Fast and Loose Semantics for Computational Cognition

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

Abstract

Psychological evidence supporting the profound effortlessness (and often substantial carelessness) with which human cognition copes with typical daily life situations abounds. In line with this evidence, we propose a formal semantics for computational cognition that places emphasis on the existence of naturalistic and unpretentious algorithms for representing, acquiring, and manipulating knowledge. At the heart of the semantics lies the realization that the partial nature of perception is what ultimately necessitates — and hinders — cognition. Inexorably, this realization leads to the adoption of a unified treatment for all considered cognitive processes, and to the representation of knowledge via prioritized implication rules. Through discussion and the implementation of an early prototype cognitive system, we argue that such fast and loose semantics may offer a good basis for the development of machines with cognitive abilities.

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 relating to the debate on cognitivism vs. behaviorism aside, it is hard not to acknowledge that modern-day Artificial Intelligence research has pushed forward on the latter front more than on the former. As readily seen from the exceptionally powerful machine learning algorithms and the ingeniously crafted reasoning algorithms, more emphasis has been placed on the performance of the algorithms, and less emphasis on their design to replicate certain features of the human intelligence.

Making progress on the former front presumably necessitates the abandonment of rigid semantics and convoluted algorithms, and the investigation of naturalistic and unpretentious solutions, guided by psychological evidence on human cognition. At the same time, and in contrast to the specialization of much modern-day Artificial Intelligence research, a holistic view of cognition should be taken, whereby the processes of perception, reasoning, and learning are considered in a unified framework that facilitates their close interaction.

In this spirit, we present a simple unified semantics and an early prototype cognitive system built on top of this semantics. The questions that arise, the alternatives that one could have adopted, and the prior works that are relevant to such an endeavor, are much too numerous to be reasonably covered in the constraints of one paper. We view this work as a starting point for a fuller investigation of this line of research.

Perception Semantics

We assume that the environment determines at each moment in time a ***state*** that fully specifies what holds. An agent never fully perceives these states. Instead, the agent uses some prespecified language to assign finite names to ***atoms***, which are used to represent concepts related to the environment. The set of all atoms is not explicitly provided upfront. Atoms are encountered through the agent's interaction with 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 that 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 ***coherent*** if there exists a scene that is an expansion of each scene in S.

In simple psychological terms, a scene can be thought of as corresponding to the contents of an agent's working memory, where the agent's perception of the environment state, and any inferences relevant to this perception, are recorded for further processing. In line with psychological evidence, scenes actually used by an agent can be assumed to have a severely-limited number of specified atoms (Miller 1956).

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

When convenient, we represent a scene by the set of all literals (atoms or their negations) that are true in it, which suffices to fully capture the information available in a scene.

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 ***proba-***

bility distribution dist over states, capturing the possibly complex and unknown dynamics with which states are produced; a stochastic ***perception process*** perc determining for each state a probability distribution over a coherent subset of scenes. The agent has only oracle access to $\langle \texttt{dist}, \texttt{perc} \rangle$ such that: in unit time the agent senses its environment and obtains a ***percept*** s, resulting by an unknown state t being drawn from dist, and scene s then drawn from $\texttt{perc}(t)$. A percept s represents, thus, what the agent senses from state t.

An agent's ultimate goal is to act in a manner appropriate for the current state of the environment. We shall not attempt to specify here how this appropriateness is defined or quantified. Suffices to say that this decision making problem is sufficiently arduous, and that the agent can be aided if more information on the current state of the environment becomes available. Reasoning serves precisely this role: *to complete information that is not explicitly available in a percept*, and to facilitate, in this manner, the decision making process. To do so, it utilizes knowledge that the agent has been given, or has acquired, on certain regularities in the environment.

Reasoning Semantics

Since reasoning aims to complete information, it should be computationally easy to determine what inferences follow from the agent's knowledge, and inferences should follow often. We present our chosen representation for this knowledge, and defer the discussion of its properties for later.

