An Active Logic Approach to Moore's Paradox

Shomir Wilson

Abstract

Moore's Paradox is a statement of the form "p and I do not believe that p", which can be true but is always absurd for an agent to assert. Such a statement poses a problem in that an agent is generally able to assert a true statement, but asserting Moore's Paradox implies a contradiction of beliefs. I present an interpretation of Moore's Paradox that preserves its absurdity while recasting it as a natural response to contradictory old and new information in an agent's knowledge. A byproduct of this approach is an intuitive formalization of the relationships between assertion, belief, and observation.

1 Introduction: Moore's Paradox

Moore's Paradox (from G. E. Moore) is a statement of the form, for some proposition p,

(1) p and I do not believe that p

or similarly

(2) p and I believe that not p

Its paradoxical aspects have made it a topic of study for artificial intelligence, philosophy, the cognitive sciences and other fields. Ludwig Wittgenstein, a contemporary of Moore's, considered it to be Moore's most impressive contribution to philosophy [7]. Its paradoxical nature is both initially obvious and deceptively deep. This is because, while it might be a true assertion, for a reasoning agent to assert it seems absurd. Consider, for clear illustration, an instantiation of (1):

(3) It is raining and I do not believe that it is raining

It could certainly be the case that, while it is raining, I have been in a room with no windows for several hours and do not expect rain in the weather forecast. However, for me to assert (3) via speech act or even cognizant affirmation produces an absurd situation. The first conjunct implies not only that it is raining but also that I believe it to be raining, because I could not assert it otherwise

(assuming I speak honestly and from knowledge, which seems reasonable). This stands in contradiction to the second conjunct, that I do not believe it to be raining.

A mere change of the point of view in (3) to third-person produces a more plausible statement:

(4) It is raining and Elizabeth does not believe that it is raining

As above, it is simple to conceive of a situation where it is raining and Elizabeth does not believe it to be so; and as long as Elizabeth is not the agent asserting (4), it is not problematic. Another ordinarily inconsequential change, a shift to past tense, also removes the problem:

(5) It was raining and I did not believe that it was raining

The above is a statement that I may assert without problem; it merely draws attention to the fact that I did not believe it to be raining, when in fact it was.

Instantiation and analysis of (2) produces essentially similar results.

Thus, it is clear that the problem with Moore's Paradox is that, unlike statements of most any other form, even when true, it cannot be asserted by a rational agent without implying a contradiction in its knowledge base. I will explain how this true but unassertable quality can come to be, and although that quality persists, I will show how it is significant in a plausible scenario in an agent's thought process.

2 Criteria for an Explanation

To better frame the nature of the desired explanation, I present some criteria that a good explanation of Moore's Paradox should fulfill. These are criteria that should make a solution meaningful, illustrative of more general issues, and nontrivial. The first and second below are similar to Sorensen's criteria [13] for an explanation, but the third differs markedly.

- A good explanation should distinguish Moore's Paradox from a simple contradiction. Since the first conjunct of (1) or (2) is a bare assertion (one without any qualification such as "believe"), and given that a rational agent only asserts what it believes, one could argue that those two statements are respectively equivalent to

  (6) I believe p and I do not believe that p

  (7) I believe p and I believe that not p

  The former of these statements is a contradiction in assertion, and the latter is an assertion of a contradiction in one's beliefs. Unlike Moore's Paradox, neither of these statements holds an elusive absurdity. Since

neither (6) nor (7) is possible for a rational agent to assert and maintain[1], it no longer seems the case that they are potentially true and capable of being asserted. The paradox reduces to a mere contradiction, due to the similar natures of belief and assertion. The reader may also note that, if an assertion of belief is equivalent to a bare assertion, then (6) and (7) can be rewritten concisely as "p and not p".

However, the reductive merit of this analysis is also its Achilles' heel. The unique absurdity of Moore's Paradox receives no treatment, and is simply shed in the transformation to a mere contradiction. A better explanation will show why at least some aspect of Moore's Paradox *is* a paradox, preferably in a way that also illustrates why (6) and (7) are not fully equivalent to (1) and (2) respectively.

- A good explanation should preserve our intuition that Moore's Paradox contains a contradiction. This would seem to frustrate efforts to follow the first criterion, but the roots of the absurdity in the paradox should not be lost: asserting it suggests the agent believes and disbelieves something at the same time. To satisfy both of these first two criteria, an explanation ought to illuminate our intuitions about the nature of assertion and belief, particularly how they are similar yet not identical.

- A good explanation should illuminate why some simple adjustments to Moore's Paradox, such as a change in person or tense, make it acceptable. Among frequently studied epistemological paradoxes, of which there are several [12], this paradox is unusual in that those simple changes in perspective suddenly render it a plausible assertion.

