Intention Is Commitment with Expectation

James Creel
Department of Computer Science
Texas A&M University

Christopher Menzel
Department of Philosophy
Texas A&M University

Thomas Ioerger
Department of Computer Science
Texas A&M University

Abstract

Philosophers and scientists interested in Artificial Intelligence emphasize the role of intention in rational communication and interaction. The papers on rational agency by Cohen and Levesque are among the first to develop a logical theory of intention that accords with a large body of philosophical work, and provide the standard reference on BDI logics. However, Singh shows the theory to have certain logical inconsistencies and permit certain absurd scenarios. We present a modification of the theory that preserves the desirable aspects of the original while addressing the criticism of Singh. This modification is achieved by the refinement of certain assumptions, the introduction of an additional operator describing the achievement of expectations, and new, clarified definitions of intention. The amended theory fulfills a multitude of philosophical desiderata for intention, allowing for the representation of prior intention and intention-in-action and appropriately constraining agents' action and deliberation. Most importantly, the criticisms of Singh are shown to motivate an additional desiderata for intention: that action should not be intended for its own sake, but rather to bring about a desired proposition or situation.

Introduction

Much work has been devoted to integrated theories of rational agency (Cohen & Levesque 1990; Herzig & Longin 2004; van der Hoek, van Linder, & Meyer 1997; Rao & Georgeff 1991; Singh & Asher 1993). Though intentional states like belief, desire, and obligation figure prominently in such theories, intention itself stands out perhaps as the most interesting due to its relationships with the other notions and its integral role in action, planning, communication, and philosophy of mind. Cohen and Levesque's (henceforth C&L) influential theory of intention as a persistent goal (Cohen & Levesque 1990) is based on a possible worlds semantics with belief and choice primitively given, and is the standard reference on BDI logics (Woolridge 2002). The theory has been employed for the development of theories of joint intentions (Cohen, Levesque, & Nunes 1990) and speech acts (Cohen & Levesque 1995). Furthermore it provides some motivation of Jennings' industrial robotics application (Jennings 1995) and Tambe's multiagent system architecture, STEAM (Shell for TEAMwork) (Tambe 1997).

Despite its prevalence, the theory suffers from certain logical problems identified by Singh (Singh 1992). In particular, agents cannot intentionally do actions of the same type twice consecutively, agents cannot simultaneously hold and act on multiple intentions without knowing fully in advance how to interleave the steps in the intentions' respective plans, and intentions permit certain absurd scenarios. We show how agents can intend the same action repeatedly by expecting their intentions to produce different desired outcomes. Then, we show how agents can maintain multiple intentions when they are not required to have perfect knowledge of their plan in advance. Finally, the absurd scenarios described by Singh occur because an agent's intentions are assumed to eventually come true under certain circumstances that do not require any action on the part of the agent. This problem is handled by an assumption that agents do not infinitely procrastinate.

The original theory and the revised theory both suffer from some vestige of the side-effect problem (Orilia 1996) and the full fledged logical omniscience problem (Woolridge 2002). However, the revised theory seems to concur with certain conclusions of (Searle 1983). Searle argues that action as a whole consists of the following series of events: prior intention causes intention-in-action which causes movement. In the theory given here, future-directed (weak) intention and present-directed (strong) intention approximately correspond to prior intention and intention-in-action respectively.

To the best of our knowledge, the explicit solution of Singh's criticisms has not been addressed in the context of C&L's theory, although (Singh & Asher 1993) proposes a logic of belief and intention based on Discourse Representation Theory rather than a possible worlds approach, avoiding the logical omniscience problem. The theories' profoundly different bases prevent a detailed treatment or comparison in this space.

The improvement of C&L's original work has also been taken up by Herzig and Longin, who provide a sound and complete propsitional logic that is a simplification of the quantified modal logic of C&L (Herzig & Longin 2004). That work redefines intentions without reference to actions. Here we consider action as integral to the definition of intention.