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 *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 might not make this always explicit, the rules in ϱ are named, and the priority relation $\succ$ is defined over their names. Thus, 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** by rule r_2 on scene s_i if r_1, r_2 are applicable on s_i and conflicting, and $r_1 \not\succ r_2$.*

With the basic notions of rule qualification at hand, we can now proceed to define the reasoning semantics.

Definition 2 (Step Operator). *The **step operator** for a knowledge base κ and a percept s is a mapping $s_i \xrightarrow{\kappa, s} s_{i+1}$ 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 by a rule in κ on s_i; 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.[1] 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.

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 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 $s_i \xrightarrow{\kappa, s} s_{i+1}$ for each integer $i \geq 0$. The **inference frontier** of a knowledge base κ on a percept s is the minimal set $\texttt{front}(\kappa, s)$ of scenes found in a suffix of $\texttt{trace}(\kappa, s)$.*

Lemma 1 (Properties of Inference Frontier). $\texttt{front}(\kappa, s)$ *is unique and finite, for any knowledge base κ and percept s.*

In general, the inference frontier includes multiple scenes, and one can define many natural entailment notions.

Definition 4 (Entailment Notions). *A knowledge base κ applied on a percept s **entails** a formula ψ if ψ is:*
(E_1) true in a scene in $\texttt{front}(\kappa, s)$;
(E_2) true in a scene in $\texttt{front}(\kappa, s)$ and not false in others;
(E_3) true in every scene in $\texttt{front}(\kappa, s)$;
(E_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.

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

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

Note the subtle difference between the entailment notions E_3 and E_4: under E_4 an entailed formula needs to be not only true in every scene in the inference frontier (as required under E_3), but true *for the same reason* in every scene. This

[1] Overriding percepts can be accounted for by introducing rules that map perceived atoms to an internal version thereof, which are thereafter amenable to endogenous qualification by other rules.

excludes reasoning by case analysis, where an inference follows if it does in each of a set of collectively exhaustive cases. For a knowledge base κ and a percept s such that $\texttt{front}(\kappa, s) = \{\{\alpha\}, \{\beta\}\}$, for instance, the formula $\alpha \vee \beta$ is true in every scene in the inference frontier by case analysis, and is entailed under E_3, but is not entailed under E_4.

In the special case where the inference frontier includes a single scene, all proposed entailment notions coincide. We restrict our attention in the sequel, and define our entailment notion to be applicable 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 $(\kappa, s) \Vdash \psi$, if ψ is true in the resolute conclusion of κ on s.*

Definition 6 (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, $\texttt{front}(\kappa_1, s) = \texttt{front}(\kappa_2, s)$.*

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 3 (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. Initially, set $\kappa_2 := \kappa_1$. For every rule $r : \varphi \rightsquigarrow \lambda$ in κ_1, introduce in κ_2 the rule $f_1(r) : \varphi \rightsquigarrow \neg\lambda$ with a fresh name $f_1(r)$. For every rule $r : \varphi \rightsquigarrow \lambda$ in κ_0, introduce in 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, introduce in κ_2 the priority $f_0(r_i) \succ_2 f_0(r_j)$. The claim follows. □

The result above shows, in particular, that even if an agent is initially programmed with certain rules, the acquisition of new rules — through learning, or otherwise — could nullify their effect, if this happens to be desirable, without requiring a "surgery" to the existing knowledge (McCarthy 1998).

Illustration of Reasoning

For illustration, 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 : *Wings* $\rightsquigarrow$ *Bird*
r_5 : *Antarctica* $\wedge$ *Bird* $\wedge$ *Funny* $\rightsquigarrow$ *Penguin*
r_6 : *Flying* $\rightsquigarrow$ *Feathers*

and the priority $r_1 \succ r_2$. Consider applying κ on a percept $s = \{$*Antarctica*, *Funny*, *Wings*$\}$. Then, $\texttt{trace}(\kappa, s) =$

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

$(\kappa, s) \Vdash \neg$*Flying*, κ is resolute on s, and $\texttt{front}(\kappa, s) = \{\{$*Antarctica*, *Funny*, *Wings*, *Bird*, $\neg$*Flying*, *Penguin*$\}\}$.

