Quotation for Real-Time Metacognition

Matthew D. Goldberg MDGOLD@CS.UMD.EDU
University of Maryland, College Park

Darsana Josyula DARSANA@CS.UMD.EDU
Bowie State University

Don Perlis PERLIS@CS.UMD.EDU
University of Maryland, College Park

Abstract

Meta-reasoning about entities internal to agents themselves (such as their own beliefs) is a well-recognized ability, and leads to the need to quantify over propositions or formulas, especially in time-varying settings. In this position paper we sketch desiderata and a potential solution, via integrating a quotation-based, syntactic approach in the time-sensitive context of active logic.

1. Introduction

AI has emphasized general reasoning principles that can apply to a wide range of external entities. Thus if an object's support is removed, it can be expected to fall. While this is best seen as a default with potential exceptions, it applies to many entities unspecified in the principle itself. Similar considerations apply to reasoning about entities internal to an agent such as beliefs themselves, especially as they evolve over time. This has been widely recognized, and leads to the need to quantify over propositions or formulas. Traditional approaches, syntactic and modal, have been largely assessed in terms of avoiding paradoxes of self-reference. Here, we describe several kinds of reasoning that an ideal real-time and metacognitive agent ought to support. In particular, we propose a syntactic approach via quotation, in the context of an agent-like time-sensitive logic, where for instance contradictory (or paradoxical) inferences are seen as normal events to be reasoned about rather than disasters indicative of a badly designed formalism.

2. Autoepistemic Reasoning for Real-Time Metacognitive Agents

Broadly speaking, to enable wide-ranging metacognitive behavior, sentences from an agent's own knowledge base (KB) must be accessible to its reasoning. The agent should be able not only to use beliefs (e.g., to derive new beliefs), but also to mention (make assertions about) those beliefs (e.g., that one belief contradicts another, or that a belief was held until a given time). Thus it should be able to refer directly to current knowledge and to perform actions revising these beliefs. This in turn

depends on the ability to perform introspection on general sentence patterns (i.e., to determine their presence or absence among its beliefs over time).

A general agent certainly must also be able to continue reasoning in the presence of direct contradictions, without the "usual" ex contradictione quodlibet — from a contradiction everything follows. Indeed, a computational reasoner should note the contradiction, and enact an appropriate response (e.g., to prefer the more recent of the two contradictands). It should be able to handle contradictory information through abilities such as: to know where the contradiction lies, to identify each formula that contributes to inconsistency, and to consider which assumption(s) involved should no longer be believed.

Moreover, an agent that resolves a contradiction should be able to recount that process. There is utility in being able to note how an agent came about the new beliefs of its KB in the current moment: that at a previous point in time it was grappling with a contradiction, and that it held different beliefs, some of which were revised into current ones. In addition to involving reference to expressions of the agent's beliefs, this reasoning is also inherently time-based. That is, it's about the change in beliefs over time as the KB evolves: reasoning that there was once a contradiction, but this belief is no longer held in the present KB, and instead is recalled as believed in the past. Deadline reasoning also reflects both of these aspects. An agent with a deadline must focus on producing and enacting a plan while also taking into account that all reasoning requires time, and that both its planning and metacognition consume that resource.

Many of these kinds of reasoning involve (evolving) time, as well as (implicit) quotation or some other mechanism for expressing properties of beliefs. There are two standard approaches to the latter: syntactic and modal. We argue in favor of a syntactic approach where a form of quotation is natural, as in $\texttt{believed}(\text{“}\texttt{at}(\texttt{C},\texttt{L})\text{”},\texttt{T})$ — at time T, I had the belief $\texttt{at}(\texttt{C},\texttt{L})$, that car C is at location L. The evolving temporal aspect is additional and has to our knowledge been insufficiently addressed in the literature. However, as metacognition about beliefs will frequently be coupled with reasoning unfolding in time, we claim using a theory of belief for a commonsense agent without accommodating temporal evolution cannot suffice. Here we explore a syntactic approach in the context of our time-sensitive active logic.