Cohen and Levesque's Theory of Intention

Cohen and Levesque propose a linear time temporal logic integrated with a KD conative logic and KD45 doxastic logic [1]. Syntax and semantics are introduced in tables 1 and 2 in the context of the proposed amendments to the theory, but those syntax and semantics are consistent with and provide clarification for the current discussion. The notation $[A \rightarrow B]$ is meant to denote the set of all functions from A to B. C&L define a model M as a structure $\langle U, Agt, T, B, C, \Phi \rangle$. Here, U, the universe of discourse consists of the following three sets: Θ a set of things, P a set of agents, and E a set of primitive event types. $T \subseteq [Z \rightarrow E]$ is a set of linear courses of events (intuitively, possible worlds) specified as a function from the integers (intuitively, times) to elements of E. $B \subseteq T \times P \times Z \times T$ is the belief accessibility relation which is Euclidean, transitive and serial. $C \subseteq T \times P \times Z \times T$ is the choice (i.e. goal)[2] accessibility relation, and is serial. The constraint of *realism* is imposed on the accessibility relataions: $C \subseteq B$. Φ interprets predicates, that is $\Phi \subseteq [Predicate^k \times T \times Z \times D^k]$ where $D = \Theta \cup P \cup E^*$.

For lack of space, we do not include every definition found in C&L's extensive theory. Informally, $(\mathsf{HAPPENS}\ a)$ means action expression a occurs at the present time. $(\mathsf{DONE}\ a)$ means action expression a just happened. One may optionally specify the event's single agent x, as in $(\mathsf{HAPPENS}\ x\ a)$. Note that the semantics of $(\mathsf{DONE}\ a)$ and $(\mathsf{HAPPENS}\ a)$ rely on the relation $\| \ \|$ defined in table 2 numbers 12, 13, and 14 which describes when an action occurs between two points in time. The semicolon ; is used to indicate consecutive occurrence of actions. The test action $\phi?$ occurs instantaneously if ϕ is the case.

The symbol $\Diamond$ is an abbreviation for "eventually" as in $\Diamond\phi \equiv \exists e, (\mathsf{HAPPENS}\ e; \phi?)$. The symbol $\Box$ is an abbreviation for "always" as in $\Box\phi \equiv \neg\Diamond\neg\phi$. The concept of "later" is defined as eventually but not currently. That is, $(\mathsf{LATER}\ \phi) \equiv \neg\phi \wedge \Diamond\phi$. To say that formula ϕ comes true before formula ψ (if ψ comes true) we write $(\mathsf{BEFORE}\ \phi\ \psi)$.

An achievement goal is a goal to bring about some as yet untrue condition, as defined by $(\mathsf{A\text{-}GOAL}\ x\ \phi) \equiv (\mathsf{CHOOSE}\ x\ (\mathsf{LATER}\ \phi)) \wedge (\mathsf{BEL}\ x\ \neg\phi)$. C&L define a persistent goal, which captures the notion of commitment, as a type of achievement goal.

$$
\begin{array}{l}
(\mathsf{P\text{-}GOAL}\ x\ \phi) \equiv (\mathsf{CHOOSE}\ x\ (\mathsf{LATER}\ \phi)) \wedge \\
\quad (\mathsf{BEL}\ x\ \neg\phi) \wedge \\
\quad [(\mathsf{BEFORE}\ ((\mathsf{BEL}\ x\ \phi) \vee (\mathsf{BEL}\ x\ \Box\neg\phi)) \\
\qquad \neg(\mathsf{CHOOSE}\ x\ (\mathsf{LATER}\ \phi))]
\end{array}
$$

That is, a commitment is an achievement goal that must be believed to be either achieved or impossible before being dropped.

Intentions toward actions and propositions are then defined as P-GOALs to have brought something about immediately after believing one was about to do it. Intention toward action is defined as

$$
\begin{array}{l}
(\mathsf{INTEND}_1\ x\ a) \equiv (\mathsf{P\text{-}GOAL}\ x \\
\quad (\mathsf{DONE}\ x\ (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ a))?; a))
\end{array}
$$

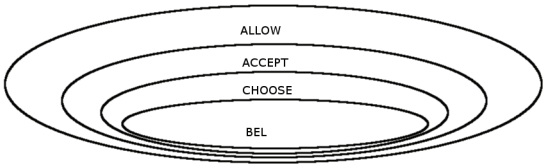
where a is any action expression. That is, an intention toward an action is a commitment to have brought about that action immediately after having believed it was about to occur. C&L define intention toward a proposition like so:

$$
\begin{array}{l}
(\mathsf{INTEND}_2\ x\ p) \equiv \\
\quad (\mathsf{P\text{-}GOAL}\ x\ \exists e, (\mathsf{DONE}\ x \\
\qquad [(\mathsf{BEL}\ x\ \exists e', (\mathsf{HAPPENS}\ x\ e'; p?)) \wedge \\
\qquad \neg(\mathsf{CHOOSE}\ x\ \neg(\mathsf{HAPPENS}\ x\ e; p?))]?; e; p?))
\end{array}
$$

That is, an intention toward a proposition is a commitment to have done some action e that brings about the proposition immediately after having believed that there exists some action e' that will bring about the proposition, and having accepted that the particular action e is the one that brings about the proposition.

