. Atoms with no dominant rules are left unspecified by the associated equivalences. □

In Theorem 3 we have used the step operator with the percept $s = \emptyset$ simply as a convenient way to exclude the process of exogenous qualification, and show that endogenous qualification among prioritized rules is properly captured by the translation to equivalences. It follows, then, that if one were to define a step operator for equivalences and apply the exogenous qualification coming from an arbitrary percept s on top of the drawn inferences, one would have an equivalent step operator to the one using prioritized rules with the percept s.

What is critical, however, and is not used simply for convenience in Theorem 3, is the insistence on having a scene s_i in which every atom in rule bodies in κ is specified. Indeed, the translation works as long as full information is available, which is, of course, contrary to the perception semantics we have argued for. For general scenes the translation is problematic, as illustrated by the following two natural examples.

Example 3. *Consider a knowledge base κ with the rules r_1 : Bird $\rightsquigarrow$ Flying, r_2 : Penguin $\rightsquigarrow \neg$Flying, and the priority $r_2 \succ r_1$. The resulting equivalence is of the form: (U$\vee$Bird)$\wedge \neg$Penguin $\equiv$ Flying. By applying Theorem 3, when s_i is the scene $\{$Bird, Penguin$\}$, $\{$Bird, $\neg$Penguin$\}$, $\{\neg$Bird, Penguin$\}$, or $\{\neg$Bird, $\neg$Penguin$\}$, both the considered knowledge base κ and the resulting equivalence give, respectively, rise to the same inference $\neg$Flying, Flying, $\neg$Flying, or 'unspecified'. However, when $s_i = \{$Bird$\}$, the considered knowledge base gives rise to the inference Flying, whereas the resulting equivalence gives rise to the inference 'unspecified' for Flying.*

Since the two formalisms agree on what to infer on a fully-specified scene, they disagree on a general scene only when one infers 'unspecified' and the other does not; i.e., they never give rise to contradictory inferences *in any single step*. However, because of the multiple steps in the reasoning process, contradictory inferences may arise at the end. Further, it is not always the case that the knowledge base gives more specified inferences when the formalisms disagree in a single step.

Example 4. *Consider a knowledge base κ with the rules r_1 : $\beta \rightsquigarrow \alpha$, r_2 : $\neg\beta \rightsquigarrow \alpha$. The resulting equivalence is of the form: $\top \equiv \alpha$. On a scene s_i that specifies β, the formalisms coincide, but on the scene $s_i = \emptyset$, the considered knowledge base gives rise to the inference 'unspecified' for α, whereas the resulting equivalence gives rise to the inference α.*

Thinking that formalisms are more appropriate (in terms of completeness) if they give more specified inferences, comes from viewing them as computational processes meant to implement an underlying mathematical logic. As we have seen, however, case analysis might not be natural, and excluding it could be psychologically-warranted. In this frame of mind, it is the knowledge base that is more appropriate in both our examples, jumping to the conclusion that birds fly when no information is available on their penguin-hood, but avoiding to draw a conclusion that would follow by a case analysis.

Beyond the conceptual reasons to choose prioritized rules over equivalences, there are also certain formal reasons. First, reasoning with equivalences is an NP-hard problem: evaluating a 3-CNF formula (as the body of an equivalence) on a scene that does not specify any formula atoms amounts to deciding the formula's satisfiability [Michael, 2010; 2011]. Second, the knowledge representable in an equivalence is subject to certain inherent limitations, which are overcome only when multiple equivalences are used instead [Michael, 2014].

Of course, one could counter-argue that the case analysis, and the intractability of reasoning that we claim is avoided by using prioritized rules can easily creep in if, for instance, a knowledge base includes the rule $\varphi \rightsquigarrow \lambda$ for $\varphi = \beta \vee \neg\beta$, or φ equal to a 3-CNF formula. Our insistence on using read-once formulas for the body of rules avoids such concerns.