3. Active Logic

Active logic is an "internal" logic (Perlis et al., 2017) for agent reasoners. One of its key features models reasoning as unfolding over a discrete series of timesteps with an evolving present time point (Elgot-Drapkin & Perlis, 1990; Purang, 2001). An active logic-based agent does not at any given time T know the closure of its beliefs under a consequence relation, but instead the believed set of formulas changes from one timestep to the next as applicable inference rules expand or contract the set of beliefs from that timestep. Active logic also naturally supports nonmonotonic reasoning, as defeasible beliefs from a prior timestep may be distrusted, modified, or otherwise disinherited when transitioning to a later point in time. A key and distinguishing characteristic is given by the "clock rule": an active-logic based agent will infer $\texttt{now}(\texttt{T}+1)$ at time $\texttt{T}+1$ from the belief $\texttt{now}(\texttt{T})$ (held at time T).

Such time-stratified reasoning is also made explicit in the logic itself, and can be reasoned about as inference progresses, which has been applied to active logic real-time planning (Nirkhe, 1995). Active logic also tracks former beliefs of the reasoner in the contents of earlier timesteps that preceded the present moment. A suitable formalism for referring to beliefs would endow active logic with a long-reaching form of introspection, where reference can be made not only to current beliefs, but to past beliefs as well (Miller, 1993).

This stratification is also what allows active logic to represent and reason about contradictory wffs. Thus, from beliefs $\mathtt{P}$ and $\neg\mathtt{P}$ in the KB at the same time $\mathtt{T}$, the logic infers at time $\mathtt{T}+1$ a contradiction between them, and the two conflicting formulas will be distrusted and disinherited. Contradictions inevitable in commonsense reasoning can be effectively managed without extensive machinery for preventing inconsistency from contaminating beliefs, and contradictions are no longer disastrous but common in the course of events. Contradiction-handling also proceeds at the same knowledge level as the rest of inference (see Dannenhauer et al. (2018) for an alternate multi-level approach), with the contradictory formulas as subjects of inference.

We thus believe active logic reflects some key desiderata of a real-time agent. If augmented with a quotation device, contradictions could be represented more precisely with a wff such as $\mathtt{contra}(\text{“}\mathtt{P}\text{”}, \text{“}\neg\mathtt{P}\text{”}, \mathtt{T})$. The logic would also thereby have a simple mechanism to recover from reasoning to a self-referential paradox, avoiding what has been seen as disastrous for syntactic theories (Montague, 1963; Thomason, 1980). While a number of researchers have employed syntactic approaches (Konolige, 1980; Perlis, 1985; Haas, 1986; Morgenstern, 1986), the usual lesson is to prefer modal treatments, with modal operators for knowledge, belief, or related notions (Reiter, 2001; Levesque & Lakemeyer, 2001; Lakemeyer & Levesque, 2012). Modal approaches to reasoning about belief, however, tend to assume logical omniscience, where an agent's belief set will be the possibly-infinite closure of its axioms under a consequence relation (although some approaches address this, e.g., Whitsey (2003)). Modal approaches also suffer from potential paradoxes (Perlis, 1988). We see as even more serious the fact that modal logics use what Ray Reiter called the "external design stance," of a designer looking in on the beliefs of an agent utilizing them, in contrast to internal logics.

4. Formalism and Examples

We previously introduced some reasoning patterns involving metacognition, that a knowledge-based agent should be capable of. We will now present formalized examples, using a general first-order syntax extended with quotation. We also assume a time-sensitive reasoning capability, such as active logic provides. The primary inference rule assumed of the reasoner is a generalized modus ponens, inferring from a conjunction of literals as the antecedent.