Comments on Belief and Choice

Unlike in previous work on C&L's theory, we include the duals of BEL and CHOOSE in the syntax and semantics, these duals being ALLOW[3] and ACCEPT respectively. The axioms of the modal logic and the *realism* constraint make clear the relationships between these modalities within the space of all possible formulas, which we depict graphically below.



Thus, the allowed formulas are the largest subset of the set of all formulas, and the believed formulas are the smallest.

Problems Identified by Singh

In Singh's original criticism, three problems are noted.
(1) The *From persistence to eventualities* theorem (C&L p. 239), which links persistent goals with eventual outcomes, is excessively powerful. Specifically, agents' intentions can be brought about inappropriately, without action on the part of the agent: if an agent adopts a persistent goal toward some improbable proposition, the agent is competent with respect to the proposition [4], and the agent does not come to believe the proposition impossible before dropping the goal, then the proposition will come about. Singh notes that the agent is not required to act in order to bring the proposition about, making the theorem too powerful. The direct solution

[1] This axiom schemata for belief follows Hintikka (Hintikka 1962) and corresponds to a "Weak S5" modal logic.

[2] We follow Herzig and Longin in designating the primary conative modality as "choice" rather than "goal". This change in nomenclature is logically insignificant.

[3] We might also use SUSPECT but suspicion seems to carry the connotation that a formula is in the case in *most* worlds accessible via B rather than at least one such world.

[4] If an agent is competent toward a proposition, then the agent is always correct in believing the proposition is the case.

we employ here is to modify the assumptions of the theory, properly prohibiting both infinite persistence and infinite deferral.

(2) When an agent executes the same action (i.e. actions of the same primitive event type) twice consecutively, it is impossible that the second action be intentional. This is because INTEND_1, intention toward action, is a persistent goal that some action has been done under certain circumstances. A persistent goal requires that an agent believe the proposition is not true. Thus, an agent cannot have the persistent goal (much less the intention) to have done what he has just done. One may solve this problem by requiring that agents form intentions toward actions in the context of plans, so that actions are expected to bring about a desired outcome. Then an agent may intentionally perform the same type of action twice consecutively *as long as the intended outcome of the second action is different from the actual outcome of the first action*. At first glance one may consider this inadequate for repetitive cases such as chopping down a tree or ringing an alarm repeatedly. Further consideration makes it clear that if a lumberjack intended exactly the same outcome from each chop, then he would be intending the same cut every time, when in fact he requires a different, deeper cut with each repetition. Likewise, when one rings an alarm for the second time, say one second after the first ringing, one intends to warn those within earshot at that time, and not those who were in earshot one second ago.

(3) Agents are unable to maintain multiple intentions simultaneously without knowing in advance how to interleave the steps of the intentions' respective plans. This is a manifestation of the more general problem that, according to the definition of INTEND_1, if an agent does not know in advance every action leading up to the fulfillment of an intention before executing the actions, then the agent cannot satisfy the intention, even if the end result is achieved. The third problem makes clear the need for a definition of intention that corresponds to the weaker sense of the word "intention" without the connotation that the agent knows every detail of his plan to fulfill the intention. We refer to such intentions as *future-directed*. Agents may then maintain multiple weak intentions, and interleave partial plans toward achieving these intentions. Only in the strongest, most bona-fide sense of intention (*present-directed* intention) must the agent know in advance the entire action sequence supposed to bring about the intention.

Solutions to the Problems

The criticism regarding repeated actions suggests that intention toward action should entail commitment to a particular outcome. One could naively indicate this like so:

$$(\mathsf{WEAK\text{-}INTEND}_1\ x\ a) \equiv \exists p, (\mathsf{P\text{-}GOAL}\ x\ (\mathsf{DONE}\ x\ a; p?))$$

However, the syntax prohibits quantification over formulas like p.