One may note some back and forth when computing the inference trace above. Initially the inference *Bird* is drawn, which then gives rise to *Flying*, and later to *Feathers*. When *Penguin* is subsequently inferred, it leads rule r_1 to oppose the inference *Flying* coming from rule r_2, and in fact override and negate it. As a result of this overriding of *Flying*, *Feathers* is no longer supported through rule r_6, and is also dropped, even though no other rule directly opposes it.

In a sense, then, although the entailment semantics itself is skeptical in nature, the *process* through which entailment is determined is credulous in nature. It jumps to inferences as long as there is sufficient evidence to do so, and if there is no immediate / local reason to qualify that inference. When and if reasons emerge later that oppose an inference drawn earlier, then those are considered as they are made available.

This fast and loose aspect of the reasoning semantics is in line with well-investigated psychological theories for human cognition (Collins and Loftus 1975), which can guide its further development (e.g., by introducing a decreasing gradient in activations that limits the length of the inference trace).

Why Not Equivalences?

The proposed semantics uses implication rules for its representation, and priorities among those rules that are conflicting. One would not be at fault to wonder whether our chosen representation is nothing but syntactic sugar to hide the fact that one is simply expressing a single equivalence / definition for each atom. We investigate this possibility below.

Fix an arbitrary knowledge base κ. Let $\texttt{body}(r_0)$ and $\texttt{head}(r_0)$ mean, respectively, the body and head of rule r_0. Let $\texttt{str}(r_0)$ mean all rules r_i in κ such that r_0, r_i are conflicting, and $r_0 \not\succ r_i$; i.e., rules that are stronger than r_0. Let $\texttt{exc}(r_0) \triangleq \bigvee_{r \in \texttt{str}(r_0)} \texttt{body}(r)$; i.e., the exceptions to r_0.

Let $\texttt{cond}(\lambda) \triangleq \bigvee_{\texttt{head}(r_0)=\lambda} (\texttt{body}(r_0) \wedge \neg\texttt{exc}(r_0))$; 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 prioritize rules in κ.

Theorem 4 (Prioritized Rules as Equivalences). *Consider a knowledge base κ, a percept $s = \emptyset$, and a scene s_i in which every atom in κ is specified. Then: $s_i \xrightarrow{\kappa, s} s_{i+1}$ if and only if $s_{i+1} = \{\alpha \mid \texttt{def}(\alpha) \equiv \alpha \in T[\kappa], \texttt{def}(\alpha)$ is true in $s_i\} \cup \{\neg\alpha \mid \texttt{def}(\alpha) \equiv \alpha \in T[\kappa], \texttt{def}(\alpha)$ is false in $s_i\}$.*

Proof. It can be shown that each dominant rule in $s_i \xrightarrow{\kappa, s} s_{i+1}$ will lead the associated equivalence to infer the rule's head, and that when no dominant rule exists for an atom, then the associated equivalence will leave the atom unspecified. □

In Theorem 4 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 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 4, is the insistence on having a scene s_i in which every atom 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. In the case of general scenes, the translation is problematic, as illustrated by the following two simple examples.

As a first example, consider a knowledge base κ with the rules $r_1 : \textit{Bird} \rightsquigarrow \textit{Flying}$, $r_2 : \textit{Penguin} \rightsquigarrow \neg\textit{Flying}$, and the priority $r_2 \succ r_1$. Then, the following equalities hold:

$\mathtt{str}(r_1) = \{r_2\}$, and $\mathtt{exc}(r_1) = \mathtt{body}(r_2) = \textit{Penguin}$, and $\mathtt{cond}(\textit{Flying}) = \mathtt{body}(r_1) \wedge \neg\mathtt{exc}(r_1) = \textit{Bird} \wedge \neg\textit{Penguin}$.

$\mathtt{str}(r_2) = \emptyset$, and $\mathtt{exc}(r_2) = \bot$, and $\mathtt{cond}(\neg\textit{Flying}) = \mathtt{body}(r_2) \wedge \neg\mathtt{exc}(r_2) = \textit{Penguin} \wedge \neg\bot = \textit{Penguin}$.

