BDI Logics

John-Jules Ch. Meyer, Jan Broersen and Andreas Herzig

September 4, 2012

Abstract

This paper presents an overview of so-called BDI logics. Starting out from the basic ideas about BDI by Bratman, we consider various formalizations in logic, such as the approach of Cohen and Levesque, slightly remodelled in dynamic logic, Rao & Georgeff's influential BDI logic based on the branching-time temporal logic CTL*, the KARO framework and BDI logic based on stit (seeing to it that) logic.

1 Introduction

In this article we present an overview of so-called BDI logics, that is, logics that describe the mental attitudes of intelligent agents [55, 52] in terms of folk-psychological notions of beliefs, desires and intentions. This theory is based on the work of Bratman [6, 7]. The paper is organized as follows: we start with some of the basic ideas in Bratman's philosophy, which is about practical reasoning (the reasoning about performing actions) on the basis of the agent's beliefs, desires and, very importantly, intentions, which are special desires to which the agent is committed. Then a number of formalizations of BDI theory in logic is reviewed, the so-called BDI logics. Starting out with Cohen & Levesque's approach, slightly reworked in a dynamic logic setting by Andreas Herzig and colleagues. Then we look at Rao & Georgeff's BDI logic, based on the branching-time temporal logic CTL*. Next we discuss the KARO framework which was based on dynamic logic since its conception. We sketch as a small excursion how the KARO framework, which was devised to capture the behaviour of (rational) intelliegent agents, also can be use for describing emotional behaviour of agents. Finally we end with a relatively new approach to BDI, based on so-called stit (seeing to it that) logic.

2 Bratman's theory of Belief-Desire-Intention

> “What happens to our conception of mind and rational agency when we take seriously future-directed intentions and plans and their roles as inputs into further practical reasoning? This question drives much of this book."

This is how the preface of Michael E. Bratman's famous book “Intention, Plans, and Practical Reason" [6] starts. In this book the author lays down the foundations of what later would be called the BDI (Belief-Desire-Intention) theory of agency, a folk-psychological theory of how humans make decisions and take action (referred to as practical reasoning after Aristotle), and which would lead to a new computing paradigm, agent-oriented programming or agent technology more in general, when AI researchers started to apply it to the specification and implementation of artificial agents.

The main new ingredient in Bratman's theory is that of *intention*. Beliefs and desires were already known to be of importance in human behaviour. For instance, Daniel Dennett's intentional stance, the strategy of interpreting the behaviour of an entity by treating it as if it were a rational agent that governed its choice of action by a consideration of its beliefs and desires, already mentions the role of beliefs and desires [17]. But Bratman claimed that to fully understand the practical reasoning of humans also the notion of an intention is needed. An intention is not just a mere desire but something the agent is committed to, that is, not given up too soon by the agent. For instance, if I have an intention to give a lecture in Amsterdam tomorrow, it is not a mere wish to do so, but I'm really taking measures (making plans, e.g. canceling other plans or making sure my laptop will be in my bag) to do it and unless something happens that seriously interferes with my intention to give that lecture tomorrow, I really will do so. Thus Bratman takes intention to be a first-class citizen, and not something that can be reduced to beliefs and desires. Thus a reduction of intention to a theory of only beliefs and desires is rejected by Bratman ([6], p. 6 - 9).

Bratman focuses on *future-oriented intentions*. Such intentions differ from present-directed intentions, alias intentions-in-action, which accompany an agent's actions (more precisely, an agent's intentional actions).

To explain the differences between beliefs, desires and intentions Bratman introduces the notion of a *pro-attitude*. A pro-attitude is an agent's mental attitude directed toward an action under a certain description [5]. It plays a motivational role. So desires and intentions are both pro-attitudes while beliefs typically are not. But although desires and intentions are both pro-attitudes they differ. Intentions are conduct-controlling pro-attitudes, while ordinary desires are merely potential influences of action. The '*volitional*' dimension of the commitment involved in future-directed intentions comes from the conduct-controlling nature of intentions: as a conduct-controlling pro-attitude an intention involves a special commitment to action that ordinary desires do not.