We introduce another set of entities into the universe of discourse as surrogates for formulas. We define $J \subseteq \wp(T)$ corresponding to a pre-defined set of possible expected outcomes defined as sets of possible worlds. J, then, consists of sets of possible worlds (timelines). We introduce

1. $\langle$**ActionVariable**$\rangle$ ::= $a, a_1, a_2, \ldots, b, b_1, b_2, \ldots, e, e_1, e_2, \ldots$
2. $\langle$**AgentVariable**$\rangle$ ::= $x, x_1, x_2, \ldots, y, y_1, y_2, \ldots$
3. $\langle$**RegularVariable**$\rangle$ ::= $i, i_1, i_2, \ldots, j, j_1, j_2, \ldots$
4. $\langle$**JustificationVariable**$\rangle$ ::= $s, s_1, s_2, \ldots, t, t_1, t_2, \ldots$
5. $\langle$**Variable**$\rangle$ ::= $\langle$AgentVariable$\rangle$ | $\langle$ActionVariable$\rangle$ | $\langle$RegularVariable$\rangle$ | $\langle$JustificationVariable$\rangle$
6. $\langle$**Numeral**$\rangle$ ::= $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$
7. $\langle$**Predicate**$\rangle$ ::= $(\langle$PredicateSymbol$\rangle\langle$Variable$\rangle_1, \ldots, \langle$Variable$\rangle_n)$
8. $\langle$**PredicateSymbol**$\rangle$ ::= $Q, Q_1, Q_2, \ldots$
9. $\langle$**Wff**$\rangle$ ::= $\langle$Predicate$\rangle$ | $\neg\langle$Wff$\rangle$ | $\langle$Wff$\rangle \vee \langle$Wff$\rangle$ | $\exists\langle$Variable$\rangle\langle$Wff$\rangle$ | $\forall\langle$Variable$\rangle\langle$Wff$\rangle$ | $\langle$Variable$\rangle = \langle$Variable$\rangle$ | $\mathsf{HAPPENS}\langle$ActionExpression$\rangle$) | $(\mathsf{DONE}\langle$ActionExpression$\rangle)$ | $(\mathsf{AGT}\langle$AgentVariable$\rangle\langle$ActionVariable$\rangle)$ | $(\mathsf{BEL}\langle$AgentVariable$\rangle\langle$Wff$\rangle)$ | $(\mathsf{ALLOW}\langle$AgentVariable$\rangle\langle$Wff$\rangle)$ | $(\mathsf{CHOOSE}\langle$AgentVariable$\rangle\langle$Wff$\rangle)$ | $(\mathsf{ACCEPT}\langle$AgentVariable$\rangle\langle$Wff$\rangle)$ | $\langle$TimeProposition$\rangle$ | $\langle$ActionVariable$\rangle \leq \langle$ActionVariable$\rangle$
10. $\langle$**TimeProposition**$\rangle$ ::= $\langle$Numeral$\rangle$
11. $\langle$**ActionExpression**$\rangle$ ::= $\langle$ActionVariable$\rangle$ | $\langle$ActionExpression$\rangle;\langle$ActionExpression$\rangle$ | $\langle$Wff$\rangle$? | $\langle$JustificationVariable$\rangle$¿

Table 1: Syntax

$\langle JustificationVariable\rangle$s whose denotations are elements of J. An element $j \in J$ may be used as T in constructing a model. Intuitively, J specifies sets of "desired" formulas, namely those compatible with its elements. We introduce semantics of a test action for expectations instead of formulas. The "¿" test action is analogous to the "?" test action, except it succeeds when the current world is an element of the "justification", or possible-worlds-set specification of the desired end state.

In the definition of a model, U therefore, is redefined to include J, the set of justifications. D, which is necessary for the interpretation of predicates by Φ, is redefined as $D = \Theta \cup P \cup E^* \cup J$. The syntax of C&L's theory needs no changes except the introduction of the "¿" test action, and variables whose denotations are elements of J. The amended syntax is given in table 1. Table 2 gives the satisfaction conditions. The semantic rules are given relative to a model M, a $\sigma \in T$, an integer n, and a set v. This v is a set of bindings of variables to objects in D such that if $v \in [Variable \rightarrow D]$, then v^d_x is that function which yields d for x and is the same as v elsewhere. If a model has a certain world σ that satisfies a Wff w at a given time under a certain binding, we write $M, \sigma, v, n \models w$.