$\mathtt{def}(\textit{Flying}) = (\mathtt{U} \vee \mathtt{cond}(\textit{Flying})) \wedge \neg\mathtt{cond}(\neg\textit{Flying}) = (\mathtt{U} \vee (\textit{Bird} \wedge \neg\textit{Penguin})) \wedge \neg\textit{Penguin} = (\mathtt{U} \vee \textit{Bird}) \wedge \neg\textit{Penguin}$.

Thus, the resulting equivalence is $(\mathtt{U} \vee \textit{Bird}) \wedge \neg\textit{Penguin} \equiv \textit{Flying}$. As it follows from Theorem 4, when s_i is the scene $\{\textit{Bird}, \textit{Penguin}\}$, $\{\textit{Bird}, \neg\textit{Penguin}\}$, $\{\neg\textit{Bird}, \textit{Penguin}\}$, or $\{\neg\textit{Bird}, \neg\textit{Penguin}\}$, both the considered knowledge base κ and the resulting equivalence give, respectively, rise to the same inference $\neg\textit{Flying}$, *Flying*, $\neg\textit{Flying}$, or 'unspecified'. However, when s_i does not specify both *Bird* and *Penguin*, the inferences of the formalisms may no longer coincide. When $s_i = \{\textit{Bird}\}$, the considered knowledge base gives rise to the inference *Flying*, whereas the resulting equivalence gives rise to the inference 'unspecified' for *Flying*.

Note that since the formalisms agree on all fully-specified scenes, they may disagree on general scenes only because one of them infers 'unspecified' when the other does not; the formalisms will never give rise to contradictory inferences *when used in a single step*. However, because of the multiple steps in the reasoning process, contradictory inferences may arise at the end. We omit the presentation of a formal result.

Based on the preceding example, one may hasten to conjecture that the knowledge base always gives more inferences, and that it is the equivalences that infer 'unspecified' when the formalisms disagree. This is not always the case.

As a second example, consider a knowledge base κ with the rules $r_1 : \beta \rightsquigarrow \alpha$, $r_2 : \neg\beta \rightsquigarrow \alpha$. It can be shown, then, that $\mathtt{def}(\alpha) = \mathtt{U} \vee \beta \vee \neg\beta = \top$, with a resulting equivalence $\top \equiv \alpha$. On scenes s_i that specify 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.

In either of the two examples, one's intuition may suggest that the formalism with an inference that specifies atoms is more appropriate than the other one. However, this intuition comes from viewing the two formalisms as computational tools aiming to draw inferences according to the underlying mathematical-logic view of the rules. We have argued, however, that case analysis is not necessarily a naturalistic way of reasoning, and that excluding it might be psychologically-warranted. In this frame of mind, it is now the knowledge base that draws the more appropriate inferences in both examples, jumping to the conclusion that birds fly when no information is available on their penguin-hood, and avoiding to draw a conclusion that would follow by a case analysis, when the scene on which the rules are applied does not provide concrete information for an inference to be drawn.

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 precisely to deciding the formula's satisfiability (Michael 2010; 2011b). Second, the knowledge representable in an equivalence is subject to certain inherent limitations, which are overcome when multiple rules are used instead (Michael 2014). We shall not elaborate any further, other than to direct the reader to the cited works for more information.

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.

Why Not Argumentation?

Abstract argumentation (Dung 1995) has revealed itself as a powerful formalism, within which 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 7 (Arguments). *An **argument** A for the literal λ given a knowledge base κ and a percept s is a minimal set of explanation-conclusion pairs of the form $\langle e, c\rangle$ that can be ordered so that: c equals λ for a pair in A; 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.*

Definition 8 (Attacks Between Arguments). *Argument A_1 for λ_1 **attacks** argument A_2 for 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, c_2 are the negations of each other, $c_1 = \lambda_1$, and either $e_1 = s$ or both e_1, e_2 are rules in κ such that $e_2 \not\succ e_1$.*