Besides identifying intentions as conduct-controlling pro-attitudes, Bratman argues that intentions also have other properties: they have inertia and they serve as inputs into further practical reasoning. By the former is meant that intentions resist reconsideration. Once an intention has been formed (and a commitment to action has been made) the intention will normally remain intact until the time of action: it has a characteristic stability / inertia. By this Bratman means that intentions made influence further reasoning about (decisions about) action, where also refinements of intentions (intentions to do more concrete actions) may play a role. For example, if I have the intention to speak in Amsterdam tomorrow, I can form a more refined intention to take the car driving to Amsterdam in order to speak. As a consequence after the second intention it won't be rational anymore to consider time tables for trains going to Amsterdam, while it was so after the first intention. All this has led to seeing intentions as distinctive states of mind, distinct from beliefs and desires, and to a belief-desire-intention model rather than a desire-belief model of practical reasoning.

As we have seen, Bratman describes how prior intentions and plans provide a filter of admissibility on options. This is what later by Cohen & Levesque [13] has been called a 'screen of admissibility'. The basis of this role of intentions in further practical reasoning is the need for consistency in one's web of intentions and beliefs, as Bratman calls it: other things being equal, it should be possible for me to do all that I intend in a world in which my beliefs are true. But as Bratman explains this is not as simple as it looks. In particular "not every option that is incompatible with what the agent already intends and believes is inadmissible." In short, it depends on whether beliefs can be forced to be changed by the new intention so that the inconsistency disappears or not. In the former case the new intention is admissible, in the latter it is not. (For more on this rather subtle issue, including an elaborate example, we refer to [6], pages 41–42.)

In Bratman's view there is an intrinsic relation between intentions and plans. Plans are intentions. They share the properties of intentions: they resist reconsideration and have inertia, they are conduct controllers and not merely conduct influencers, and they provide crucial inputs for further practical reasoning and planning. But they have increased complexity as compared to simple intentions: they are typically partial in the sense of incomplete (typically I have a partial plan to do something and fill in the details later) and have a hierarchical structure (plans concerning ends embed plans concerning means and preliminary steps, and more general intentions embed more specific ones).

To sum it up, according to Bratman, future-oriented intentions have the following characteristics:

- An intention is a *high-level plan.*
- An intention guides deliberation and triggers further planning: it typically leads to the *refinement* of a high-level plan into a more and more precise plan.
- An intention comes with the agent's *commitment* to achieve it.
- An agent abandons an intention only under the following conditions:
  - the intention has been achieved;
  - he believes it is impossible to achieve it;
  - he abandons another intention for which it is instrumental.

Let us illustrate Bratman's future-oriented intentions by an example. Suppose we are in autumn and I desire to go to Paris next spring. Under certain conditions —such as the importance of that desire and my beliefs about its feasibility—, that desire will make me form an intention to travel to Paris next spring. This is a very high-level plan: I do not settle the exact dates, I do not decide by which means of transportation I am going to go to Paris, and I do not know where to stay yet. I am however committed to that plan: during the following months I will stick to my intention to go to Paris, unless I learn that it is impossible to go to Paris in spring (say because my wife wants to spend our spring holidays in Spain, or because I changed my mind due to an invitation to give a talk at an important conference). During the next months I am going to refine my high-level plan: I will decide to go on a particular weekend, I will decide to go by train and not by plane, and I will book a hotel for the weekend under concern. This more elaborated plan is going to be refined further as time goes by: I decide to take the 7am train and not the 9am train, and I decide to go to the train station by metro and not by taxi, etc. Finally, once I have spent that weekend in Paris I no longer pursue that goal and drop it.

Bratman's theory might be called semi-formal: while he isolates the fundamental concepts and relates them, he does not provide a formal semantics. This was both undertaken by Phil Cohen and Hector Levesque and, more or less at the same time, by Anand Rao and Michael Georgeff in much-cited papers that won, respectively, 2006 and 2007 IFAAMAS Awards for Influential Papers in Autonomous Agents and Multiagent Systems [13, 45].

In the next section we will go into the details of how they casted Bratman's theory into a logic of intention and we present some subsequent modifications and extensions of their original logic. This will be followed by a section on Rao and Georgeff's approach.

