A Model of Intention with (Un)Conditional Commitments*

Dongmo Zhang

Intelligent Systems Laboratory,
University of Western Sydney, Australia
d.zhang@uws.edu.au

Abstract. This paper proposes a model of intention with conditional/ unconditional commitments based on Cohen and Levesque's (C&L's for short) framework of intention. We first examine C&L's framework with a well-known philosophical puzzle, the Toxin Puzzle, and point out its insufficiency in modelling conditional and unconditional commitments. We then propose a model theory of a specific modal logic with modalities representing the typical mental attributes as well as action feasibility and realisibility. Instead of defining intention as a persistent goal, we define an intention as a conditional/unconditional commitment made for achieving a goal that is believed to be achievable. Finally we check our framework with the Toxin Puzzle and show our solution to the puzzle.

1 Introduction

With the motivation of modelling autonomous agents and multi-agent systems, a great number of logical frameworks, just to name a few, Cohen and Levesque's formalism of intention [1], Rao and Georgeff's BDI logic [2], and Meyer et al.s KARO logic [3], were proposed in the past two decades, attempting to capture the rationality of human agency and imitate rational behaviour of human being for implementation of software agents.

Most of the early works on agent modelling in the AI literature were stimulated by philosophical investigations on human mental states, most significantly Bratman's theory of intention [4]. It has been widely accepted that human rational behaviour is highly effected, if not determined, by their mental attitudes, such as knowledge, belief, desire and intention. However, formalising each of the mental attitudes and their relationships have been proved to be hard and tend to be complicated. This is because any separation of these attitudes leads to insufficiency of explanation in human rational behaviour. The central of these attitudes is human intention, which is in general intertwined with other attitudes, such as belief, desires and commitments, involving reasoning about actions and time [1, 4]. Any formal analysis of intention must explicate the relationships among those facets of mental state. This explains why the existing logics of intention are highly complicated [1, 2, 5, 3, 6–10]. Even worse, there is a theoretical boundary that prevents us from putting all the facets into a single logic system. Schmidt and Tishkovsky showed that if we combine propositional dynamic logic, used for specifying actions and time,

* This research was partially supported by the Australian Research Council through Discovery Project DP0988750.

P. Anthony, M. Ishizuka, and D. Lukose (Eds.): PRICAI 2012, LNAI 7458, pp. 698–709, 2012.

with doxastic modal logics, used for specifying knowledge and beliefs, the outcome logic system collapses to propositional dynamic logic if we admit substitution (a fundamental rule for using axiom schemata) [11]. Certain compromise or isolation of concepts has to be made in order to get a logic of intention with manageable complexity.

Another tricky part of modelling intention is its intimate connection to the concept of commitment. On the one hand, an agent intending to do an action would mean that the agent is committing to the action to be done now or in the future. On the other hand, an agent may continuously weigh his competing goals and choose one that is most desirable to achieve, which means that the agent does not have to bond to his commitments [4]. To deal with such a dilemma, Cohen and Levesque model intention as a kind of persistent goal: *an agent intending to do an action will form a persistent goal with which the agent acts until either the action is done or is believed that it can never be done* [1]. However, as we will illustrate in the next section, such a definition of intention is insufficient to capture the concept of intention with conditional/unconditional commitments, which often occur in the real world.

In this paper, we propose a model of intention based on C&L's framework but reformulate the concept of intention and commitment. Instead of defining intention as a persistent goal, we define an intention as a conditional/unconditional commitment made for achieving a goal that is believed to be achievable. We check our framework by providing a solution to a well-known philosophical problem, the Toxin Puzzle [12].

The paper is organised as the following. In the next section, we briefly recall the basic concepts of C&L's framework and using their language to describe the Toxin Puzzle. Section 3 presents our model of intention and discuss its properties. Section 4 gives our solution to the Toxin Puzzle before we conclude the paper in Section 5. Due to space limit, we omit all the proofs and some technique lemmas.

2 C&L's Model of Intention

C&L's model combines doxastic logics and segments of dynamic logic in model-theoretic fashion. Agent's beliefs and goals are represented by modalities, BEL and GOAL, and are specified via possible world semantics. It is assumed that BEL satisfies KD45 (thus its accessibility relation is Euclidean, transitive and serial) and that GOAL satisfies KD (its accessibility relation is then serial). Actions are represented by action expressions which compound primitive events via program connectives: ; (*sequential*), — (*nondeterministic choice*), ? (*test*) and * (*iteration*). Time is modeled in linear structure with infinite future and infinite past. Each course of event represents a time unit. Therefore a time period is identical to a sequence of events that happen during the period. A possible world is then an infinite sequence of events with two open ends, representing the events happened in the past and the events happen now or in the future.