Definition 9 (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.*

Definition 10 (Grounded Extension). *A set $\Delta \subseteq \mathbb{A}$ of arguments is the unique **grounded extension** of an argumentation framework $\langle \mathbb{A}, \mathbb{R}\rangle$ if it can be constructed iteratively*

as follows: initially, set $\Delta := \emptyset$ and $\Omega := \mathbb{A}$; move an argument A_1 from Ω to Δ if there is no argument $A_2 \in \Omega$ such that $\langle A_2, A_1 \rangle \in \mathbb{R}$; remove an argument A_1 from Ω if there is an argument $A_2 \in \Delta$ such that $\langle A_2, A_1 \rangle \in \mathbb{R}$; repeat the last two steps until neither leads to any changes in Δ and Ω.

Definition 11 (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 $\langle \mathbb{A}, \mathbb{R} \rangle$.*

Perhaps counter to expectation, this natural translation to argumentation does not preserve the set of entailed formulas.

Theorem 5 (Incomparability with Argumentation Semantics). *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) 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. 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 $\langle r_2, b \rangle$. But, then $\langle A_1, A_4 \rangle \in \mathbb{R}$. Hence, Δ is a subset of the grounded extension of the argumentation framework $\langle \mathbb{A}, \mathbb{R} \rangle$. Clearly, $\langle \mathbb{A}, \mathbb{R} \rangle$ entails a. Note, also, that $\langle \mathbb{A}, \mathbb{R} \rangle$ is well-founded.

The claim follows immediately by letting $\psi = a$. □

At a conceptual level, the incomparability — even assuming simultaneously a resolute knowledge base and a well-founded argumentation framework — can be traced back to the fact that the argumentation semantics does not directly lend itself to a naturalistic reasoning algorithm. Unlike the original semantics, the *process* of considering arguments and attacks to compute the grounded extension proceeds in a very skeptical and rigid manner. It meticulously chooses which argument to include in the grounded extension, and thus which new inferences to draw, after ensuring that the choice is globally appropriate and will not be later retracted.

We conjecture that under additional natural assumptions, the two considered semantics coincide. But more important are the cases where any two given semantics diverge. Even if we were to accept that the grounded extension semantics is the "right" way to draw inferences, the goal here is not to define semantics that capture some ideal notion of reasoning, but semantics that are naturalistic and close to human cognition. Ultimately, only psychological experiments can reveal which semantics is more appropriate in this regard.

Learning Semantics

The knowledge base used for reasoning encodes regularities in the environment, which can be naturally acquired 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 $s_P \triangleq \{\lambda \mid \lambda \in s \text{ and } (\lambda \in P \text{ or } \neg\lambda \in P)\}$; the ***P*-projection** of a set S of scenes is the set $S_P \triangleq \{s_P \mid s \in S\}$.

Definition 12 (Projected Resolute). *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 13 (Projected Complete). *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 14 (Projected Accurate). *Given a knowledge base κ, a coherent 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-accurate** on s against S if every atom that is true (resp., false) in s_i is either true (resp., false) or unspecified in each scene in the subset S.*

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

Definition 15 (Knowledge Base Evaluation Metrics). *Given an environment $\langle \texttt{dist}, \texttt{perc} \rangle \in \mathbb{E}$, and a set P of atoms, a knowledge base κ is **$(1-\varepsilon)$-resolute**, **$(1-\varepsilon)$-complete**, or **$(1-\varepsilon)$-accurate** on $\langle \texttt{dist}, \texttt{perc} \rangle$ with **focus** P if with probability $1-\varepsilon$ 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-accurate on s against S, where S is the coherent subset of scenes determined by $\texttt{perc}(t)$.*

It would seem unrealistic that one universally-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. Formal results in previous work (Michael 2008; 2014) show that one cannot be expected to provide explicit completeness guarantees when learning from partial percepts, and that one should focus on accuracy, letting the reasoning process with the multiple rules it employs to improve completeness to the extent allowed by the perception process perc that happens to be available. This view is formalized in the cited works, as an extension of the Probably Approximately Correct semantics for learning (Valiant 1984).