2.1 Cohen and Levesque's approach to intentions

We have seen that the concepts of belief, desire, time and action play an important role in Bratman's theory of intention. A logical analysis of that theory should involve combining a logic of belief, a logic of desire, a logic of time and a logic of action.

Belief, time, and action play a fundamental role in Cohen and Levesque's logic. However, the concept of desire is somewhat neglected: Cohen and Levesque rather base their logic on the concept of *realistic preference*. The latter can be viewed as a desire that has already been filtered by the agent's beliefs about its realisability. This is highlighted by the property that belief implies realistic preference: when I am convinced that φ is true then I also have to prefer that φ is true. (I might however prefer that φ be false at some point in the future.)

Cohen and Levesque's analysis amounts to a *reduction* of the concept of intention to those of belief, realistic preference, time and action: they define intention in terms of the latter four concepts. The reader may note that this is actually a surprising move, given that Bratman had strongly argued that intentions are independent of desires and cannot be reduced to them.

In the next two sections we present the four building blocks of Cohen and Levesque's logic, grouping together action and time on the one hand, and belief and realistic preference on the other.

Action and time

The basic building block of Cohen and Levesque's logic is a linear version of propositional dynamic logic PDL. The semantics of linear PDL allows to also interpret the temporal operators of linear-time temporal logic LTL.

Standard PDL. Standard PDL is not about actions but about events. It has a set $\mathcal{A}$ of atomic event names. Cohen and Levesque add agents to the picture and provide an *agentive version* of PDL. Let us write i, j, etc. for agents from some set of individuals $\mathcal{I}$. Then atomic actions are elements of $\mathcal{I} \times \mathcal{A}$. We write them $i{:}\alpha$ where $\alpha \in \mathcal{A}$ is an atomic event and $i \in \mathcal{I}$. Formulas of the language of PDL are built from atomic formulas and atomic actions by means of modal operators $\texttt{Poss}_\pi$, where π is an action. The formula $\texttt{Poss}_\pi\varphi$ reads "there is a possible execution of π after which φ is true".[1] This reading highlights that the standard version of PDL allows for several possible executions of π in order to account for indeterminism.

While $\texttt{Poss}_\pi$ quantifies existentially over the executions of π, the dual modal operator $\texttt{After}_\pi$ quantifies universally. It is definable from $\texttt{Poss}_\pi$ as follows:

$$\texttt{After}_\pi\varphi \overset{\text{def}}{=} \neg\texttt{Poss}_\pi\neg\varphi$$

Let us consider the case where φ is truth $\top$ or falsity $\bot$: $\texttt{Poss}_\pi\top$ has to be read "π is executable", while $\texttt{After}_\pi\bot$ has to be read "φ is inexecutable".

The semantics of PDL is based on *transition systems* where an atomic action $i{:}\alpha$ can be interpreted as a set of edges. Such a transition system is a couple $\langle W, R\rangle$ where W is a non-empty set of possible worlds and R maps every action π to an accessibility relation $R_\pi \subseteq W \times W$ relating possible worlds. An edge from world w to world u that is labeled π means that it is possible to execute π in w and that u is a possible outcome world when π is executed. The set of all these π edges makes up the accessibility relation R_π interpreting the action π.

A PDL *model* is a transition system together with a valuation V mapping atomic formulas p from the set of propositional variables $\mathcal{P}$ to their extension $V(p) \subseteq W$, i.e., to the set of worlds $V(p)$ where p is true.

Models allow to give truth values to formulas. In particular, $\texttt{Poss}_\pi\varphi$ is true at a world w if there is a couple (w, w') in R_π such that φ is true at world w':

$$M, w \models \texttt{Poss}_\pi\varphi \quad \text{iff} \quad \text{there is a } u \in W \text{ such that } wR_\pi u \text{ and } M, u \models \varphi$$

The formula $\texttt{Poss}_\pi\varphi$ therefore expresses a weak notion of ability: the action π might occur and φ could be true afterwards.

[1] The standard notation is $\langle\pi\rangle\varphi$; we here deviate in order to be able to distinguish actual action from potential action.