Two temporal operators HAPPENS and DONE are defined to indicate the actions that are about to happen and the actions that have just been done, respectively. Other temporal operators can be defined accordingly:

$$\Diamond\varphi =_{def} \exists e\ \mathtt{HAPPENS}\ e;\varphi?$$
$$\Box\varphi =_{bef} \neg\Diamond\neg\varphi$$

$$\texttt{LATER}\ \varphi =_{def} \neg\varphi \wedge \Diamond\varphi$$
$$\texttt{BEFORE}\ \varphi\ \psi =_{def} \forall e(\texttt{HAPPENS}\ e;\psi? \rightarrow \exists e'(e' \leq e \wedge \texttt{HAPPENS}\ e';\varphi?))$$

where $e \leq e'$ means that e' happens before e.

Based on the concepts, they introduce the concept of persistent goals to capture "one grade" of commitments. Intuitively, a goal of an agent is a *persistent goal* (P-GOAL) if the goal will not give up once it is established unless the agent believes that it has been achieved or it can never be achievable.

$$\begin{aligned}\texttt{P-GOAL}\ \varphi =_{def}\ & (\texttt{GOAL LATER}\ \varphi) \wedge (\texttt{BEL}\ \neg\varphi) \wedge \\ & (\texttt{BEFORE}\ (\texttt{BEL}\ \varphi \vee \texttt{BEL}\ \Box\neg\varphi)\ \neg(\texttt{GOAL LATER}\ \varphi))\end{aligned} \quad (1)$$

With the concept of commitment, intention can be simply defined as a persistent goal to perform an action:

$$\texttt{INTEND}\ \alpha =_{def} \texttt{P-GOAL}\ (\texttt{DONE}\ ([\texttt{BEL}\ (\texttt{HAPPENS}\ \alpha)]?;\alpha)) \quad (2)$$

where α is an action expression. Intuitively, an agent intending to do an action is a commitment (persistent goal) to have done this action if he believes the action is executable.

The INTEND operator can be overloaded to take a proposition as argument:

$$\begin{aligned}\texttt{INTEND}\ \varphi =_{def}\ & \texttt{P-GOAL}\ \exists e(\texttt{DONE}\ [(\texttt{BEL}\ \exists e'(\texttt{HAPPENS}\ e';\varphi?))\wedge \\ & \neg(\texttt{GOAL}\neg\texttt{HAPPENS}(e;\varphi?))]?;e;\varphi?)\end{aligned} \quad (3)$$

where φ is a proposition. It reads that, to intend to bring about φ, an agent sets up a persistent goal that is to find a sequence of events after doing which himself, φ holds. For a full understanding of C&L's theory of intention, including comments, criticisms and follow-ups, the reader is referred to Cohen and Levesque (1990) [1], Singh (1992) [13], Rao and Georgeff [14], Hoek and Wooldridge [6], and etc.

2.1 The Toxin Puzzle

Gregory Kavka challenges theories of intention by inventing the following thought experiment, known as *the Toxin Puzzle* [12]:

> *Suppose that a billionaire offers you on Monday a million dollars if on Tuesday you intend to drink a certain toxin on Wednesday. It is common knowledge that drinking this toxin will make you very sick for a day but will not have a lasting effect. If you do so intend on Tuesday, the million dollars will be irreversibly transferred to your bank account; this will happen whether or not you actually go ahead and drink the toxin on Wednesday.* [1]

Although the puzzle sounds a bit unrealistic, we can find large amount of similar scenarios in the real-world, such as government funded research grants, scholarships, and high-risk investments.

The Toxin Puzzle has induced tremendous debates in the philosophical literature [15–18]. It challenges the theories of intentions from two aspects: *the nature of intentions* and *the rationality of intentions*. On the one hand, if intentions were inner

[1] This simplified version of the Toxin Puzzle is cited from [4] p.101.

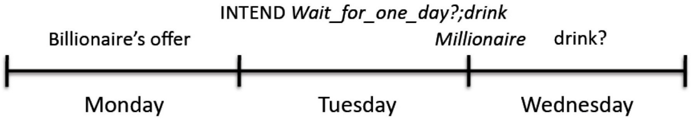
performances or self-directed commands, an agent would have no trouble to form an intention. On the other hand, when reasons for intending and reasons for acting diverge, as they do in the puzzle, confusion often reigns: either giving way to rational action, rational intention, or aspects of the agent's own rationality [12].