We now present a definition of weak intention, or per-

1. $M, \sigma, v, n \models Q(x_1, \ldots, x_k) \Leftrightarrow$
   $\langle v(x_1) \ldots v(x_k) \rangle \in \Phi[Q, \sigma, n]$
2. $M, \sigma, v, n \models \neg\alpha \Leftrightarrow$
   $M, \sigma, v, n \not\models \alpha$
3. $M, \sigma, v, n \models (\alpha \vee \beta) \Leftrightarrow$
   $M, \sigma, v, n \models \alpha$ or $M, \sigma, v, n \models \beta$
4. $M, \sigma, v, n \models \exists x, \alpha \Leftrightarrow$
   $M, \sigma, v^x_d, n \models \alpha$ for some d in D
5. $M, \sigma, v, n \models \forall x, \alpha \Leftrightarrow$
   $M, \sigma, v^x_d, n \models \alpha$ for every d in D
6. $M, \sigma, v, n \models (x_1 = x_2) \Leftrightarrow$
   $v(x_1) = v(x_2)$
7. $M, \sigma, v, n, \models \langle \text{TimeProposition} \rangle \Leftrightarrow$
   $v(\langle \text{TimeProposition} \rangle) = n$
8. $M, \sigma, v, n, \models (e_1 \leq e_2) \Leftrightarrow$
   $v(e_1)$ is an initial subsequence of $v(e_2)$
9. $M, \sigma, v, n, \models (\mathsf{AGT}\ x\ e) \Leftrightarrow$
   $AGT[v(e)] = v(x)$
10. $M, \sigma, v, n, \models (\mathsf{HAPPENS}\ a) \Leftrightarrow$
    $\exists m, m \geq n,$ such that $M, \sigma, v, n \| a \| m$
11. $M, \sigma, v, n, \models (\mathsf{DONE}\ a) \Leftrightarrow$
    $\exists m, m \leq n,$ such that $M, \sigma, v, m \| a \| n$
12. $M, \sigma, v, n \| e \| n + m \Leftrightarrow$
    $v(e) = e_1 e_2 \ldots e_m$ and $\sigma(n + i) = e_i, 1 \leq i \leq m$
13. $M, \sigma, v, n \| a; b \| m \Leftrightarrow$
    $\exists k, n \leq k \leq m,$ such that $M, \sigma, v, n \| a \| k$ and
    $M, \sigma, v, k \| b \| m$
14. $M, \sigma, v, n \| \alpha? \| n \Leftrightarrow$
    $M, \sigma, v, n \models \alpha$
15. $M, \sigma, v, n \| s_{\text{¿}} \| n \Leftrightarrow$
    $v(s) = j$ and $\sigma \in j \in J \subseteq \wp(T)$
16. $M, \sigma, v, n \models (\mathsf{BEL}\ x\ \alpha) \Leftrightarrow$
    $\forall \sigma^*$ such that $\langle \sigma, n \rangle B[v(x)]\sigma^*,\ M, \sigma^*, v, n \models \alpha$
17. $M, \sigma, v, n \models (\mathsf{ALLOW}\ x\ \alpha) \Leftrightarrow$
    $\exists \sigma^*$ such that $\langle \sigma, n \rangle B[v(x)]\sigma^*, M, \sigma^*, v, n \models \alpha$
18. $M, \sigma, v, n \models (\mathsf{CHOOSE}\ x\ \alpha) \Leftrightarrow$
    $\forall \sigma^*$ such that $\langle \sigma, n \rangle C[v(x)]\sigma^*, M, \sigma^*, v, n \models \alpha$
19. $M, \sigma, v, n \models (\mathsf{ACCEPT}\ x\ \alpha) \Leftrightarrow$
    $\exists \sigma^*$ such that $\langle \sigma, n \rangle C[v(x)]\sigma^*, M, \sigma^*, v, n \models \alpha$

Table 2: Semantics

sonal commitment. Such an intention, when directed toward action, is defined as a commitment to have done the action, bringing about a desired condition or state. Agents weakly intend that which they are committed to bringing about themselves, regardless of having a plan to do so. However, in the case of intention toward action, an agent will indeed have at least a partial plan, this being the action itself.

$$(\mathsf{WEAK\text{-}INTEND}_1\ x\ a) \equiv \exists s, (\mathsf{P\text{-}GOAL}\ x\ (\mathsf{DONE}\ x\ a; s_{\text{¿}}))$$

$$(\mathsf{WEAK\text{-}INTEND}_2\ x\ p) \equiv (\mathsf{P\text{-}GOAL}\ x\ \exists a, (\mathsf{DONE}\ x\ a; p?))$$

By requiring the agent to commit that the action will have a particular outcome, we enable contextualization of the action within larger plans. This logic is compatible with means-ends planning under which an action is a candidate for execution if its preconditions are met, and the postcondition of each action is either the precondition of the next action in the given plan or the end of the plan. We have not explored the compatibility of this logic with hierarchical task network planning.