Our analysis above reveals that the choice of representation follows inexorably from the partial nature of perception. Prioritized rules are easy to check, while allowing expressivity through their collectiveness, and easy to draw inferences with, while avoiding non-naturalistic reasoning patterns.

3.3 Why Not Argumentation?

Abstract argumentation [Dung, 1995] has revealed itself as a powerful formalism, within which several forms of defeasible reasoning can be understood. We examine the relation of our proposed semantics to abstract argumentation, by considering a natural way to instantiate the arguments and their attacks.

Definition 6 (Arguments)**.** *An **argument** A for the literal λ given a knowledge base κ and a percept s is a subset-minimal set of explanation-conclusion pairs of the form $\langle e, c\rangle$ ordered such that: if e equals s, then c is a literal that is true in s; if e equals a rule r in κ, then c is the head of the rule, and the rule's body is classically entailed by the set of conclusions in the preceding pairs in A; c equals λ for the last pair in A.*

We consider below two natural notions for attacks.

Definition 7 (Attack Notions)**.** *An argument A_1 for literal λ_1 **attacks** an argument A_2 for literal λ_2 given a knowledge base κ and a percept s if there exist $\langle e_1, c_1\rangle \in A_1$ and $\langle e_2, c_2\rangle \in A_2$ such that $c_1 = \lambda_1$, $c_2 = \neg\lambda_1$, e_2 is a rule in κ, and either $e_1 = s$ or: (N_1) e_1 is a rule in κ and $e_2 \not\succ e_1$; (N_2) for every $\langle e, c\rangle \in A_1$ such that e is a rule in κ, it holds that $e_2 \not\succ e$.*

Definition 8 (Argumentation Framework)**.** *The **argumentation framework** $\langle \mathbb{A}, \mathbb{R}\rangle$ associated with a knowledge base κ and a percept s comprises the set $\mathbb{A}$ of all arguments for any literal given κ and s, and the attacking relation $\mathbb{R} \subseteq \mathbb{A} \times \mathbb{A}$ such that $\langle A_1, A_2\rangle \in \mathbb{R}$ if A_1 attacks A_2 given κ and s.*

Most typical semantics for abstract and logic-based argumentation frameworks give rise to multiple extensions, and differ from our formalism either because they produce credulous inferences, or because determining their skeptical inferences requires checking all such extensions. We show below that the *grounded semantics* can also be differentiated from our proposed formalism, even if not on these same grounds.

Definition 9 (Argumentation Framework Entailment)**.** *A set Δ of arguments **entails** a formula ψ if ψ is classically entailed by the set $\{\lambda \mid A \in \Delta$ is an argument for $\lambda\}$ of literals. An argumentation framework $\langle \mathbb{A}, \mathbb{R} \rangle$ **entails** a formula ψ if ψ is entailed by the grounded extension of $\langle \mathbb{A}, \mathbb{R} \rangle$.*

Theorem 4 (Incomparability with Argumentation)**.** *There exists a knowledge base κ, a percept s, and a formula ψ such that: (i) κ is resolute on s, and $(\kappa, s) \models \neg\psi$, (ii) for either attack notion N_1, N_2, the argumentation framework $\langle \mathbb{A}, \mathbb{R} \rangle$ associated with κ and s is well-founded, and $\langle \mathbb{A}, \mathbb{R} \rangle$ entails ψ.*

Proof. Let $\psi = a$. Consider a knowledge base κ with the rules

$$r_1 : \top \rightsquigarrow a \quad r_2 : \top \rightsquigarrow b \quad r_3 : \top \rightsquigarrow c \quad r_4 : c \rightsquigarrow \neg b$$
$$r_5 : b \rightsquigarrow \neg a \quad r_6 : b \rightsquigarrow d \quad r_7 : d \rightsquigarrow \neg a \quad r_8 : \neg a \rightsquigarrow d$$

and the priorities $r_4 \succ r_2$, $r_5 \succ r_1$, $r_7 \succ r_1$. Consider the percept $s = \emptyset$, on which κ is resolute. Indeed, $\texttt{trace}(\kappa, s)$ equals $\emptyset, \{a, b, c\}, \{\neg a, \neg b, c, d\}, \{\neg a, \neg b, c, d\}, \ldots,$ and $\texttt{front}(\kappa, s) = \{\{\neg a, \neg b, c, d\}\}$. Clearly, $(\kappa, s) \models \neg a$.