2.2 Formalising the Toxin Puzzle in C&L's Language

We describe the Toxin Puzzle in C&L's language. Let $wait\_a\_day$ denote the action "wait for one day" and $drink$ the action "drink the toxin". We use $millionaire$ to denote the fact "given a million dollars". Then the billionaire's offer, which was made on Monday, can be expressed as:

$$(\texttt{INTEND}\ wait\_a\_day; drink) \rightarrow millionaire \tag{4}$$

Assume that an agent had a goal to be a millionaire thus he formed an intention $\texttt{INTEND}\ wait\_a\_day; drink$ on Tuesday. As the billionaire had promised, the agent was given one million dollars and became a millionaire on the same day. The question is: *Whether the agent will drink the toxin on the next day as the billionaire expected?*



According to C&L's definition (Equation 2), $\texttt{INTEND}\ wait\_a\_day; drink$ represents the following persistent goal:

$$\texttt{P-GOAL}[\texttt{DONE}((\texttt{BEL}(\texttt{HAPPENS}\ wait\_a\_day; drink))?; wait\_a\_day; drink)] \tag{5}$$

which says that the agent has a persistent goal to drink the toxin on the next day as long as he believes it can happen in due course. Obviously the agent does not have to believe on Tuesday that the action will happen because an agent is allowed to withdraw his commitment if there is other competing goals and he does not know whether there is any competing goals on next day. Therefore $(\texttt{BEL}(\texttt{HAPPENS}\ wait\_a\_day; drink))?$ fails on Tuesday. Then the persistence goal is carried over to Wednesday. However, execution of the action $wait\_a\_day; drink$ does not make sense. No matter it is executed or not, the outcome is not what is expected. In fact, since a persistent goal does not require when the goal is achieved, any time-restricted intention cannot be rightly characterised using persistent goals.

As the matter of fact, there is no standard answer to the Toxin Puzzle. The puzzle is used to illustrate the divergence of reasons for intending and reasons for acting [12]. The reason for the agent to form the intention is to be a millionaire but the reasons to act, if it happens, is due to his commitment. Both sides (the billionaire and the agent) agree on the first reason but diverge on the second reason if the agent does not drink the toxin. The billionaire expects the agent to form a *unconditional commitment* while the agent might have formed a *conditional commitment*. Nevertheless, none of conditional or unconditional commitments can be properly specified in C&L's model.

3 The Model of Intention

In this section we propose a model of intention with condition/unconditional commitments. The model is built our on C&L's framework. We do so because C&L's framework is well-known in AI community and relatively simple. We will introduce a formal language that consists of two doxastic modalities BEL (for belief) and DES (for desire), and three action modalities ATH, FEASIBLE and COMMITTED. ATH stands for "*About To Happen*". Similar to C&L's framework, we shall present our model in possible-words semantics with linear-time model. Each world represents a linear sequence of events. However, our time is future-directed without referring to the past. Another important difference from the C&L model is that, for this paper, we assume that any events that happens are due to the actions performed by a single agent. The beliefs, desires, intentions are the mental states of the same agent. Commitments are meant self-commitment. Any changes of worlds are resulted from the agent's actions. We admit that such a simplification would lead to the insufficiency of modeling social behaviours of rational agency. However, such a sacrifice of completeness can be worth if a simple theory of intention can help us to understand the concept deeper.

3.1 Syntax

Consider a propositional language with modalities BEL, DES, ATH, FEASIBLE, COMMITTED, and regular action expressions. Let $\varPhi$ be a countable set of atomic formulas, denoted by $\phi, \phi_1, \phi_2, \cdots$, and Act_P be a countable set of primitive events or actions, denoted by $a, b, \cdots$. The set of all formulas generated from $\varPhi$, Act_P, and the modalities, are denoted by Fma. Its members are denoted by φ, ψ, φ_1, $\cdots$, etc. Act is the set of all compound actions through action connectives ; , $\cup$, $*$, with typical members denoted by $\alpha, \beta, \cdots$. We use $\top$ and $\bot$ to denote the propositional constants "**true**" and "**false**", respectively.

Formally, formulas ($\varphi \in Fma$) and actions ($\alpha \in Act$) are defined by the following BNF rules:

$\varphi ::= \phi \mid \neg\varphi \mid \varphi_1 \rightarrow \varphi_2 \mid \mathtt{BEL}\varphi \mid \mathtt{DES}\varphi \mid \mathtt{ATH}\,\alpha \mid \mathtt{FEASIBLE}\,\alpha \mid \mathtt{COMMITTED}\,\alpha$

$\alpha ::= a \mid \alpha_1;\alpha_2 \mid \alpha_1 \cup \alpha_2 \mid \alpha^* \mid \varphi?$

where $\phi \in \varPhi$ and $a \in Act_P$.