The seemingly ill-defined requirement to achieve accuracy against the *unknown* subset S — effectively, the underlying state t of a percept s — can be achieved optimally, in a defined sense, by (and only by) ensuring that the inferences that one draws are consistent with the percept itself (Michael 2010). In the language of this work, as long as the rules used

are not exogenously qualified during the reasoning process, one can safely assume that the inferences drawn on atoms not specified in a percept are optimally accurate against S.

Although the reasoning process can cope with exogenous qualification, this feature should be used only as a last resort and in response to unexpected / exceptional circumstances. 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.

Examining learnability turns out to provide arguments for and against the proposed perception and reasoning semantics and knowledge representation. On the positive side, it is known that learning in settings where the relevant atoms are not determined upfront remains possible for many learning problems, and enjoys naturalistic learning algorithms (Blum 1992). Priorities between implications can be identified by learning default concepts (Schuurmans and Greiner 1994) or 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; 2011b). Noisy percepts can 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, the proposed semantics and representation would seem to lie at the very edge between what is and what is not (known to be) learnable. This realization can be viewed as favorable evidence for our proposal, since, one could argue, evolutionary pressure would have pushed for such an optimal choice for the cognitive processing in humans as well.

A Cognitive System

In line with our analysis so far, we give here some desiderata for the design and development of cognitive systems.

A cognitive system should have *(D1) perpetual and sustainable operation*, without suppositions on the existence of bounded collections of atoms or rules with which the system deals, nor on the existence of a specified time-horizon after which the system's behavior changes. In particular, a cognitive system should undergo *(D2) continual improvement and evaluation*, without designated training and testing phases, but rather with the ability to improve — if not monotonically, then roughly so — its performance (cf. Definition 15). Acquisition of knowledge in a cognitive system should *(D3) employ autodidactic learning*, avoiding to the extent possible the dependence on external supervision (Michael 2008; 2010). The constituent processes for perception, reasoning, and learning should be *(D4) integrated in a holistic architecture*, naturally interacting with, and feeding from, each other (Michael 2014). Lastly, a cognitive system's workings should *(D5) avoid rigidness and be robust* to moderate perturbations of its current state, allowing the system to gracefully recover from externally / internally-induced errors.

Embracing these desiderata, we have implemented a prototype cognitive system that perpetually obtains a new percept, reasons to draw inferences, evaluates its performance, and learns. Learning introduces new rules when new atoms are encountered, gradually corrects existing rules that are exogenously qualified to maintain accuracy, and produces mutations of rules to improve completeness. Rules that are introduced / produced are always assigned lower priority than existing rules to aid in completeness without affecting accuracy as a side-effect, and are maintained in a sandbox part of the memory until sufficient confidence is gained for their appropriateness. Rules that remain inapplicable for sufficiently long are removed through a process of garbage collection.

We are currently evaluating and comparing variants of the prototype system empirically, as a means to explore further the semantics proposed in this work, and to guide the formulation of results that one may prove in relation to the system's operation. Initial empirical evidence indicates that the learned knowledge bases are almost 1-resolute, and achieve surprisingly high completeness and accuracy, given the minimal structure that is available in the simulated environment that we have been using. The learned knowledge bases seem to comprise significantly more rules than one might have hoped for, an indication that a more aggressive garbage collection and / or a less aggressive introduction / production of new rules might be warranted; or, perhaps, an indication that when playing it fast and loose, the plurality of concepts and associations is not something to be dismissed, but embraced.

Further Thoughts

For everyday cognitive tasks (as opposed to problem solving tasks), humans resort to fast thinking (Kahneman 2011), for a form of which we have offered a first formalization herein. Closely related is a neuroidal architecture that uses relational implications and learned priorities (Valiant 2000a), but does not examine the intricacies of reasoning with learned rules on partial percepts. Extending our framework to relational rules can proceed via known reductions (Valiant 2000b).