4.1 Referring to and updating an agent's current knowledge

An agent reasoner may be integrated within a physical system (for example, situated as the means of controlling a robot and issuing executable action commands). One way to enact actions from reasoning is by designating atomic "impetus" formulas (or iwffs) as corresponding to primitive actions. Once an iwff is detected by other components, the system attempts its action. Iwffs are not

persistent knowledge, but should remain only until prompting the rest of the system; accordingly, other components should update iwff knowledge when the desired action is attempted.

Consider a reasoner that has concluded it ought to move left: do(move(left)). This sentence, which may be a command to the wheels to move a certain amount to its left, is detected by other components, and once action is initiated the reasoner should be instructed to update the transient formula so that the agent knows an action is now being done. A command to update a formula into another would nest two quoted formulas inside: update("do(move(left))", "doing(move(left))"). Similarly, an update from *doing* to *done* may occur when the action completes.

4.2 Reasoning about the presence or absence of specific knowledge

Quotation can enable more expressive rules, which specify patterns of inference that would have otherwise required different instances of schemata.

Positive introspection — consider the following formula: goal(G) ∧ fulfills(C, G) ∧ know(C) ⇒ completed(G). If the agent adopts a goal G, and a condition C fulfills the goal, the agent notes its goal as met when that condition is believed. In this case, the *fulfills* predicate's first argument is a condition that must become true, while *know* is true if its argument occurs in the agent's KB. To keep the theory first-order and not present a variable where a predicate would occur, positive introspection uses the *know* predicate as shorthand. The truth value of *know* is based on searching the KB for a unifying formula, once its argument variable has been bound to a constrained formula structure. For instance, when C is color(ball_256, blue) and G is paint(ball_256, blue)), then during inference if the color formula exists in the KB, the completion of the goal would follow. In the absence of quotation devices, a schema instantiating a new rule for each predicate to appear in place of C would be required.

Generalized positive introspection — introspection should not be restricted to checking only ground formulas; the truth value of *know* should be determined by whether any formula is known that can unify with its argument. Consider a KB that in addition to the first rule of (4.2) above has goal(paint(ball_256, blue)) and fulfills(color(Obj, Color), paint(Obj, Color)). In this example, if the success condition for painting an object is the object's color matching the paint's, then the introspective lookup will proceed with the expression color(Obj, Color), and would retrieve a ground term like color(ball_256, blue) as a known matching fact.

Negative introspection — in a logic without a closed world assumption inference rule, reasoning about the absence of certain formulas may still be desired. For instance, an agent may have complete knowledge of its actions, but should not assert numerous facts in the KB about the actions it is not performing at each point in time. Negative introspection applied to formulas that should not appear in the KB gives a flexible means of inferring about their absence: empty_adjacent(X) ∧ ¬know(moving(X)) ⇒ can_move(X).

4.3 Distinguishing experience and quoted expressions

Information may arrive in the KB from observation, such as by robotic sensors, that carries a level of trust from direct experience. It also may arrive from sources that require reflection before trusting, such as fallible fellow agents. A system that has acquired information to evaluate must be able to

track it without considering a particular truth value carried. Treating these statements as quoted expression terms achieves this.

So, an agent whose microphone records an alarm would believe sound(alarm). But if the agent instead does not directly experience the alarm, but hears about it via a message from an unfamiliar source, the agent can instead record heard("sound(alarm)"). What is heard here certainly will not lead to the same inferences as the first sentence. Only when an evaluation procedure has run, and the contents are withdrawn from quotation and asserted as true (or false), would they have the same status as experienced information. An agent believing the following rule might then evacuate only in the case of a directly experienced alarm: sound(alarm) $\Rightarrow$ do(initiate_evacuation_procedure). Behavior when the quoted version was heard will depend on the results of any evaluation procedures triggered.