In the stronger sense of the word, one has a plan to carry out what one intends. This stronger notion of intention is applicable to both intention toward propositions INTEND_2 and intention toward actions INTEND_1. Therefore, the new definition of intention consists of a weak intention and what the agent *thinks* is a sure-fire plan to presently bring it about.

$$(\mathsf{INTEND}'_1\ x\ a) \equiv \exists s, (\mathsf{P\text{-}GOAL}\ x\ (\mathsf{DONE}\ x\ a; s_{\text{¿}})) \wedge (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ a; s_{\text{¿}}))$$

$$(\mathsf{INTEND}'_2\ x\ p) \equiv \exists e, (\mathsf{WEAK\text{-}INTEND}_2\ x\ p) \wedge (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ e; p?))$$

In this very strong sense of intention, the agent does not allow for any futures under which the intention does not come true. The agent believes that other outcomes are ruled out by the inevitability of the present action. However, the agent may freely change his mind about this fact as conditions develop (this belief is not restrained by any BEFORE clause like that of P-GOAL), and thus readily update his intentions, which would nevertheless maintain a degree of consistency due to the agent's commitment.

We turn now to the matter of the theory's assumptions. As examined earlier, the *No infinite persistence* assumption produces undesirable results when combined with the definition of P-GOAL. To prevent a scenario where an agent's commitments must come about regardless of action, we need to ensure that agents to not perpetually procrastinate. Therefore, we also adopt the assumption of *No infinite deferral*, defined as

$$\models (\mathsf{P\text{-}GOAL}\ x\ p) \wedge \neg[\mathsf{BEFORE}\ (\mathsf{BEL}\ x\ \Box\neg p)\ \neg(\mathsf{CHOOSE}\ x\ (\mathsf{LATER}\ p))] \rightarrow \Diamond(\mathsf{INTEND}'_2\ x\ p)$$

This definition allows for the case where the agent realizes his commitment has been derailed by uncontrollable events in the world, such as the intervention of other agents.

Under these assumptions, as long as an agent is competent with respect to his belief that he is carrying out the intention, then the intention must come true. That is, if the agent x never believes $(\mathsf{HAPPENS}\ x\ e; p?)$ with e and p as in the definition of INTEND'_2 unless it is true, then the strong intention's plan will succeed. One may define a notion of capability whereby the conditions are satisfied for *No infinite deferral*, and the agent is competent with respect to the belief component of the strong intention that comes about.

Meeting the Desiderata for Intention

The revised assumptions and definitions complete the modifications needed to address all of the formal criticisms of Singh. C&L show that their theory meets many desiderata

motivated by (Bratman 1987). The revised theory satisfies these desiderata as well.

Intentions normally pose problems for the agent; the agent needs to determine a way to achieve them: In the case of intention toward action, the agent already plans to achieve the goal by performing the intended action. In the case of intention toward propositions, by the assumption of *No infinite deferral* and the definition of INTEND'_2 clearly an agent will try to come up with a plan once committed to doing something.

Intentions provide a "screen of admissibility" for adopting other intentions: If an agent intends b, as in $(\mathsf{INTEND}'_1\ b)$, and always believes that doing a forever prevents doing b, as in

$\Box(\mathsf{BEL}\ (\mathsf{HAPPENS}\ a) \rightarrow \Box\neg(\mathsf{HAPPENS}\ b))$,

then the agent cannot intend to do a before b in any sequence of actions. Suppose that the agent did $(\mathsf{INTEND}'_1\ x\ a; b)$. Then the agent would believe a would occur, forever preventing b. But the agent would also believe b would occur, a contradiction. Therefore, we may formally say

$$\begin{array}{l}\models\ \forall x, (\mathsf{INTEND}'_1\ x\ b) \\ \quad \wedge \Box(\mathsf{BEL}\ x\ [(\mathsf{DONE}\ x\ a)\ \rightarrow \\ \quad\quad \Box\neg(\mathsf{DONE}\ x\ b)])\ \rightarrow \\ \quad \neg(\mathsf{INTEND}'_1\ x\ a; b)\end{array}$$