Consider, now, the set $\Delta = \{A_1, A_2\}$ with the arguments $A_1 = \{\langle r_3, c\rangle, \langle r_4, \neg b\rangle\}$, $A_2 = \{\langle r_1, a\rangle\}$. Observe that no argument A_3 is such that $\langle A_3, A_1\rangle \in \mathbb{R}$. Furthermore, any argument A_4 such that $\langle A_4, A_2\rangle \in \mathbb{R}$ includes either $\langle r_5, \neg a\rangle$ or $\langle r_7, \neg a\rangle$, and necessarily $\langle r_2, b\rangle$. Thus, $\langle A_1, A_4\rangle \in \mathbb{R}$, and therefore Δ is a subset of the grounded extension of $\langle \mathbb{A}, \mathbb{R} \rangle$. Clearly, $\langle \mathbb{A}, \mathbb{R} \rangle$ entails a. Also, $\langle \mathbb{A}, \mathbb{R} \rangle$ is well-founded. □

The incomparability — even for resolute knowledge bases and well-founded argumentation frameworks — is traceable to the skeptical and rigid semantics of argumentation, which meticulously chooses an argument (and thus a new inference) to include in the grounded extension, after ensuring that the choice is *globally* appropriate and will not be later retracted.

Such a treatment that reasons ideally and explicitly from premises to conclusions is dismissed by Bach [1984], as not even being a good cognitive policy. Rather, he stipulates that: *"When our reasoning to a conclusion is sufficiently complex, we do not survey the entire argument for validity. We go more or less step by step, and as we proceed, we assume that if each step follows from what precedes, nothing has gone wrong[.]"*. Our framework makes concrete exactly this point of view.

4 Learning Semantics

Bach [1984] aptly asks: *"Jumping to conclusions is efficient, but why should it be reliable?"*. We respond by positing that the reliability of a knowledge base can be guaranteed through a process of learning. An agent perceives the environment, and through its partial percepts attempts to identify the structure in the underlying states of the environment. How can the success of the learning process be measured and evaluated?

Given a set P of atoms, the ***P*-projection** of a scene s is the scene $s_P \triangleq \{\lambda \mid \lambda \in s$ and the atom of λ is in $P\}$; the ***P*-projection** of a set S of scenes is the set $S_P \triangleq \{s_P \mid s \in S\}$.

Definition 10 (Projected Resoluteness)**.** *Given a knowledge base κ, a percept s, and a set P of atoms, κ is **P-resolute** on s if the P-projection of $\texttt{front}(\kappa, s)$ is a singleton set; then, the unique scene in the P-projection of $\texttt{front}(\kappa, s)$ is the **P-resolute conclusion** of κ on s.*

Definition 11 (Projected Completeness)**.** *Given a knowledge base κ, a percept s, and a set P of atoms such that κ is P-resolute on s, and s_i is the P-resolute conclusion of κ on s, κ is **P-complete** on s if s_i specifies every atom in P.*

Definition 12 (Projected Soundness)**.** *Given a knowledge base κ, a compatible subset S of scenes, a percept s, and a set P of atoms, such that κ is P-resolute on s, and s_i is the P-resolute conclusion of κ on s, κ is **P-sound** on s against S if $\{s_i\} \cup S$ is compatible; i.e., there is no atom that is true (resp., false) in s_i and false (resp., true) in some scene in S.*

The notions above can, then, be used to evaluate the performance of a given knowledge base on a given environment.