For any formula $\varphi \in Fma$ and action $\alpha \in Act$, we write α_φ to represent another action that is yielded from α by replacing each primitive action a in α with $\varphi?; a$. It is easy to give an inductive definition of α_φ based on α's structure so omitted.

Given an action α, we define all the computation sequences of α by induction on the structure of α as follows (see [19]):

$CS(a) =_{def} \{a\}$, where $a \in Act_P$, an atomic program

$CS(\varphi?) =_{def} \{\varphi?\}$

$CS(\alpha;\beta) =_{def} \{\sigma';\sigma'' : \sigma' \in CS(\alpha), \sigma'' \in CS(\beta)\}$

$CS(\alpha \cup \beta) =_{def} CS(\alpha) \cup CS(\beta)$

$CS(\alpha^*) =_{def} \bigcup_{n \geq 0} CS(\alpha^n)$.

$CS = \bigcup_{\alpha \in Act} CS(\alpha)$.

where $\alpha^0 = \top?$ and $\alpha^{n+1} = \alpha;\alpha^n$.

We remark that with non-deterministic choice and iteration, we can express "very rough" plans. Suppose that we have only finite number of primitive actions $a_1, a_2, \cdots,$ a_n. Then the action **any** $= (a_1 \cup \cdots \cup a_n)^*$ can be any action the agent can perform except test actions. Therefore, whenever we talk about an action description, it actual mean a "rough plan" or a "partial plan" in Bratman's terminology [4] because the execution of such a auction relies on a high-level planner to find an executable computation sequence of primitive events that implements the action.

3.2 Semantic Model

A model M is a structure $(E, W, \bar{\sigma}, B, D, C, V)$, where E is a set of primitive event types[2]. $W \subseteq [\mathbb{N} \to E]$ is a set of possible courses of events or worlds specified as a function from the natural numbers, representing time points, to elements of E, $\bar{\sigma} \in W$ is a special world (the real world) representing the actual events that have happened or will happen, $B \subseteq W \times \mathbb{N} \times W$ is the accessibility relation for belief modality, $D \subseteq W \times \mathbb{N} \times W$ is the accessibility relation for desire modality, $C : W \times \mathbb{N} \to \wp(CS)$ is a function that assign a set of computation sequences to each world at each time[3], and $V : W \times \mathbb{N} \to \wp(\Phi)$ is the valuation function of the atomic formulae.

Definition 1. Given a model M, let $\sigma \in W$ be any possible world and $n \in \mathbb{N}$ a natural number. We define the satisfaction relation $M, \sigma, n \models \varphi$ as follows:

1. $M, \sigma, n \models p$ iff $p \in V(\sigma, n)$, where $p \in \Phi$.
2. $M, \sigma, n \models \neg\varphi$ iff $M, \sigma, n \not\models \varphi$.
3. $M, \sigma, n \models \varphi \to \psi$ iff $M, \sigma, n \models \varphi$ implies $M, \sigma, n \models \psi$.
4. $M, \sigma, n \models \mathtt{BEL}\ \varphi$ iff for all σ' such that $(\sigma, n, \sigma') \in B$, $M, \sigma', n \models \varphi$.
5. $M, \sigma, n \models \mathtt{DES}\ \varphi$ iff for all σ' such that $(\sigma, n, \sigma') \in D$, $M, \sigma', n \models \varphi$.
6. $M, \sigma, n \models \mathtt{ATH}\ \alpha$ iff for all $i \leq n$, $\sigma(i) = \bar{\sigma}(i)$ and $\exists m, m \geq n$ such that $M, \bar{\sigma}, n[\![\alpha]\!]m$.
7. $M, \sigma, n \models \mathtt{FEASIBLE}\ \alpha$ iff there exists σ' such that for all $i \leq n$, $\sigma(i) = \sigma'(i)$ and $\exists m, m \geq n$ such that $M, \sigma', n[\![\alpha]\!]m$.
8. $M, \sigma, n \models \mathtt{COMMITTED}\ \alpha$ iff $CS(\alpha) \subseteq C(\sigma, n)$.
9. $M, \sigma, n[\![a]\!]n+1$ iff $a = \sigma(n+1)$, where a is a primitive action.
10. $M, \sigma, n[\![\alpha \cup \beta]\!]m$ iff $M, \sigma, n[\![\alpha]\!]m$ or $M, \sigma, n[\![\beta]\!]m$.
11. $M, \sigma, n[\![\alpha; \beta]\!]m$ iff $\exists k, n \leq k \leq m$, such that $M, \sigma, n[\![\alpha]\!]k$ and $M, \sigma, k[\![\beta]\!]m$.
12. $M, \sigma, n[\![\varphi?]\!]m$ iff $M, \sigma, n \models \varphi$.
13. $M, \sigma, n[\![\alpha^*]\!]m$ iff $\exists n_1, \cdots, n_k$, where $n_1 = n$ and $n_k = m$ and for every i such that $1 \leq i \leq k$, $M, \sigma, n_i[\![\alpha]\!]n_{i+1}$.