In fact, the peculiar acceptability of (4) and (5) suggests that an explanation could take advantage of such changes in perspective. Next, I will propose a logical formalism in which to examine Moore's Paradox that provides this ability.

3 Active Logic Preliminaries

Active logic [3] is a time-sensitive logical formalism designed for commonsense reasoning. It consists of a first order logical language, a set of inference rules capable of reasoning in time, and an observation function to represent environment knowledge as first-order formulas. It also contains syntactic features that allow a rule-based agent to isolate and recover from contradictions in its knowledge base [5]. Full presentations of active logic can be found other places, such as [1]; what follows is brief description of the salient details for the problem at hand.

Active logic is a step logic; that is, for a given (discrete) time step t there is a finite set of first order propositions that hold at that time. Formulas have

[1] *Maintaining* the assertion of a contradiction is perhaps the true difficulty. I will explain later a plausible role of contradictory mental states in a reasoner.

access to the current time step via a special predicate, $now(t)$, which is true if and only if t is equivalent to the current time. The update of $now(t)$ from one timestep to the next is done by active logic's clock rule:

$t_i : Now(i)$
$t_{i+1} : Now(i+1)$

Active logic employs an inference rule similar to *modus ponens*, which resembles reasoning from premises to conclusions in realtime:

$t_i : A, A \rightarrow B$
$t_{i+1} : A, A \rightarrow B, B$

There is also an inheritance rule that carries knowledge forward from one timestep to the next:

$t_i : A$
$t_{i+1} : A$

under the conditions that $\neg A$ is not in the knowledge base at step i and A is not equivalent to $Now(i)$. These rules enable a reasoner to keep an accessible history of how and when it acquires knowledge and arrives at conclusions.

Especially important to an analysis of Moore's Paradox is a simple active logic rule for contradiction detection and handling:

$t_i : P, \neg P$
$t_{i+1} : Contra(i, P, \neg P)$

The contradictands are pulled out of the knowledge base and placed in a meta-predicate, along with the time at which the contradiction was detected. This prevents the reasoner from deriving anything else from them, which could further sully its knowledge base with inconsistent conclusions. Other inference rules more appropriate for the reasoning environment can then handle the contradiction, perhaps by discarding both contradictands, reinstating the more reliable one, or further investigating the source of the contradiction.

4 Assumptions and Cognitive Framework

Before continuing, some assumptions and observations must be made about the nature of belief and bare assertion in an active logic-based agent:

- Current observations are more trustworthy of being bare assertions than past observations. If the agent is currently observing A, then A is at least as likely—and possibly more so—than if the agent observed A some time ago. This carries with it the more basic assumption that an agent's observations reflect accurately on the state of the world.

- When an observation of A ceases, but $\neg A$ remains unobserved, without further knowledge about the persistence of A, it is reasonable to continue holding A as a belief, but not necessarily a bare assertion. This can be considered a modified version of the Commonsense Law of Inertia [9],

altered to conservatively distinguish between certain "present" knowledge and what has been seen in the past. For instance, consider when an agent observes A for some period of time but then finds its ability to observe interrupted or obscured. If A is likely to change (e.g., a traveling train reaches its destination and stops, or a weather forecast calls for the temperature to drop) then that information should govern what the agent believes next. However, if little or no specific knowledge is available about the future of A (e.g., a block sits unattended on a table), then it is reasonable to believe—even if only for a limited time, or with fading confidence—that it remains the same.

- Asserted propositions are a subset of believed propositions. This follows mostly from the above two assumptions, but is worth stating separately. Since bare assertion expresses greater confidence than asserted belief, and a current observation justifies a bare assertion, observing A implies both a bare assertion of A and an asserted belief in A.

- When beliefs from two different observations in the past directly contradict, and no further information is available, the more recent observation is a better candidate for the current truth. Again, this follows partially from the first two assumptions. A skeptical agent could distrust both present and past observations, but in a world of changing conditions, consistency is sometimes hard to find. To be able to change its mind, an agent must have some criterion for resolving contradictory observations, and it seems reasonable for newer ones to supersede older ones.

These four points are formalized in the rules below, and added to them is a facility for contradiction detection and resolution. Let $Observe(A, t)$ hold when the agent observes circumstance A at time t. Let $Assert(A)$ and $Retract(A)$ have the effect of adding and subtracting (respectively) A from the agent's knowledge base in the next time step. Let $MoreRecent(A, B)$ hold when A has been asserted at least as recently as B in the reasoner's knowledge history. Finally, $Believe(A)$ holds when the agent believes A, and $Now(t)$ and $Contra(A, B)$ remain as presented in the previous section.