Definition 13 (Knowledge Base Evaluation Metrics)**.** *Given an environment $\langle \texttt{dist}, \texttt{perc} \rangle \in \mathbb{E}$, a set P of atoms, and a real number $\varepsilon \in [0, 1]$, a knowledge base κ is **ε-resolute**, **ε-complete**, or **ε-sound** on $\langle \texttt{dist}, \texttt{perc} \rangle$ with **focus** P if with probability at least ε an oracle call to $\langle \texttt{dist}, \texttt{perc} \rangle$ gives rise to a state t being drawn from $\texttt{dist}$ and a scene s being drawn from $\texttt{perc}(t)$ such that, respectively, κ is P-resolute on s, κ is P-complete on s, or κ is P-sound on s against S, where S is the compatible subset of scenes determined by $\texttt{perc}(t)$.*

It would seem unrealistic that a single globally-appropriate tradeoff between these evaluation metrics should exist, and that a learner should strive for a particular type of knowledge base independently of context. Nonetheless, some guidance is available. Prior work [Michael, 2014] shows that one cannot be expected to provide explicit completeness guarantees when learning from partial percepts, and that one should focus on soundness, letting reasoning over the multiple rules being considered to improve completeness to the extent allowed by the perception process $\texttt{perc}$ that happens to be available.

The seemingly ill-defined requirement to ensure soundness against the *unknown* compatible subset S — effectively, the state t that underlies percept s — can be achieved optimally in some defined sense by (and only by) ensuring that the drawn inferences are consistent with the percept s itself [Michael, 2010]; or, in the language of this work, that the rules used are not exogenously qualified during the reasoning process.

Note that although the reasoning process can cope with exogenous qualification, this ability should be used in response to unexpected / exceptional circumstances, and only as a last resort. It is the role of the learning process *to minimize the occurrences of exogenous qualifications, and to turn them into endogenous qualifications*, through which the agent internally can explain why a certain rule failed to draw an inference.

Interestingly, the position above echoes evidence from the behavioral and brain sciences, asserting that the human brain is ultimately a predictive machine that learns (and even acts) in a manner that will minimize surprisal in its percepts [Clark, 2013]. Our analysis reveals that surprisal minimization is not necessarily an *end* in itself and a goal of the learning process, but rather a *means* to the reliability of the reasoning process.

Examining learnability turns out to offer arguments for and against our proposed formalism. On the positive side, learning when the atoms are not determined upfront remains possible and enjoys naturalistic algorithms for several problems [Blum, 1992]. Priorities between implications can be identified by learning default concepts [Schuurmans and Greiner,

1994] or learning exceptions [Dimopoulos and Kakas, 1995]. On the negative side, partial percepts hinder learnability, with even decision lists (hierarchical exceptions, bundled into single equivalences) being unlearnable under typical worst-case complexity assumptions [Michael, 2010; 2011]. Noisy percepts also critically hinder learnability [Kearns and Li, 1993].

Back on the positive side, environments without adversarially chosen partial and noisy percepts undermine the non-learnability results. The demonstrable difference of a collection of prioritized implications from a single equivalence further suggests that the non-learnability of the latter need not carry over to the former. Back on the negative side again, learning from partial percepts cannot be decoupled from reasoning, and one must simultaneously learn and predict to get highly-complete inferences [Michael, 2014]. Efficiency concerns, then, impose restrictions on the length of the inference trace, which, fortuitously, can be viewed in a rather positive light as being in line with psychological evidence on the restricted depth of human reasoning [Balota and Lorch, 1986].

Overall, our framework would seem to lie at the edge between what is or is not (known to be) learnable. This realization can be viewed as favorable evidence, since, one could argue, evolutionary pressure would have pushed for such an optimal choice for the cognitive processing in humans as well.