Agents "track" the success of their attempts to achieve intentions: Agents maintain their intentions after failure, contingent upon being able to come up with a plan. Suppose it is the case that

$$\begin{array}{l}(\mathsf{DONE}\ x[(\mathsf{INTEND}'_1\ x\ a) \wedge \\ \quad (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ a; p?))]?; e; \neg p?)\end{array}$$

that is the agent's intended action did not occur just now when expected. Further suppose that

$$\begin{array}{l}(\mathsf{BEL}\ x\ \neg(\mathsf{DONE}\ x\ a; p?)) \wedge \\ \quad \neg(\mathsf{BEL}\ x\ \Box\neg(\mathsf{DONE}\ x\ a; p?))\end{array}$$

which means the agent is aware the intention did not succeed, but still believes it possible to succeed. By the BEFORE clause, it is obvious that the P-GOAL or weak intention component of the strong intention conjunct remains true. The agent at this point will be disposed to form a new belief about what action to take; presumably, he would like to try again. In order to try again, he must resolve to do so; the agent would adopt $(\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ a; p?))$. As we would expect, successful maintenance of strong intentions therefore depends on agents' ability to maintain plans. This case meets all the conditions for INTEND'_1.

If the agent intends action a, then the agent believes it is possible to do action a: By definition an agent cannot adopt intentions considered impossible (always not fulfilled). Since when adopting any commitment the agent must make a choice (i.e. adopt a goal), by the *realism* constraint, the agent allows that the intention could possibly succeed.

If the agent intends action a, sometimes the agent believes he will in fact do a: Under from the definition of $\mathsf{WEAK\text{-}INTEND}_1$, the agent has an persistent goal and therefore an achievement goal toward having done the action: $(\mathsf{A\text{-}GOAL}\ (\mathsf{DONE}\ a))$; hence the agent will choose that the intended act comes true later (by definition of A-GOAL) meaning $(\mathsf{CHOOSE}\ (\mathsf{LATER}\ (\mathsf{DONE}\ a)))$. Since C is serial, by the D axiom we have

$(\mathsf{ALLOW}\ (\mathsf{LATER}\ (\mathsf{DONE}\ a)))$,

meaning that the agent believes it possible that a occurs. This is not quite so strong as belief, but then again the agent has not made definitive plans to execute a.

On the other hand, in the case of strong intention under INTEND'_1, quite simply the agent believes he is doing the intended act.

If the agent intends action a, the agent does not believe he will never do a: This follows from the *realism* constraint.

Agents need not intend all the expected side-effects of their intentions: This follows (admittedly "for the wrong reasons") from the lack of consequential closure of the P-GOAL, discussed at length in (Cohen & Levesque 1990) and (Orilia 1996).

Dropping futile intentions: The final theorem in (Cohen & Levesque 1990), which appears at the end of section 7, states that "if an agent believes anyone else is truly going to achieve p, then either the agent does not intend to achieve p himself, or he does not believe p can be achieved only once. Contrapositively, if an agent intends to achieve p, and always believes p can be achieved only once, the agent cannot simultaneously believe someone else is going to achieve p"

$$\begin{array}{l}\models\ \forall x, \forall y,\ (y \neq x) \wedge \\ \quad (\mathsf{BEL}\ x\ \Diamond\exists e, (\mathsf{DONE}\ y\ \neg p?; e; p?))\ \rightarrow \\ \quad \neg(\mathsf{INTEND}'_2\ x\ p) \wedge \\ \quad \neg(\mathsf{BEL}\ x\ [\exists e, (\mathsf{DONE}\ y\ \neg p?; e; p?) \\ \quad\quad \rightarrow\ \Box\neg\exists e, (\mathsf{DONE}\ x\ \neg p?; e; p?)])\end{array}$$

This holds for the new definition as well. If the agent x were to believe y would achieve p, and intended to achieve p himself as well, then he would not believe that p could be achieved only once. If instead the agent x believed that y would achieve p and p could be achieved only once, then x could never adopt an intention to achieve p due to the fact that the requisite achievement goal could not hold: in all of x's belief accessible worlds, p is achieved only once, by y.

The Case Due to Chisholm: C&L rightly observe that their definition of INTEND_2 handles the following case, due to Chisholm (Chisholm 1966), paraphrased by C&L (page 248): "An agent intends to kill his uncle. On the way to his uncle's house, this intention causes him to become so agitated that he loses control of his car, and runs over a pedestrian, who happens to be his uncle. Although the uncle is dead, we would surely say that the action that the agent did was not what was intended." Under the definition of $\mathsf{WEAK\text{-}INTEND}_2$, which we characterize as personal commitment, an avunculicidal agent would fulfill his intention in this case. Thus, agents are able to fulfill their commitments (weak intentions) accidentally. However, an unexpected accident would generally not occur in any of the agent's allowed (belief accessible) worlds, so the agent could not have

a strong intention, goal, or even allowance for the occurrence.