1. $(Observe(A,t)\&\neg Observe(A,t-1)\&Now(t)) \rightarrow$
   $(Assert(A)\&Assert(Believe(A)))$
2. $(\neg Observe(A,t)\&Observe(A,t-1)\&Now(t)) \rightarrow (Retract(A))$
3. $(Contra(Believe(A), \neg Believe(A))\&MoreRecent(A, \neg A)) \rightarrow$
   $(Retract(Contra(Believe(A), \neg Believe(A)))\&Assert(Believe(A)))$
4. $(Contra(Believe(A), \neg Believe(A))\&MoreRecent(\neg A, A)) \rightarrow$
   $(Retract(Contra(Believe(A), \neg Believe(A)))\&Assert(\neg Believe(A)))$

Informally, these can be restated as:

1. When A is observed and it was not observed in the previous time step, assert A and believe A.

2. When A is no longer observed and was observed in the previous time step, retract A (while retaining the belief in A).

3. When A is both believed and disbelieved, and A was asserted more recently than $\neg A$, retract the contradiction and believe A.

4. When A is both believed and disbelieved, and $\neg A$ was asserted more recently than A, retract the contradiction and believe $\neg A$.

5 The Paradox Revisited

With this framework in place, we can examine Moore's Paradox in a more formal fashion than (1) or (2). Respectively, they can be represented as

(8) $p\&Believe(\neg p)$
(9) $p\&\neg Believe(p)$

which are, for our purposes, equivalent[2]. By itself, this formalization reveals nothing; the contradiction is no more or less apparent than before, and the absurdity remains. However, given the above four formulas for reasoning about belief, we can construct a plausible scenario that produces Moore's Paradox in an active logic-based reasoner. Assume that those formulas are present in the knowledge base of a reasoner at every step. In the following scenario, the reasoner begins with the knowledge $\neg Believe(A)$ at time step t_0 and then encounters the contrary observation $Observe(A)$ at step t_1. The remaining steps illustrate the impact of the observation and belief formulas, although the formulas are omitted from the listing for clarity.

$t_0 : \neg Believe(A)$
$t_1 : \neg Believe(A), Observe(A)$
$t_2 : \neg Believe(A), A, Believe(A), Observe(A)$
$t_3 : A, contra(Believe(A), \neg Believe(A)), Observe(A)$
$t_4 : A, Believe(A), Obseve(A)$
$t_5 : A, Believe(A)$
$t_6 : Believe(A)$

The beginning of the observation of A in step t_1 triggers Formula 1, which adds $Believe(A)$ and A to the knowledge base for step t_2. Active logic's contradiction rule isolates $Believe(A)$ and $\neg Believe(A)$ in t_3, and Formula 3 resolves

[2]A more articulate semantic treatment would probably argue that they should be treated differently. For the moment, consider the two conjunctions $Believe(p)\&\neg Believe(p)$ and $Believe(p)\&Believe(\neg p)$. Both seem to be contradictions (though the second requires interpretation in a knowledge base), their first conjuncts are identical, and their second conjuncts have an intuitive relationship to each other.

the contradiction in t_4. In t_5, the observation of A has ended[3], and in t_6 Formula 2 has retracted the bare assertion of A.

More generally, this sequence of events represents the thought process of an agent changing its mind when presented with new information. The agent begins with a disbelief in A, due to some prior observation or reasoning. It then observes contradictory information, resolves the contradiction by prioritizing the more recent input, and ends with a belief of A. The formalized version of Moore's Paradox is present in step t_2, for the brief time between the processing of the new observation and the detection of a contradiction. The paradox is a symptom of the imminent need for knowledge repair.

The first criterion in Section 3 is satisfied by incorporating belief, bare assertion, and observation as necessary ingredients in the creation of the paradox. Although believed propositions are a superset of asserted ones in the given framework, the two sets are not identical and have crucial differences. The explanation also satisfies the second criterion by arriving at the contradiction $Believe(A)$ and $\neg Believe(A)$ in step t_2. Although one of those contradictands is not present in a literal isolation of Moore's Paradox, the explanation shows how Moore's Paradox fits into a larger picture that requires both of them.

This explanation also shows why the first-person present-tense perspective is essential to the absurdity of Moore's Paradox, fulfilling the third criterion. Looking back on its history of reasoning, an agent can assert that it believed one thing while the opposite was true, and it can also explain why it did that. A third person observer, as in (4), can assert the same. The paradox was the indication of a problem, and while its assertability is questionable, it is the essential trigger for the agent to realize that its knowledge is inconsistent.