A formula φ is satisfiable if there is at least one model M, world σ and index n such that $M, \sigma, n \models \varphi$. A formula φ is valid iff for every model M, world σ, and index n, $M, \sigma, n \models \varphi$.

It is not hard to find that our semantic model is a variation of C&L's model. The semantics for BEL is exactly the same as C&L's. The one for DES are similar to C&L's

[2] For simplicity, we will identify each primitive event symbol with its type. Equivalently, we assume that there is a unique one to one mapping from Act_P to E.

[3] $C(\sigma, n)$ can be interpreted as the searching space of the event sequences for planning purposes.

semantics for GOAL operator but with less constraints (see more details in next subsection). We shall introduce GOAL as a composite concept defined by BEL and DES.

ATH α represents the statement that action α is about to happen in the real world ($\bar{\sigma}$). Note that it is different from C&L's HAPPENS for its truth value relies on the real world $\bar{\sigma}$, which is unique to a model.

FEASIBLE α represents the feasibility of action α performed by the agent under consideration. That α is feasible in a world σ at time n means that there is a possible world σ' which has the same history as σ but could branch at time n such that α is executable in the world σ'.

The operator COMMITTED captures the key feature of agent commitments. $CS(\alpha)$ lists all the alternative computation paths of action α. The set $C(\sigma, n)$ collects all the computation sequences the agent has committed to at each time n and each world σ. $CS(\alpha) \subseteq C(\sigma, n)$ means that no matter which path the agent takes to execute α, the path has been committed.

3.3 Constraints and Properties

In this subsection, we introduce a few constraints on the semantic conditions for BEL, DES and COMMITTED. Similar to C&L's model, we assume that B is transitive, serial and Euclidean. Therefore BEL is a doxastic modality that satisfies KD45.

We assume that DES is a modal operator that satisfies K and define goal as a composite concept from belief and desire:

Definition 2. GOAL $\varphi =_{def}$ DES $\varphi \wedge \neg$BEL φ.

In words, φ is a goal of an agent if the agent desires or wishes φ to be true and does not believe it is true now. Assume that you have a goal to be a millionaire. It means that you desire to be a millionaire but you believe that currently you are not a millionaire. The idea is inspirited by Meyer *et al.*'s KARO logic[4].

The semantic condition for COMMITTED is much more complicated. Intuitively, the accessibility relation $C(\sigma, n)$ collects all the computation sequences the agent committed to at time n in world σ. The agent will choose a computation sequence from this collection to execute. We assume that C satisfies the following conditions:

(c1) *If $\delta \in C(\sigma, n)$, then for any initial subsequence δ' of δ, $\delta' \in C(\sigma, n)$.*
In other words, $C(\sigma, n)$ is closed under initial subsequences.

(c2) *If $\varphi?; \delta \in C(\sigma, n)$, then $\delta \in C(\sigma, n)$.*
Similar to C&L's model, we assume that a test action does not take time. Thus tests have to be carried out in conjunction with one primitive event.

(c3) *For any sequences $\delta \in C(\sigma, n)$ and any world $\sigma' \in W$, if for all $i \leq n$ $\sigma(i) = \sigma'(i)$ and $\delta = \sigma(n); \delta'$, then $\delta' \in C(\sigma', n+1)$.*
In other words, if one committed to performing a sequence of actions, then he should carry over the commitment (the remaining computation sequences) to the next state of any possible world he might move to once he fulfils the initial primitive action in the sequence. This conditions captures the phenomenon of commitment persistence.

[4] In [3], Meyer et al. recognise beliefs and wishes as the primitive mental attributes instead of belief-desire. We shall not discuss the philosophical differences between wish and desire.

Example 1. *Given an action* $\alpha = (a \cup b); c; (p?; d \cup (\neg p); e)$. *Then* α *has the following computation sequences:* $a; c; p?; d$, $a; c; (\neg p)?; e$, $b; c; p?; d$, *and* $b; c; (\neg p)?; e$. *Assume that an agent has committed to performing* α *at time* 0 *and makes no other commitments further on. Firstly, the agent takes action* a. *Thus he has to carry over the committed sequences* $c; p?; d$ *and* $c; (\neg p)?; e$ *but drop the other two. After done* c*, if* p *is true, he shall continues with his committed action* d*; otherwise, he shall do* e. *Assume the latter case holds, the actual execution sequence will be* $a; c; (\neg p)?; e$*, which fulfils his commitment.*

According to our semantics, the following property is easy to verify,

Proposition 1. $\models \mathtt{ATH}\, \alpha \rightarrow \mathtt{FEASIBLE}\, \alpha$.