Conclusion

C&L's theory of intention as a persistent goal has furnished a great foundation for new theory and application. Theories of joint intentions and theories of speech acts can both be built from its constructs, and Tambe and Jennings have provided us with creditable implementations. However, Singh makes disturbing observations about the original theory. In particular, agents' commitments can be automatically (and inappropriately) brought about, agents cannot intend the same action twice consecutively, and agents are unable to maintain and act on multiple intentions without planning in advance exactly how to interleave the steps of those intentions.

In the amended theory presented here, C&L's desiderata for intention remain fulfilled, and the criticisms of Singh have been addressed. As discussed, agents' commitments will never be brought about inappropriately (without action on the part of an agent). Furthermore, agents may intend actions of the same type repeatedly by expecting their actions to have distinct outcomes. Finally, agents may maintain multiple weak (future-directed) intentions and use these to form whatever strong (present-directed) intention is appropriate at a particular time, thus allowing agents to act opportunistically and interleave execution of different recipes.

The original theory and the revised theory place some useful restrictions on agent architectures. Clearly, agents should have a particular mental state corresponding to CHOOSE and a particular mental state corresponding to BEL, with relationships as discussed. Furthermore, agents should commit to certain goals, with the results that these goals persist through time. Also, Singh's criticism has revealed that agents should not intend actions for their own sake, but rather to have some effect. This comes as no surprise to those who study planning and means-ends reasoning. Furthermore, if the desired outcomes of intended actions are motivated by recipes, agents can easily be made to avoid over-commitment and under-commitment. Specifically, an agent may conclude that an intention is impossible when all recipes have been exhausted.

By capturing these notions, and implicitly tying agents' beliefs to their actions, the theory presented here gives us certain insights into rational agency. First of all, agents become more rational by being intentional, because they can persistently concentrate on goals and plans, and this facilitates means-ends reasoning. In addition, agents intend things to occur in the future with unspecified plans, but agents always know exactly what they are doing when they intentionally do it. These insights concur with certain conclusions of Searle and Bratman. The revised theory then, like the original, has philosophical appeal. Future work involving the theory should involve an explicit integration of belief and choice revision in the logic, which would strengthen the connection to planning theory. Finally, the notion captured here by justifications (sets of possible worlds) could also be captured by situations as in the theory of (Barwise 1989) which promises to move us beyond some of the inadequacies of possible worlds semantics.

References

Barwise, J. 1989. *The Situation in Logic*. Stanford, CA: CSLI.

Bratman, M. 1987. *Intentions, Plans, and Practical Reason*. Harvard University Press, Cambridge.

Chisholm, R. 1966. *Freedom and Determinism*. Random House, New York. chapter Freedom and action.

Cohen, P., and Levesque, H. 1990. Intention is choice with commitment. *Artificial Intelligence* 42:213–261.

Cohen, P., and Levesque, H. 1995. Communicative actions for artificial agents. In *First International Conference on Multi-agent Systems*, 65–72. AAAI Press, Monolo Park, CA.

Cohen, P.; Levesque, H.; and Nunes, J. 1990. On acting together. In *AAAI*.

Herzig, A., and Longin, D. 2004. C&l intention revisited. In *Proceedings of the Ninth International Conference on Knowledge Representation and Reasoning*, 527–535.

Hintikka, J. 1962. *Knowledge and Belief*. Cornell University Press.

Jennings, N. 1995. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *Artificial Intelligence* 75(2).

Orilia, F. 1996. Side-effects and Cohen's and Levesque's theory of intention as a persistent goal. *From the Logical Point of View (ICE, Institute of Philosophy, Praga)* 3(94):1–19.

Rao, A., and Georgeff, M. 1991. Modeling rational agents within a BDI-architecture. In *Proceedings of Knowledge Representation and Reasoning*, 473–484.

Searle, J. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Singh, M., and Asher, N. 1993. A logic of intentions and beliefs. *Journal of Philosophical Logic* 22(5):513–554.

Singh, M. 1992. A critical examination of the Cohen-Levesque theory of intentions. In *Proceedings of the European Conference on Artificial Intelligence*.

Tambe, M. 1997. Towards flexible teamwork. *Journal of Artificial Intelligence Research* 7:83–124.

van der Hoek, W.; van Linder, B.; and Meyer, J.-J. C. 1997. An integrated modal approach to rational agents. In *Proc. 2nd AISB Workshop on Practical Reasoning and Rationality.*, 123–159.

Woolridge, M. 2002. *An Introduction to Multiagent Systems*. John Wiley and Sons, New York.