4.1 Boundaries of Learnability

Unsurprisingly, then, establishing the formal learnability of a knowledge base should be viewed as a major open challenge, and one that might need to be guided by a deeper understanding of how humans come to acquire their world knowledge. Nonetheless, we are able to provide some initial directions.

We consider certain complexity metrics for any knowledge base $\kappa = \langle \varrho, \succ \rangle$. The ***breadth*** of κ is the maximum number $\mathtt{b}$ of body atoms in a rule $r \in \varrho$. The ***depth*** of κ is the maximum number $\mathtt{d}$ such that $r_0 \succ r_1 \succ \ldots \succ r_{\mathtt{d}}$ for rules $r_i \in \varrho$. A Boolean (logic) circuit ***implements*** κ over a set P of atoms if for every percept s_P on which κ is P-resolute, the circuit on s_P outputs the P-resolute conclusion of κ on s_P. The ***circuit complexity*** of κ over P is the minimum size of such a circuit.

Theorem 5 (Unlearnability of Unbounded-Breadth Knowledge Bases)**.** *For any positive integer* $\mathtt{b}$*, and any set* $P_{\mathtt{b}}$ *of atoms, there exists a set* $\mathbb{E}_{\mathtt{b}}$ *of environments on each of which there exists a target knowledge base with breath* $\mathtt{b}$ *and depth* 1 *that is* 1*-resolute,* 1*-complete, and* 1*-sound with focus* $P_{\mathtt{b}}$*.*

Under cryptographic assumptions, and for any $\varepsilon > 0$*, there is no algorithm that, given* $\mathtt{b}$ *and oracle access to an environment* $\langle \mathtt{dist}, \mathtt{perc} \rangle \in \mathbb{E}_{\mathtt{b}}$*, runs in time polynomial in the circuit complexity of the target knowledge base over* $P_{\mathtt{b}}$*, and returns with probability at least* ε *a knowledge base that is* ε*-complete and* $(1/2+\varepsilon)$*-sound on* $\langle \mathtt{dist}, \mathtt{perc} \rangle$ *with focus* $P_{\mathtt{b}}$*.*

Proof. A circuit C over $\mathtt{b}$ input variables can be implemented within a knowledge base κ with breadth $\mathtt{b}$ and depth 1: κ includes the rule $r_0 : \top \rightsquigarrow \neg\alpha$; for each disjunct φ in the DNF representation of C, κ includes the rule $r : \varphi \rightsquigarrow \alpha$ and the priority $r \succ r_0$. The proof rests on the unlearnability of circuits under standard cryptographic assumptions [Kearns and Vazirani, 1994] and proceeds roughly analogously to existing unlearnability results; e.g., [Michael, 2014, Theorem 8]. □

Does a result analogous to Theorem 5 hold for knowledge bases with unbounded depth? The question is inapplicable if one disallows duplicate rules with different names (and priorities), since breadth bounds imply depth bounds. With duplication things are less clear, and the question remains open.

Our discussion points to the main research problem: establishing the existence of naturalistic learning algorithms for bounded-breadth and bounded-depth knowledge bases.

4.2 Does Learning Suffice?

One may wonder whether updating a knowledge base through a process of learning suffices, or whether extra revision processes are needed (e.g., removing parts of the knowledge base through belief revision [Peppas, 2008]). We show that learning new rules suffices to nullify the effect of existing parts of the knowledge base, if this happens to be desirable, without a "surgery" to the existing knowledge [McCarthy, 1998].

Definition 14 (Knowledge Base Equivalence)**.** *Knowledge bases* κ_1, κ_2 *are* ***equivalent*** *if for every percept* s *(on which both* κ_1 *and* κ_2 *are resolute),* $\mathtt{front}(\kappa_1, s) = \mathtt{front}(\kappa_2, s)$*.*

Below we write $\kappa_1 \subseteq \kappa_2$ for two knowledge bases $\kappa_1 = \langle \varrho_1, \succ_1 \rangle, \kappa_2 = \langle \varrho_2, \succ_2 \rangle$ to mean $\varrho_1 \subseteq \varrho_2$ and $\succ_1 \subseteq \succ_2$.