It shows that any action that is about to happen must be feasible for the agent to execute.

Proposition 2. *Properties of* COMMITTED *operator:*

1. $\models \mathtt{COMMITTED}(\alpha; \beta) \rightarrow [\mathtt{COMMITTED}\, \alpha \ \wedge (\mathtt{ATH}\, \alpha \rightarrow \mathtt{ATH}\, (\alpha; (\mathtt{COMMITTED}\, \beta)?)]$.
2. $\models \mathtt{COMMITTED}(\alpha \cup \beta) \rightarrow \mathtt{COMMITTED}(\alpha) \wedge \mathtt{COMMITTED}(\beta)$.
3. $\models \mathtt{COMMITTED}(\alpha^*) \rightarrow [\mathtt{COMMITTED}(\alpha) \ \wedge (\mathtt{ATH}\, \alpha \rightarrow \mathtt{ATH}\, (\alpha; (\mathtt{COMMITTED}\, \alpha^*)?)]$.

These statements shows that a commitment is carried over along the path of an execution.

3.4 Implementation Assumption and Conditional Commitments

Suppose that an agent has committed to performing an action α. Doing this action is also feasible for him. Then we should expect that the action is about to happen. This ideas can be described as the following assumption.

Implementation Assumption (IA) $\models (\mathtt{FEASIBLE}\, \alpha \wedge \mathtt{COMMITTED}\, \alpha) \rightarrow \mathtt{ATH}\, \alpha$.

Figure 1 illustrates the relation between commitment, feasibility and actual execution. Assume that an agent has committed to do action α which has five possible computation sequences a, b, c, d, e. The current state is at node d1. Among all the computation sequences, the paths a, c, e are not feasible (marked in red). The agent is actual taking

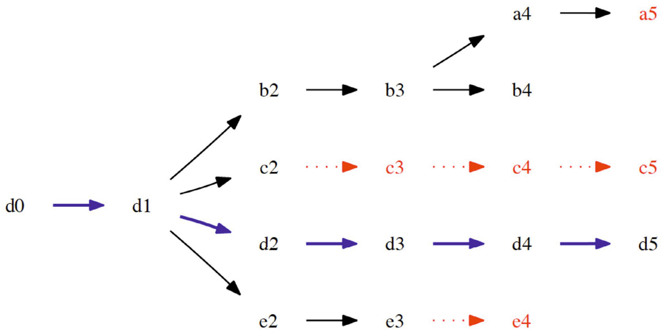


Fig. 1. Promise, feasibility and execution

path d (marked in bold and blue). Note that Implementation Assumption assumes that an agent has the ability of choosing a committed path from all feasible computation paths and execute the path.

This assumption is similar to C&L's assumption on persistent goal: *if a goal is not achieved, the agent will keep the goal until he believes that it is not achievable.* However, for action commitments, the assumption sounds a bit strong because if the commitment of performing an action was to achieve a goal and the goal has achieved or is believed not achivable, the commitment for performing the action might be dropped if the commitment is conditional. For example, assume that I intend to go to the airport tomorrow to pick my friend up from the airport. So I commit myself to going the airport tomorrow. However, if the friend is not coming tomorrow, the commitment to going the airport should be dropped. We call such a commitment a *conditional commitment.* I committed to going airport is because I wanted to pick my friend up. Without such a demand, I would not make the commitment. We define conditional commitment as follows:

Definition 3. *For any formula φ and action α,*
$\mathtt{COMMITTED}_{\varphi}\ \alpha =_{def} \mathtt{COMMITTED}\ \alpha_{\varphi}$

$\mathtt{COMMITTED}_{\varphi}\ \alpha$ means "commit to α under condition φ". Recall that α_{φ} is generated from α by replacing all primitive action a with $\varphi?; a$. The condition φ is checked all the way during the execution of α. Note that we can view $\mathtt{COMMITTED}\ \alpha$ as $\mathtt{COMMITTED}_{\top}\ \alpha$. Therefore the original COMMITTED operator can be treated as unconditional commitment. In this sense, **IA** is not strong.

3.5 Definition of Intentions

C&L consider intending an action and intending a proposition as separate concepts. As many philosophers argue, "[i]t is (logically) impossible to perform an intentional action without some appropriate reason." ([20] p264). Meanwhile we also think it is not a valid intention of achieving a goal without refer to a (rough) plan to achieve the goal. In this paper, we define an intention as a binary operator with the arguments: the intended action and the reason for doing the action. Similar idea also appears in [5, 3, 7]. Different from C&L's definition again, we treat conditional intentions and unconditional intentions as two different objects. We will give separate definition for each of them: *unconditional intention* and *conditional intention*. We start with the simply one first.

