In Proceedings of the European Conference on Artificial Intelligence (ECAI), 1992

A Critical Examination of the Cohen-Levesque Theory of Intentions

Munindar P. Singh*
MCC, 3500 W. Balcones Center Drive, Austin, TX 78759, USA
Department of Computer Sciences, University of Texas, Austin, TX 78712, USA
msingh@mcc.com, msingh@cs.utexas.edu

Abstract. We examine the formal theory of intentions recently proposed by Cohen & Levesque [2]. We evaluate the assumptions made by this theory and their consequences relative to fairly general intuitions about intentions, especially as they are applied in AI domains. We show that the theory is conceptually problematic and that certain technical claims made by the authors are false or counterintuitive in most natural scenarios. Keywords: intentions, logic in AI.

1 Introduction

It is of great use to a scientific discipline when its theories are critically examined and their unstated assumptions, and the unexpected consequences of their explicit assumptions, are brought out. Unfortunately, there is not as much evaluation of claims and theories in AI as in other, more classical, disciplines. This is changing now, as AI becomes established itself. This paper contributes to this trend by critically examining a theory of intentions proposed by Cohen & Levesque (henceforth, C&L) [2].

The importance of intentions and actions in AI is obvious, but relatively few significant attempts have been made to formalize them. Intentions are involved in several subareas of AI: Planning, Natural Language, Multiagent Systems, and others. C&L's paper builds on Bratman's theory of intentions, from which they derive their main requirements for a theory of intentions [1]. These requirements include (a) the mutual consistency of intentions and beliefs, and (b) the tendency of intentions to persist.

C&L's theory is quite complex. We can discuss only its major features in the space available. Most of our remarks are conceptual. However, since C&L claim that their theory satisfies certain conceptual criteria on the basis of technical claims, we have to consider the latter also. We evaluate C&L's intuitions and assumptions, as well as the unexpected conceptual consequences of some of their technical results.

In §2, we describe the core concepts and definitions of C&L's theory to make this note self-contained. In §3, we evaluate the result of combining an assumption that all goals are dropped with the definition of a *persistent goal*. In §4, we consider (a) an incorrectly claimed property of intentions, and (b) how C&L's theory makes it impossible for an agent to succeed with, and often even to have, multiple intentions. Lastly, in §5, we evaluate C&L's theory relative to some general conceptual issues pertaining to intentions, and their formalization.

2 An Overview of C&L's Paper

C&L's theory begins with the notion of GOAL as primitively given. The GOALs of an agent specify the alternatives that he implicitly chooses. GOALs are assumed to be mutually consistent and consistent with the agent's beliefs. C&L define a *persistent goal*, or , as a goal that an agent persists with until he comes to believe that it has occurred or comes to believe that it has become impossible. Intentions are then defined as s for special conditions. Only one event may happen at a time (p. 225); there are no environmental events or processes.

The essential components of the theory are a model, M, which includes a set of linear courses of events or possible worlds (i.e., a function from integers to events), and a domain of objects ("things," "persons," and "events"). v binds variables to objects in the domain. Φ interprets predicates at different courses of events and at different time indices (i.e., integers). The semantics is given relative to a model, M, a course of events, σ, a valuation, v, and a time index, n. $M, \sigma, v, n \models p$ indicates that p is satisfied at the given point in the model under the binding v; $\models p$ indicates that p is satisfied at all points in all models under all bindings. We have collected below the relevant definitions to make this note self-contained. The

*This work was supported by the National Science Foundation (via grant # IRI-8945845 to the Center for Cognitive Science, U. Texas), and by MCC, Austin, Texas.

reader can easily skim over them on a first reading: the English gloss will suffice for most purposes here.

1. $M, \sigma, v, n \models P(x_1, \ldots, x_k)$ iff $\langle v(x_1), \ldots, v(x_k)\rangle \in \Phi[P, \sigma, v]$ (p. 225).

2. The conditions for $\neg, \vee, \exists, =$ are standard (p. 225).

3. $M, \sigma, v, n \models (\ a)$ iff $\exists m, m \geq n$, such that $M, \sigma, v, n[\![a]\!]m$ (p. 226). That is, a describes a sequence of events that happens next.

4. $M, \sigma, v, n \models (\ a)$ iff $\exists m, m \leq n$, such that $M, \sigma, v, m[\![a]\!]n$ (p. 226). That is, a describes a sequence of events that just happened.