Finally, one likely criticism of this explanation of Moore's Paradox is that it is too *ad hoc* to address the problem. The formulas and circumstances that create the paradox in the above explanation were selected specifically to do so, and one might object to that. However, I would argue that they should be judged on their own merit, as intuitive (even if simplistic) representations. The basic idea of an agent reasoning in time about conflicting information should be sufficiently general to support this explanation of the paradox.

6 Related Work

Several other treatments of Moore's Paradox have been written. Shoemaker [11] examines the paradox as a product of constituent relations between first and second-order beliefs. Kriegel [6] takes a similar approach but argues that conscious beliefs are always partly self-referential, and that it follows that Moore's Paradox is self-contradictory. Sorensen [13] characterizes the paradox as a *blindspot*: a proposition that, although possibly true, cannot be rationally accepted. However, little (if any) material exists on the situational possibility

[3]An earlier end to the observation would not have affected the reasoner's contradiction handling. I show the observation as persisting for a few timesteps to illustrate a scenario in which observability is not instantaneous.

of Moore's Paradox in an agent's knowledge base, or its temporal aspects, as presented in this paper.

Active logic has been used previously for automated reasoning about statements that require self-reflection on belief and knowledge. R. Moore's Brother Problem [8] is a statement of the form

(10) Since I do not know I have a brother, I must not.

Elgot-Drapkin and Perlis have shown that the step-logic and metareasoning features of active logic make it a natural environment for dealing with such statements [2].

An active logic reasoner has been implemented in software [10] and subsequently used to create ALFRED [4], an intelligent human-computer dialog system. ALFRED (Active Logic for Reason-Enhanced Dialog) serves as a natural language interface to a domain that the user wishes to control, such as trains on a railroad network. It uses active logic's contradiction tolerance to identify misunderstandings in dialog, such as underspecified commands or unknown words. ALFRED then initiates knowledge repair, either internally or by asking the user for clarification. This is an example of the more general process of noting an anomaly, identifying a failure, and guiding a solution into place, a process named the Metacognitive Loop, or MCL [1]. The contradiction repair process discussed in this paper can also be thought of as an applied instance of the MCL paradigm.

7 Conclusion

The active logic-based explanation of Moore's Paradox presented here sustains the notion that asserting it is absurd for a rational agent to do. The absurdity stems from a peculiar contradiction that is not fully apparent in the paradox, as it arises from the relationships between belief, bare assertion, and observation. However, absurdity in this treatment does not imply impossibility, as a rational agent that reasons in time about contradictions can use them as an indication of knowledge inconsistency. A simple logical framework for belief is necessary, but the assumptions behind it are arguably fair. The explanation builds on previous similar uses of active logic, and differs markedly from other existing treatments of the paradox.

References

[1] Michael L. Anderson and Donald R. Perlis. Logic, self-awareness and self-improvement: The metacognitive loop a nd the problem of brittleness. *Journal of Logic and Computation*, 15(1), 2005.

[2] J. Elgot-Drapkin and D. Perlis. Reasoning situated in time I: basic concepts. *J. of Experimental and Theoretical Artificial Intelligence*, 2:75–98, 1990.

[3] Jennifer Elgot-Drapkin, Sarit Kraus, Michael Miller, Madhura Nirkhe, and Donald Perlis. Active logics: A unified formal approach to episodic reasoning. Technical Report CS-TR-4072, 1999.

[4] Darsana Josyula, Michael L. Anderson, and Don Perlis. Towards domain-independent, task-oriented, conversational adequacy. In *Proceedings of IJCAI-2003 Intelligent Systems Demonstrations*, pages 1637–8, 2003.

[5] Darsana P. Josyula. *A Unified Theory of Acting and Agency for a Universal Interf acing Agent*. PhD thesis, Department of Computer Science, University of Maryland, College Park, 2005.

[6] Uriah Kriegel. Moore's paradox and the structure of conscious belief. *Erkenntnis*, 61:99–121, 2004.

[7] Norman Malcolm. *Ludwig Wittgenstein: A Memoir*. Oxford University Press, 1958.

[8] Robert Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25:75–94, 1985.

[9] Erik T. Mueller. *Commonsense Reasoning*. Morgan Kaufmann, 2006.

[10] K. Purang. Alma/carne: implementation of a time-situated meta-reasoner. In *Proceedings of the 13th International Conference on Tools with Artificial Intelligence*, pages 103–110, 2001.

[11] S. Shoemaker. Moore's paradox and self-knowledge. *Philosophical Studies*, 77:211–228, 1995.

[12] Roy Sorensen. Epistemic paradoxes. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2006.

[13] Roy A. Sorensen. *Blindspots*. Oxford University Press, 1988.