Definition 4. (Unconditional Intention)
$\mathtt{INTEND}_1\ \alpha\ \varphi =_{def} \mathtt{GOAL}\ \varphi \ \wedge\ \mathtt{BEL}(\mathtt{FEASIBLE}\ \alpha; \varphi?) \wedge \mathtt{COMMITTED}\ \alpha.$

Intuitively, I intend to do α to achieve a goal φ iff the following conditions are satisfied:

- φ is one of my goals;
- I believe that it is feasible to bring about φ via conducting α.
- I commit to performing α.

Note that for unconditional intention, any committed action has to been fulfilled no matter whether the goal has achieved, in which case the intention no longer exists. This is the intention the billionaire understood.

Definition 5. (Conditional Intention)
$\texttt{INTEND}_2\ \alpha\ \varphi =_{def} \texttt{GOAL}\ \varphi\ \wedge\ \texttt{BEL}(\texttt{FEASIBLE}\ \alpha; \varphi?) \wedge \texttt{COMMITTED}_{\texttt{DEC}\varphi}\ \alpha.$

The only difference between unconditional intention and conditional intention is the form of commitments. A conditional intention commits to perform the intended action as long as the goal is still desirable.

Example 2. Consider the example of chopping a tree (see Cohen and Levesque [1]). The intention of knocking down a tree can be expressed as follows:

$$\texttt{INTEND}_2\ (chop)^*\ (tree\_down)$$

To form the intention, it is required:

1. Knocking down the tree is my goal;
2. I believes that if I chop the tree repeatedly, the tree can be knocked down eventually (I may not know how many chops are actually needed).
3. I commit to doing the action as long as I still desire to knock down the tree (I may stop without knocking down the tree).

Note that our representation of this example using the concept of conditional intention is much clearer and simpler than C&L's version, thanks the specific definition of α_φ (Section 3.1).

4 Solution to the Toxin Puzzle

Now we are ready to present our solution to the Toxin Puzzle. Assume that the agent forms a unconditional intention on Tuesday as follows:

$$\texttt{INTEND}_1\ (wait\_a\_day; drink)\ millionaire \tag{6}$$

which means

1. The agent has a goal to be a millionaire, i.e., $\texttt{GOAL}\ millionaire$.
2. The agent believes that $wait\_a\_day; drink$ can bring about the goal, i.e., $\texttt{BEL}[\texttt{FEASIBLE}(wait\_a\_day; drink; millionaire?)]$.
3. The agent commits to doing the action $wait\_a\_day; drink$ unconditionally, i.e., $\texttt{COMMITTED}(wait\_a\_day; drink)$.

The billionaire's offer can be represented by the following statement:

$$[\texttt{INTEND}_1\ (wait\_a\_day; drink)\ \varphi] \to millionaire \tag{7}$$

where φ can be any reason as long as the billionaire can accept. Since the intention has been formed, the agent will receive a payment on Tuesday. The agent's goal is then achieved and the intention no longer exists. However, he has commit to drinking the toxin on next day, i.e., $\texttt{COMMITTED}(wait\_a\_day; drink)$. By Proposition 2, after done the action $wait\_a\_day$, the agent needs to keep the rest of commitment, i.e., $\texttt{COMMITTED}(drink)$. According to **IA**, as long as drinking the toxin is feasible to him, he has to make it happen on Wednesday, i.e., $\mathbf{ATH}(drink)$. This explains why the billionaire does not care too much about what is the goal of the agent to form the intention.

Assume instead that the agent forms the following conditional intention on Tuesday:

$$\texttt{INTEND}_2\ (wait\_a\_day; drink)\ millionaire \tag{8}$$

and, assume (but not too sure!) that the billionaire agrees on the intention and pay the agent one million dollars. The agent then have no goal to be a millionaire. However, whether the agent keeps the commitment of drinking the toxin will depend on whether he still keep the desire to be a millionaire. Either way can be rational. As main philosophers have pointed out, the reason for an agent to do an action, here is a million, can be different from the reason for the agent to drink. Whether the agent should keep the commitment does not necessarily reply on whether the agent has been a millionaire but on the agent's desire.