Theorem 6 (Additive Elaboration Tolerance)**.** *Consider two knowledge bases* κ_0, κ_1*. Then, there exists a knowledge base* κ_2 *such that* $\kappa_1 \subseteq \kappa_2$ *and* κ_0, κ_2 *are equivalent.*

Proof. Set $\kappa_2 := \kappa_1$. For each rule $r : \varphi \rightsquigarrow \lambda$ in κ_1, add to κ_2 the rule $f_1(r) : \varphi \rightsquigarrow \neg\lambda$ with a fresh name $f_1(r)$. For each rule $r : \varphi \rightsquigarrow \lambda$ in κ_0, add to κ_2 the rule $f_0(r) : \varphi \rightsquigarrow \lambda$ with a fresh name $f_0(r)$. Give priority to rule $f_0(r)$ over every other rule that appears in κ_2 because of κ_1. For every priority $r_i \succ_0 r_j$ in κ_0, add to κ_2 the priority $f_0(r_i) \succ_2 f_0(r_j)$. □

5 Conclusions

For everyday cognitive tasks (as opposed to problem solving tasks), humans resort to fast thinking [Kahneman, 2011], a form of which we have formalized. The formalism has been implemented in Mathematica, and is being used for an empirical exploration of possible learning strategies and other extensions. A Prolog meta-interpreter (supporting two natural representations of prioritized implications) for the reasoning semantics has also been implemented to evaluate reasoning.

Related to our work is a neuroidal architecture that exploits relational implications and learned priorities [Valiant, 2000a], but does not examine the intricacies of reasoning with learned rules on partial percepts. Extending this work to use relational rules can proceed via known reductions [Valiant, 2000b].

In addition to other possible extensions (e.g., asynchronous and / or probabilistic application of rules, decreasing gradient in rule activations, time-stamped atoms for temporal reasoning, coherence mechanism to ensure cognitive economy), and further formal analysis (e.g., establishing learnability and complexity results), we believe that, ultimately, the challenge is the design and development of a cognitive system with the following properties — towards which our formalism makes a concrete step, and following which it can be further extended:

(1) perpetual and sustainable operation, without suppositions on pre-specified and bounded collections of atoms or rules;

(2) continual improvement and evaluation, without designated training and testing phases for its learning process;

(3) autodidactic learnability, avoiding any dependence on some form of external human supervision [Michael, 2010];

(4) a holistic architecture, integrating seamlessly perception, reasoning, and learning in a coherent whole [Michael, 2014];

(5) non-rigidness and robustness, accommodating a graceful recovery from externally and / or internally-induced errors,

a point raised by von Neumann [1961] when envisioning the differences of a future logical theory of computation from formal logic: *"1. The actual length of 'chains of reasoning' [...] will have to be considered."* and *"2. The operations of logic [...] will all have to be treated by procedures which allow exceptions (malfunctions) with low but non-zero probabilities."*.

In mechanizing human cognition, it might be that getting the behavior right offers too little feedback [Levesque, 2014], and that looking into the human psyche is the way to go.

References

[Bach, 1984] Kent Bach. Default Reasoning: Jumping to Conclusions and Knowing When to Think Twice. *Pacific Philosophical Quarterly*, 65(1):37–58, 1984.

[Balota and Lorch, 1986] David A. Balota and Robert F. Lorch. Depth of Automatic Spreading Activation: Mediated Priming Effects in Pronunciation but Not in Lexical Decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(3):336–345, 1986.

[Blum, 1992] Avrim Blum. Learning Boolean Functions in an Infinite Attribute Space. *Machine Learning*, 9(4):373–386, 1992.

[Clark, 2013] Andy Clark. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[Collins and Loftus, 1975] Allan M. Collins and Elizabeth F. Loftus. A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, 82(6):407–428, 1975.