$(\ x\ a)$ and $(\ x\ a)$ mean that x is the agent of a and that $(\ a)$ and $(\ a)$ hold, respectively (p. 230). The semantics of beliefs and goals are given using the alternativeness relations, B and G, respectively, each of which relates a world, an index, and a person to a world (p. 225). It is assumed that B is euclidean, transitive and serial, and G is serial. It is also assumed that $G \subseteq B$ (p. 227).

5. $M, \sigma, v, n \models (\ x\ \alpha)$ iff for all σ^* such that $\langle \sigma, n\rangle B[v(x)]\sigma^*$, $M, \sigma^*, v, n \models \alpha$ (p. 226). That is, α is true in all worlds accessible via B at σ and n.

6. $M, \sigma, v, n \models (\ x\ \alpha)$ iff for all σ^* such that $\langle \sigma, n\rangle G[v(x)]\sigma^*$, $M, \sigma^*, v, n \models \alpha$ (p. 226). That is, α is true in all worlds accessible via G at σ and n.

The semantics of actions is given in terms of $[\![]\!]$, which denotes that the given action takes place in the given interval (p. 226). $|$ and * are not needed here.

7. $M, \sigma, v, n[\![e]\!]n + m$ iff $v(e) = e_1, \ldots, e_m$ and $\sigma(n + i) = e_i, 1 \leq i \leq m$. The sequence of events denoted by e occurs from n to m in σ.

8. $M, \sigma, v, n[\![a; b]\!]m$ iff $\exists k, n \leq k \leq m$ such that $M, \sigma, v, n[\![a]\!]k$ and $M, \sigma, v, k[\![a]\!]m$. First a occurs after n and then b occurs after a, ending at m.

9. $M, \sigma, v, n[\![\alpha?]\!]n$ iff $M, \sigma, v, n \models \alpha$. $\alpha?$ occurs if α is true, and fails if it is false.

Further terms are defined using these primitives.

10. $(\ p\ q) \stackrel{\text{def}}{=} \forall c(\ c; q?) \rightarrow \exists a(a \leq c) \wedge (\ a; p?)$ (p. 230). If q occurs in the future, p occurs before it does; i.e., either (a) p holds in the future and $\neg q$ holds until then, or (b) neither p nor q ever holds.

11. $(p) \stackrel{\text{def}}{=} p \wedge (\ p)$ (p. 232). That is, knowledge is true belief.

12. $(p) \stackrel{\text{def}}{=} ((\ p) \supset (\ p))$ (p. 232). That is, an agent is competent about something iff if he believes it (now), he knows it (now).

13. $\Diamond\alpha \stackrel{\text{def}}{=} \exists x(\ x; \alpha?)$ (p. 227). α is eventually true if it is true after some sequence of events.

14. $\Box\alpha \stackrel{\text{def}}{=} \neg\Diamond\neg\alpha$ (p. 227). That is, α is always true iff $\neg\alpha$ is never true.

15. $(\ p) \stackrel{\text{def}}{=} \neg p \wedge \Diamond p$ (p. 230). That is, $(\ p)$ holds iff p is true in the strict future.

16. $(\ x\ p) \stackrel{\text{def}}{=} (\mathsf{GOAL}\ x\ (\ p)) \wedge (\ x\ \neg p) \wedge [\ ((\ x\ p) \vee (\ x\ \Box\neg p))\ \neg(\mathsf{GOAL}\ x\ (\ p))]$ (p. 236).

A *persistent goal* is a proposition p such that (a) it is a GOAL of the agent that p hold in the strict future, (b) the agent does not believe p now, and (c) condition (a) above will continue to hold unless the agent comes to either believe p or to believe that p will never hold.

17. $(_1\ x\ a) \stackrel{\text{def}}{=} (\ x\ [\ x\ (\ x\ (\ a))?; a])$ (p. 245).

That is, an agent intends to do action a iff he has a persistent goal to have done a immediately after believing it was about to happen. Intentions thus are special kinds of s. However, goals and persistent goals are more than just auxiliary definitions. Agents can have goals and persistent goals in addition to, and prior to, their intentions; such goals can even serve as reasons for the adoption of specific intentions. For instance, see C&L's examples on p. 255 and p. 256. Intentions are also be defined for propositions, but we concentrate on actions for reasons of space.

3 Can Persistence be Enough?

An important assumption of C&L is #3.25 (p. 233):

18. $\models \Diamond\neg(\mathsf{GOAL}\ x\ (\ p))$

This states that all goals are eventually dropped. This is an assumption about *all* agents and *all* their goals. C&L take this assumption to capture both the following restrictions: (a) agents do not persist forever with a goal, and (b) agents do not forever defer working on their goals (p. 233).