4.4 Fluents, slow and fast

A fluent expresses a condition that may or may not hold, and may change over time. As an example, imagine a car C moving from a location L over time; an agent with a belief about C's location held at time T, must revise its belief to match the changed situation at later time T′. Consider two cases, depending on the information available: (i) the agent observes at time T′ that C is no longer at that location; and (ii) it observes that C is moving between time T and time T′. We refer to these as slow and fast fluent evolution (in very loose analogy with Kahneman (2011). Our focus here is mostly on type (i), where the agent gets information about "slow" changes at discrete moments far enough apart that reasoning can keep up with the changes. So the belief at(C, L) is to be updated, without explicitly representing its negation, but replacing it with a formula about the time period during which at(C, L) held, and quotation provides a mechanism to represent and make such updates. [1]

5. Conclusion

General reasoning for an artificial agent will involve the intersection of real-time issues and metacognition, through cases such as an agent reasoning about its time-changing beliefs. We believe the contradiction-tolerant active logic extended with quotation is appropriate for addressing these, illustrated by the examples presented. However, as this paper has focused primarily on motivation and suitability, there remains much future work to be done in development of the formalism, and the realization of this new formalism in the software artifacts implementing active logic inference.

Acknowledgements

This work has been supported by a grant from DARPA, under Agreement No. HR00112090025.

1. One could include a time-parameter within these beliefs, which avoids a direct contradiction through distinct time arguments. But an agent cannot constantly observe every condition, and so must assume, as a default, that most conditions tend to continue to hold; this is the "commonsense law of inertia." So given at(C, L, T), one can reasonably assume at(C, L, T′), but if the location has changed the agent has not escaped resolving two formulas stating C is in different locations at once. Reasoning about the location fluent change still must be done.

References

Dannenhauer, D., Cox, M. T., & MuÃśoz-Avila, H. (2018). Declarative metacognitive expectations for high-level cognition.

Elgot-Drapkin, J. J., & Perlis, D. (1990). Reasoning situated in time I: basic concepts. *Journal of Experimental & Theoretical Artificial Intelligence*, *2*, 75–98.

Haas, A. R. (1986). A syntactic theory of belief and action. *Artificial Intelligence*, *28*, 245–292.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Konolige, K. (1980). A first-order formalization of knowledge and action for a multiagent planning system. SRI International Menlo Park CA Artificial Intelligence Center.

Lakemeyer, G., & Levesque, H. J. (2012). Only-knowing meets nonmonotonic modal logic. *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning* (pp. 350–357).

Levesque, H. J., & Lakemeyer, G. (2001). *The logic of knowledge bases*. MIT Press.

Miller, M. J. (1993). *A view of one's past and other aspects of reasoned change in belief*. Doctoral dissertation, University of Maryland at College Park, College Park, MD, USA.

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta Philosophica Fennica*, *16*, 153–167.

Morgenstern, L. (1986). A first order theory of planning, knowledge, and action. *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge* (pp. 99–114). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Event-place: Monterey, California, USA.

Nirkhe, M. V. (1995). *Time-situated reasoning within tight deadlines and realistic space and computation bounds*. PhD Thesis, University of Maryland at College Park, College Park, MD, USA.

Perlis, D. (1985). Languages with self-reference I: foundations. *Artificial Intelligence*, *25*, 301–322.

Perlis, D. (1988). Languages with self-reference II: knowledge, belief, and modality. *Artificial Intelligence*, *34*, 179–212.

Perlis, D., Brody, J., Kraus, S., & Miller, M. (2017). The internal reasoning of robots. *Thirteenth International Symposium on Commonsense Reasoning*.

Purang, K. (2001). *Systems that detect and repair their own mistakes*. PhD thesis, Department of Computer Science, University of Maryland, College Park, Maryland.

Reiter, R. (2001). *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT Press.

Thomason, R. H. (1980). A note on syntactical treatments of modality. *Synthese*, *44*, 391–395.

Whitsey, M. (2003). *Logical omniscience: a survey*. Technical Report NOTTCS-WP2003-2, School of Computer Science and IT, University of Nottingham.