Employing prioritized rules also, recent work on story understanding focuses on temporal aspects of perception and reasoning, which seem to call for more elaborate semantics (Michael 2013b; Diakidoy et al. 2014). Priorities are present in conditional preference networks, and work on learning the latter could be brought to bear (Dimopoulos, Michael, and Athienitou 2009; Michael and Papageorgiou 2013).

Systems that extract facts or answers from the Web are available (Carlson et al. 2010; Ferrucci et al. 2010), but work on endowing machines with websense (Michael 2013a) is considerably closer to the goal of acquiring human-readable rules (Michael and Valiant 2008; Michael 2008; 2009; 2010; 2014). Extending the framework to learn causal knowledge can build on top of existing work (Michael 2011a).

The desideratum for continual improvement and evaluation relates to the view of evolution as constrained learning (Valiant 2009), which may prove useful for designing monotonically improving (Michael 2012) learning processes.

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

Balota, D. A., and Lorch, R. F. 1986. 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.

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

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka Jr., E. R.; and Mitchell, T. M. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI 2010)*. Atlanta, GA, U.S.A.: AAAI Press.

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

Diakidoy, I.-A.; Kakas, A. C.; Michael, L.; and Miller, R. 2014. Story Comprehension through Argumentation. In *Proceedings of the 5th International Conference on Computational Models of Argument (COMMA 2014)*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, 31–42. Scottish Highlands, U.K.: IOS Press.

Dimopoulos, Y., and Kakas, A. 1995. Learning Non-Monotonic Logic Programs: Learning Exceptions. In *Proceedings of the 8th European Conference on Machine Learning (ECML 1995)*, volume 912 of *LNAI*, 122–137. Berlin: Springer.

Dimopoulos, Y.; Michael, L.; and Athienitou, F. 2009. Ceteris Paribus Preference Elicitation with Predictive Guarantees. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, 1890–1895.

Dung, P. M. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artificial Intelligence* 77(2):321–257.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3):59–79.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

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

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

McCarthy, J.; Minsky, M. L.; Rochester, N.; and Shannon, C. E. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Report, Massachusetts Institute of Technology, A.I. Lab, Cambridge, MA, U.S.A.

McCarthy, J. 1998. Elaboration Tolerance. In *Working notes of the 4th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 1998)*, 198–216.

Michael, L., and Papageorgiou, E. 2013. An Empirical Investigation of Ceteris Paribus Learnability. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*. Beijing, China: IJCAI/AAAI.

Michael, L., and Valiant, L. G. 2008. A First Experimental Demonstration of Massive Knowledge Infusion. In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR 2008)*, 378–389. Sydney, Australia: AAAI Press.

Michael, L. 2008. *Autodidactic Learning and Reasoning*. Ph.D. Dissertation, School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, U.S.A.

Michael, L. 2009. Reading Between the Lines. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, 1525–1530.

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

Michael, L. 2011a. Causal Learnability. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, 1014–1020. Barcelona, Catalonia, Spain: IJCAI/AAAI.

Michael, L. 2011b. Missing Information Impediments to Learnability. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT 2011)*, volume 19 of *JMLR Proceedings*, 825–828. Budapest, Hungary: JMLR.org.

Michael, L. 2012. Evolvability via the Fourier Transform. *Theoretical Computer Science* 462(1):88–98.

Michael, L. 2013a. Machines with WebSense. In *Working notes of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 2013)*.

Michael, L. 2013b. Story Understanding... Calculemus! In *Working notes of the 11th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 2013)*.

Michael, L. 2014. Simultaneous Learning and Prediction. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR 2014)*. Vienna, Austria: AAAI Press.

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

Schuurmans, D., and Greiner, R. 1994. Learning Default Concepts. In *Proceedings of the 10th Canadian Conference on Artificial Intelligence (AI 1994)*, 99–106.

Valiant, L. G. 1984. A Theory of the Learnable. *Communications of the ACM* 27(11):1134–1142.

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

Valiant, L. G. 2000b. Robust Logics. *Artificial Intelligence* 117:231–253.

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

Valiant, L. G. 2009. Evolvability. *Journal of the ACM* 56(1):3:1–3:21.