[Cormen *et al.*, 2009] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 3rd edition, 2009.

[Dimopoulos and Kakas, 1995] Yannis Dimopoulos and Antonis Kakas. Learning Non-Monotonic Logic Programs: Learning Exceptions. In Nada Lavrač and Stefan Wrobel, editors, *Proceedings of the 8th European Conference on Machine Learning (ECML 1995)*, volume 912 of *LNAI*, pages 122–137, Berlin, 1995. Springer.

[Dung, 1995] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artificial Intelligence*, 77(2):321–357, 1995.

[Harman, 1974] Gilbert Harman. *Thought*. Princeton University Press, Princeton, New Jersey, U.S.A., 1974.

[Kahneman, 2011] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

[Kearns and Li, 1993] Michael J. Kearns and Ming Li. Learning in the Presence of Malicious Errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[Kearns and Vazirani, 1994] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, Cambridge, Massachusetts, U.S.A., 1994.

[Levesque, 2014] Hector J. Levesque. On Our Best Behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[McCarthy *et al.*, 1955] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Report, Massachusetts Institute of Technology, A.I. Lab, Cambridge, Massachusetts, U.S.A., 1955.

[McCarthy, 1998] John McCarthy. Elaboration Tolerance. In *Working notes of the 4th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 1998)*, pages 198–216, London, England, U.K., 1998.

[Mercier and Sperber, 2011] Hugo Mercier and Dan Sperber. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 34(02):57–74, 2011.

[Michael, 2010] Loizos Michael. Partial Observability and Learnability. *Artificial Intelligence*, 174(11):639–669, 2010.

[Michael, 2011] Loizos Michael. Missing Information Impedements to Learnability. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory (COLT 2011)*, volume 19 of *JMLR: Workshop and Conference Proceedings*, pages 825–827, Budapest, Hungary, 2011.

[Michael, 2014] Loizos Michael. Simultaneous Learning and Prediction. In Chitta Baral, Laura Kovacs, Giuseppe De Giacomo, and Thomas Eiter, editors, *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR 2014)*, pages 348–357, Vienna, Austria, 2014. AAAI Press.

[Miller, 1956] George A. Miller. The Magic Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63(2):81–97, 1956.

[Murphy and Medin, 1985] Gregory L. Murphy and Douglas L. Medin. The Role of Theories in Conceptual Coherence. *Psychological Review*, 92(3):289–316, 1985.

[Open Letter, 2015] Open Letter. Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter. `http://futureoflife.org/misc/open_letter`, 2015. Accessed: 05 February 2015.

[Peppas, 2008] Pavlos Peppas. Belief Revision. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, *Handbook of Knowledge Representation*, chapter 8, pages 317–359. Elsevier Science, 2008.

[Schuurmans and Greiner, 1994] Dale Schuurmans and Russell Greiner. Learning Default Concepts. In Russell Greiner, Thomas Petsche, and Stephen José Hanson, editors, *Proceedings of the 10th Canadian Conference on Artificial Intelligence (AI 1994)*, pages 99–106, Banff, Alberta, Canada, 1994.

[Valiant, 2000a] Leslie G. Valiant. A Neuroidal Architecture for Cognitive Computation. *Journal of the ACM*, 47(5):854–882, 2000.

[Valiant, 2000b] Leslie G. Valiant. Robust Logics. *Artificial Intelligence*, 117(2):231–253, 2000.

[Valiant, 2006] Leslie G. Valiant. A Quantitative Theory of Neural Computation. *Biological Cybernetics*, 95(3):205–211, 2006.

[von Neumann, 1961] John von Neumann. The General and Logical Theory of Automata. In Abraham H. Taub, editor, *John von Neumann: Collected Works. Volume V: Design of Computers, Theory of Automata and Numerical Analysis*, chapter 9, pages 288–328. Pergamon Press, Oxford, 1961. Delivered at the Hixon Symposium, September 1948.