It is obvious that the assumption captures the former restriction, but it is not clear how it could capture the latter. This is because the assumption does not involve action in any way: it does *not* require that an agent eventually *act* on any goal, just that he drop it. In fact, it predicts that any agent, even one who cannot or would not act on a goal, will end up dropping it. Thus the assumption seems an inappropriate formalization of restriction (b) above.

Taken by itself, the assumption might seem still quite reasonable. However, in conjunction with other definitions, its effect, which we discuss next, is quite counterintuitive. C&L state the following theorem (theorem 4.5, p. 239):

"If someone has a persistent goal of bringing about p, p is within his area of competence, and, before dropping his goal, the agent will not believe p will never occur, then eventually p becomes true:

$\models (\ y\ p) \wedge \Box(\ y\ p) \wedge$

$\neg(\ (\ y\ \Box\neg p)\ \neg(\mathsf{GOAL}\ y\ (\ p))) \quad \supset \Diamond p$" The key step in the proof of this theorem is that each goal must eventually be dropped. Thus the definition of sanctions that eventually the agent come to believe p or come to believe $\Box\neg p$. However, the latter disjunct is

prevented by a premise of the theorem, so the former must occur. Competence then ensures that the given condition become true. Thus, given assumption 3.25, the proof is quite simple.

However, this theorem, though technically sound, is simply too powerful. A useful way to think of this theorem is as follows. Its antecedents correspond to different requirements that a given agent, y, must meet. **1**. $(\ y\ p)$ means that the agent, for whatever reason, has the given persistent goal. It is easy to meet this requirement for any arbitrary p while designing an agent—we can just set the agent to have that . **2**. $\neg((y\Box\neg p)\ \neg(\mathsf{GOAL}y(p)))$ is a requirement that one can satisfy in the design of an agent, by defining its belief-revision and goal-revision policies appropriately. **3**. The condition that $\Box(\ y\ p)$ is the hardest of these requirements, but is quite simple in practice. All we need to satisfy it is to ensure that the agent does not hallucinate (for the given p).

What makes theorem 4.5 excessively powerful is the fact that it relates virtually trivial requirements on agents to non-trivial conditions in the world. For example, let me be the agent and let p be my favorite implausible proposition: that Helmut Kohl is on top of Mt Everest. I can easily (1) have this , (2) for eternity not hold the belief that Herr Kohl will not ever make it to the top of Mt Everest, and (3) be always about p. Therefore, by the above theorem, Herr Kohl will get to the top of Mt Everest. He does not need to try; nor do I. He does not even need to know that his mountaineering feat had been my persistent goal.

C&L define an agent to be with respect to a proposition p such that if the agent believes p, then p actually holds. Competence, as defined, is quite easy to ensure. All that is required is that the given agent be cautious in adopting his beliefs. If the agent begins to believe a proposition only after he has seen it to be true himself, or inferred it on the basis of unimpeachable data, he will turn out to be with respect to it. For example, I am about Herr Kohl being on top of Mt Everest iff if I believe he is there, then he is indeed there. Note that cases when Herr Kohl is on top of Mt Everest, but I do not believe so do *not* affect my competence in the defined sense. That is, all that is needed is that the relevant belief not be adopted incorrectly; there may be cases when it could perhaps be adopted, but is not actually adopted.

If C&L had formalized properly the requirement that agents do not forever defer working on their goals, and had modified theorem 4.5 appropriately, it would still be too strong. This is because success cannot be guaranteed unless one additionally assumes that if an agent repeatedly attempts a goal, he will eventually succeed. But this is precisely a form of a *fairness* assumption common in standard distributed computing [3]. However, C&L explicitly reject fairness "in the computer science sense" (p. 233). Another way of properly formalizing the conditions under which agents succeed with their s would require explicit consideration of the agents' ability to achieve the given goal. Note that competence as defined by C&L is not the same as the ability to achieve something.

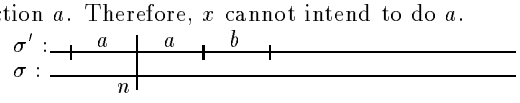
Later in the paper, C&L introduce relativized persistent goals, s as goals that need to be persisted with only as long as the agent believes that a certain relativizing condition does not hold. The relativizing condition captures the "reasons" (p. 254) for the adoption of the given persistent goal. We discuss s in detail in §5.2, but it should be sufficient to note here that versions of both theorem 4.5 and our examples can be designed for them. To continue the example above, I might never believe that my reasons for making Herr Kohl get atop Mt Everest have become false. In that case, according to the revised theory, he will still end up there.

It should be noted that, since intentions are special kinds of s, the above counterexample can be adapted to apply to them also: the propositions involved would simply have to be made correspondingly complex.