5 Conclusion

We have proposed a formal model of intention based on Cohen and Levesque's framework. We have shown that C&L's account of intention as persistent goal is inadequate to capture the key features of commitments of rational agency. We have argued that commitment plays a crucial role in intentional reasoning. With a unconditional commitment, the agent picks up a feasible computation path from a partial plan (specified by the intended action) to achieve his goal. The agent executes the actions alone the path and carries over unfulfilled actions to next state, which creates reason-giving force on the agent to execute his intended actions. With a conditional commitment, the condition is checked before executing any primitive actions. Both the accounts of commitment provide rational explanations to the Toxin Puzzle.

Conditional intention has been a traditional topic in philosophy for a long history even though it has not been well-investigated in the AI literature [21]. It has been widely accepted that a conditional intention and an unconditional intention refer to different objects and should be viewed as different intentions. In this paper, we treat the condition of a conditional intention as the way to determine whether the committed action should be carry out or stop while we treat the unconditional intention as an extreme case of conditional intention (blind commitment). More importantly, for a conditional intention, we do not assume that the commitment of the intention is automatically dropped if the intended goal has achieved. Whether an agent drops his commitments or not is determined by his desires.

A few issues are excluded from the current work. First, this work deals with single agent intentional reasoning. Commitment is meant self-commitment. However, the consideration of intention in multia-gent environment will give rise of issues on social commitments. In such a case, game-theoretic approach might be needed to model the social behaviour of rational agents. Secondly, we consider belief, desire and commitment as the primacy of intention. Changes of these mental states are left unspecified. The investigation of the dynamics of belief, desire and commitment will certainly provide a better understanding of human intention. Finally, the current work is based on propositional logic. Extending the framework to first-order language will allow us to define more temporal operator, such as $\Diamond$, $\Box$, $achievable$ and etc.

References

1. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artif. Intell. 42(2-3), 213–261 (1990)
2. Rao, A.S., Georgeff, M.P.: Modelling rational agents within BDI architecture. In: Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning, KR 1991, pp. 473–484 (1991)
3. Meyer, J.J.C., van der Hoek, W., van Linder, B.: A logical approach to the dynamics of commitments. Artif. Intell. 113(1-2), 1–40 (1999)
4. Bratman, M.E.: Intentions, Plans, and Practical Reason. Harvard University Press (1987)
5. van Linder, B., van der Hoek, W., Meyer, J.J.C.: Formalising abilities and opportunities of agents. Fundamenta Informaticae 34(1-2), 53–101 (1998)
6. van der Hoek, W., Wooldridge, M.: Towards a logic of rational agency. Logic Journal of the IGPL 11(2), 135–159 (2003)
7. Schmdit, R., Tishkovsky, D., Hustadt, U.: Interactions between knowledge, action and commitment within agent dynamic logic. Studia Logica 78, 381–415 (2004)
8. Herzig, A., Longin, D.: C&l intention revisited. In: KR, pp. 527–535 (2004)
9. Lorini, E., Herzig, A.: A logic of intention and attempt. Synthese 163(1), 45–77 (2008)
10. Lorini, E., van Ditmarsch, H.P., Lima, T.D., Lima, T.D.: A logical model of intention and plan dynamics. In: ECAI, pp. 1075–1076 (2010)
11. Schmidt, R.A., Tishkovsky, D.: On combinations of propositional dynamic logic and doxastic modal logics. Journal of Logic, Language and Information 17(1), 109–129 (2008)
12. Kavka, G.S.: The toxin puzzle. Analysis 43(1), 33–36 (1983)
13. Singh, M.P.: A critical examination of the cohen-levesque theory of intentions. In: Proceedings of the 10th European Conference on Artificial intelligence, ECAI 1992, pp. 364–368. John Wiley & Sons, Inc. (1992)
14. Rao, A.S., Georgeff, M.P.: Decision procedures for bdi logics. J. Log. Comput. 8(3), 293–342 (1998)
15. Andreou, C.: The newxin puzzle. Philosophical Studies, 1–9 (July 2007)
16. Bratman, M.E.: Toxin, temptation, and the stability of intention. In: Coleman, J.L., Morris, C.W. (eds.) Rational Commitment and Social Justice: Essays for Gregory Kavka, pp. 59–83. Cambridge University (1998)
17. Gauthier, D.: Rethinking the toxin puzzle. In: Coleman, J.L., Morris, C.W. (eds.) Rational Commitment and Social Justice: Essays for Gregory Kavka, pp. 47–58. Cambridge University Press (1998)
18. Harman, G.: The toxin puzzle. In: Coleman, J.L., Morris, C.W. (eds.) Rational Commitment and Social Justice: Essays for Gregory Kavka, pp. 84–89. Cambridge University (1998)
19. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. The MIT Press (2000)
20. Davidson, D.: Essays on Actions and Events. Clarendon Press, Oxford (1980)
21. Meiland, J.W.: The Nature of Intention. Methuen & Co Ltd. (1970)