Linear PDL. Probably Cohen and Levesque were the first to adapt PDL in order to model actual agency [13]. The modalities are interpreted in *linear* PDL models. In such models, for every possible world w there is at most one successor world u that is temporally related to w. The accessibility relation linking w to u may be labelled by several atomic actions. Formally, a transition system $\langle W, R, V\rangle$ is linear if for every world $w \in W$ such that $\langle w, u_1\rangle \in R_{\pi_1}$ and $\langle w, u_2\rangle \in R_{\pi_2}$ we have $u_1 = u_2$. An edge from world w to world u that is labeled π means that π is executed in w and that u will be the result. (The reader might note the difference with the above standard PDL.) This allows for the simultaneous performance of two different actions; they must however lead to the same outcome world. The models of linear PDL are the class of linear transition systems.

In order to distinguish the modal operators of actual action from the modal operators of possible action we write the former as $\texttt{Happ}_\pi\varphi$, read "π is going to be performed, and φ is true afterwards". Just as $\texttt{After}_\pi$ is the dual of $\texttt{Poss}_\pi$, we define a modal operator $\texttt{IfHapp}_\pi$ that is the dual of $\texttt{Happ}_\pi$ by stipulating:

$$\texttt{IfHapp}_\pi\varphi \overset{\text{def}}{=} \neg\texttt{Happ}_\pi\neg\varphi$$

$\texttt{Happ}_\pi\varphi$ and $\texttt{IfHapp}_\pi\varphi$ say different things: the first formula says that π is executable and that φ is true after it, while the second says that *if* π is executable then φ is true after it. The former should therefore imply the latter.

The truth condition for $\texttt{Happ}_\pi$ is:

$$M, w \models \texttt{Happ}_\pi\varphi \quad \text{iff} \quad \text{there is a } u \in W \text{ such that } wR_\pi u \text{ and } M, u \models \varphi$$

So it has exactly the same form as that for $\texttt{Poss}_\pi$. We changed the name of the modal operator in order to better suit the linearity of the models.

The following axiom schema characterises linear PDL models:

$$(\texttt{Happ}_{i:\alpha}\top \wedge \texttt{Happ}_{j:\alpha'}\varphi) \rightarrow \texttt{Happ}_{i:\alpha}\varphi \tag{1}$$

Beyond atomic events, PDL also has complex events such as sequential and nondeterministic composition, test, and iteration. We will however not refer to them in our present introduction.

Cohen and Levesque's logic has the temporal operators "eventually" (noted $\texttt{F}$), "henceforth" (noted $\texttt{G}$), and "until" (noted $\texttt{U}$). These operators are interpreted in linear PDL models in the obvious way. Let us give the truth condition for the 'eventually" operator: For example:

$$\begin{aligned} M, w \models \texttt{F}\varphi \quad \text{iff} \quad & \text{there is an integer } n \text{ and there are } v_1, \ldots, v_n \in W \text{ such} \\ & \text{that } v_1 = w, \langle v_k, v_{k+1}\rangle \in R_{\pi_k} \text{ for some } \pi_k, \text{ and } M, v_n \models \varphi \end{aligned}$$

Cohen and Levesque also need existential quantification over actions. We here present their account in terms of an operator fusing existential quantification $\exists$ over events α with the dynamic operator $\texttt{Happ}_{i:\alpha}$. Its truth condition is as follows:

$$\begin{aligned} M, w \models \exists\alpha\texttt{Happ}_{i:\alpha}\varphi \quad \text{iff} \quad & \text{there are } \alpha \in \mathcal{A}, u \in W \text{ such that} \\ & \langle w, u\rangle \in R_{i:\alpha} \text{ and } M, u \models \varphi \end{aligned}$$

Belief and preference

Cohen and Levesque's account of belief is standard, while their account of preference is in terms of the somewhat unusual notion of strong realistic preference.

Belief. Cohen and Levesque have modal operators of belief $\texttt{Bel}_i$, one per agent i. The modal logic of each of these operators is the standard logic of belief $\mathsf{KD45}$. Such operators can be interpreted if we add accessibility relations B_i to the transition systems of linear PDL, one per agent i. The set of worlds $B_i(w) = \{u : \langle w, u\rangle \in B_i\}$ is the set of those worlds that are possible for agent i at world w: the set of worlds that are compatible with his beliefs at w.