4 Some Properties of Intentions

4.1 Intending $a;b$ and Intending a

C&L claim without proof that "[o]ne can easily show that an agent who intends $a;b$ intends to do a" (p. 247). As a counterexample to this claim, consider a simple model in which there are two worlds, σ and σ'. At a time, n, let σ' be the only alternative belief and the only alternative goal world to σ (for the given agent); also let σ' be the only such world for itself. This satisfies the constraints on B and G quoted in §2. Since no variables are involved, the binding function, v, is irrelevant. $(_1\ x\ a;b)$ means that x has a certain as defined above. This implies that $a;b$ is done on σ' (after x believes that it is about to happen). Recall that $(\ x\ c)$ holds iff c was begun in the past and ended at the present time (see condition 2 in §2). Since $(\ x\ p)$ implies $(\ x\ \neg p)$, $(_1\ x\ a;b)$ also implies that x believes that $a;b$ is not done (after his believing that it is about to happen). Thus $a;b$ cannot currently be done on σ'. However, this is compatible with a being currently done on σ' (after x believes that it is about to happen). Thus x cannot have the appropriate about action a. Therefore, x cannot intend to do a.



Indeed, if σ had several alternative worlds, and action

a was done as above on just *one* of them, then the agent cannot intend it on σ.

It might seem that we have taken unfair advantage of the requirement that if p is a , it must be currently believed to be false. However, this requirement is a crucial component of C&L's theory. It occurs in the definition of "achievement" goals, which are the only kind of goals considered by them (p. 233). This requirement is used in proving some important results, e.g., (a) about the logical properties of s (pp. 236–237), and (b) about McDermott's *Little Nell Problem* (p. 219), which is a major motivation: "Our theory ... avoids this problem because an agent's having a requires that the goal be true later and that the agent *not believe it is currently true*" (p. 253, emphasis added). Thus the above counterexample is a natural consequence of C&L's own intuitions.

4.2 Having Multiple Intentions

A restrictive feature of C&L's theory is that it seems to be designed for agents who have no more than one intention at any time that they can act on. This is because C&L's definition of intentions states that an agent intends to do an action iff he has a to have done it *immediately* after believing it was about to happen. In other words, the entire sequence of events of which the given action is composed must occur at once without any other action by the given or by any other agent.

As a natural example, imagine an agent who runs a cafeteria. He takes orders from his customers, forms the appropriate intentions, and acts on them. When asked to serve coffee, he forms an intention to do the following complex action: pick up a cup; pour coffee into it; take the cup to the table. When asked to serve tea, he forms an intention to do the corresponding action for tea. Suppose now that two orders are placed: one for tea and the other for coffee. The agent adopts two intentions as described above. The agent initially ought to pick up a cup; let us assume that this is the action he chooses, and the one he believes he is about to do. However, at the time the agent picks up a cup, he might not have decided what action he will do after that, i.e., whether he will pour coffee or pour tea into the cup. Indeed, whether he pours coffee or tea into the cup might depend on other factors, e.g., which of the two brews is prepared, or whether other agents are blocking the route to one of the pots.

While this is a fairly ordinary state of affairs and a natural way for an agent to operate, it is disallowed by the theory. This is because the theory requires that for an intention to be *satisfied*, the agent believe beforehand that he is going to do the given action, no matter how complex it is (and then to do it). In the present example, this is not the case: the agent knows what he is doing before each subaction, but does not have a belief about a complex action *before* beginning to execute it. Also, for the agent to even *have* an intention, the theory requires that he have a to satisfy it in the above sense. Thus according to the theory our agent does *not* have either intention, in contrast to our intuitions that he has both.

Similarly, because of the requirement that only one event takes place at a time, the theory prevents an agent from succeeding with his intention, if another agent also acts. It is clear that to be of any use in AI, a theory of intentions must allow agents to have and act on several intentions and also allow different agents to act in an interleaved manner, if not concurrently. Thus C&L's definition is deeply unsatisfactory.

5 Discussion

We have looked at different components of C&L's theory of intentions and have evaluated its success in formally capturing some general intuitions about intentions. Now we evaluate it relative to some further conceptual aspects of a theory of intentions.

5.1 Commitment and Intention-Revision

The commitment-based analysis of intentions of Bratman [1, ch. 2] and Harman [4, p. 94] is a useful view of intentions for AI. In the above philosophical analysis, however, the commitment of agents to their intentions is merely a condition that holds in *normal* circumstances. That is, when real-life agents reason about their intentions, they normally persist with them. Indeed, Bratman says that intentions "resist (to some extent) revision and reconsideration" and involve "characteristic processes of reasoning and intention retention and (non)reconsideration." [1, p. 108]. This is only reasonable in light of the functional roles of intentions that he has described. However, C&L merge the agents' persistence into the very semantics of intentions.

Thus the important distinction between the *semantics* of intentions and the *policies* of intention-revision is lost. The semantics of intentions applies to all agents in all circumstances. It characterizes the states of the agents and their relationship to different possible worlds. Different policies of intention-revision apply only to some agents and only in some circumstances. They are easily overridden by other deliberation policies of the agent and factors such as the urgency of other tasks. For example, an agent who intends to get an umbrella from his closet will (probably) not continue to intend that if he realizes that

his house is on fire. On the other hand, if he intends to get his child from the bedroom, he will continue to intend to do so after his house catches fire. Similarly, if an agent who intends to go to a museum learns that the bridge he planned to drive on is closed, he might give up his original intention and go to some other museum, rather than rent an expensive helicopter.

C&L's approach thus does not capture the essential distinction between the semantics of intentions and policies of when to, and when not to, update them.

5.2 Relativized Persistent Goals

C&L define a relativized persistent goal, , as a generalization of (p. 254).

19. $(x\,p\,q) \stackrel{\text{def}}{=} (\mathsf{GOAL}\,x(p)) \wedge (x\neg p) \wedge [((x\,p) \vee (x\Box\neg p) \vee (x\,\neg q))\neg(\mathsf{GOAL}\,x(p))]$

That is, a relativized persistent goal is one that an agent will not give up until he believes it has been satisfied or believes it will never be true or believes its relativizing condition, q, is false. C&L try to use s to make it possible for agents to give up their goals. However, this attempt is quite unsatisfactory for a number of reasons. First, as explained in §3, the desired effect is not obtained. Second, it seems that this definition is provided as an afterthought—it comes after the main definitions and theorems of the paper. No theorems are given for .

Most importantly, the presence of an arbitrary condition, q, only serves to reduce the predictive power of the theory. The main motivation for a commitment-based view of intentions is that it yields intuitive properties in Bratman's theory [1]. By introducing a relativizing condition to use as a "reason" (p. 254), C&L state that an agent should persist with an intention as long as there are reasons for having it. But this is a position that Bratman anticipates and attacks as contradicting his view [1, p. 107]. Thus, in effect, C&L regress from the main motivation of their theory.

If the definition of is seen as incorporating a general policy of intention-revision (see §5.1), it only complicates the problem. This is because it prevents us from stating that an agent intends something *simpliciter:* we have to state in advance the conditions under which a given intention may be given up. Besides leading to a loss of predictive power, this is conceptually problematic. The entire set of possible exceptions cannot be specified in advance (this is akin to the *qualification* problem [5]). If these conditions are specified nonmonotonically, the semantics of intentions depends in an *ad hoc* manner on how an agent deliberates about his beliefs and intentions: whether or not an agent intends something now depends on what beliefs he would choose to adopt, or not to adopt, at a later time.

5.3 Rationality

It is not clear how the concepts of and may be reconciled with rationality. C&L's state that agents can *adopt* s (e.g., see the usage on p. 254, among other places). It seems to us, however, that s ought better to be considered as descriptions of possible states of agents. And for rational agents, these are states that they must *never* enter into. This argument applies to s as well, since they also require an agent to continue to have a goal based on criteria external to himself. Whether an agent adopts a or an , he is then at the mercy of other agents and of the world: no rational agent should voluntarily get himself into such a position. As shown in §3, assumption 3.25 and theorem 4.5 serve only to hide this inherent irrationality by making it an artificial reason for success.

6 Conclusion

C&L's theory is an important first step. However, their approach yields counterintuitive results and also suffers from conceptual shortcomings. In parts it seems unduly complicated—some of their own intuitions are not properly captured in it. To their credit, however, it is only because the theory is formal and precise that we could uncover the problems with it.

References

[1] Michael E. Bratman. *Intention, Plans, and Practical Reason.* Harvard University Press, Cambridge, MA, 1987.

[2] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[3] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B. North-Holland Publishing Company, Amsterdam, The Netherlands, 1990.

[4] Gilbert Harman. *Change in View.* MIT Press, Cambridge, MA, 1986.

[5] John McCarthy. Epistemological problems of artificial intelligence. In Matthew L. Ginsberg, editor, *Readings in Nonmonotonic Reasoning*, pages 46–52. Morgan Kaufmann, Inc., Los Altos, CA, 1987. Reprinted from IJCAI-77.