In order to be an accessibility relation for $\mathsf{KD45}$, each of these relations has to satisfy the following constraints:

- for every $w \in W$ there is at least one $u \in W$ such that $\langle w, u\rangle \in B_i$ (seriality);
- if $\langle w, u\rangle \in B_i$ and $\langle u, v\rangle \in B_i$ then $\langle w, v\rangle \in B_i$ (transitivity);
- if $\langle w, u\rangle \in B_i$ and $\langle w, v\rangle \in B_i$ then $\langle u, v\rangle \in B_i$ (Euclideanity).

These constraints make that the following implications become valid:

- $\texttt{Bel}_i\varphi \to \neg\texttt{Bel}_i\neg\varphi$ (consistency of belief, axiom D)
- $\texttt{Bel}_i\varphi \to \texttt{Bel}_i\texttt{Bel}_i\varphi$ (positive introspection, axiom $\mathsf{4}$)
- $\neg\texttt{Bel}_i\varphi \to \texttt{Bel}_i\neg\texttt{Bel}_i\varphi$ (negative introspection, axiom $\mathsf{5}$)

Preference. For Cohen and Levesque, intentions are particular *strong realistic preferences.* The latter are among the worlds that are possible for an agent there is a subset the agent prefers. There is a modal operator $\texttt{Pref}_i$, one per agent i, and the formula $\texttt{Pref}_i\varphi$ reads "i chooses φ to be true".[2] Such a notion of preference is strongly realistic in the sense that belief logically implies preference. Semantically, strong realistic preference can be modelled by accessibility relations P_i, one per agent $i \in \mathcal{I}$, such that $P_i \subseteq B_i$. The latter constraint implements realism: a world that is compatible with agent i's preferences cannot be incompatible with i' beliefs. In other words, at world w agents have to select their preferred worlds among the worlds that are epistemically possible for them at w.

The logic of action, time, belief, and preference

Let us sum up Cohen and Levesque's semantics. A *frame* is a quadruple $M = \langle W, R, B, P\rangle$ where

- W is a non-empty set of possible worlds;
- $R : (\mathcal{I} \times \mathcal{A}) \longrightarrow (W \times W)$ maps actions π to accessibility relations R_π;
- $B : \mathcal{I} \longrightarrow (W \times W)$ maps agents i to accessibility relations B_i;
- $P : \mathcal{I} \longrightarrow (W \times W)$ maps agents i to accessibility relations P_i;

These frames have to satisfy the following constraints:

- every B_i is serial, transitive and Euclidean;
- $P_i \subseteq B_i$, for every $i \in \mathcal{I}$.

Let us call $\mathcal{CL}$ that class of frames. As usual, a model is a frame together with a valuation $V : \mathcal{P} \longrightarrow 2^W$ mapping atomic formulas p to their extension $V(p) \subseteq W$. Validity and satisfiability in $\mathcal{CL}$ frames are defined as usual.

We are now ready to formulate Cohen and Levesque's reduction of intention.

[2] Cohen and Levesque's original notation is $\texttt{Goal}_i$ instead of $\texttt{Pref}_i$ (while they actually refer to it in their title as 'choice'). We moved to ours in order to avoid confusion with the concept of choice in stit theory.

Defining intention

Cohen and Levesque define a modal operator of intention by means of a cascade of definitions. We here reproduce them in a slightly simplified form. We then discuss them and finally comment on the modifications.

1. i has the *achievement goal* that φ if i prefers that φ is eventually true and believes that φ is currently false. Formally:

$$\texttt{AGoal}_i\varphi \overset{\text{def}}{=} \texttt{Pref}_i\texttt{F}\varphi \land \texttt{Bel}_i\neg\varphi$$

2. i has the *persistent goal* that φ if i has the achievement goal that φ and will keep that goal until it is either fulfilled or believed to be out of reach. Formally:

$$\texttt{PGoal}_i\varphi \overset{\text{def}}{=} \texttt{AGoal}_i\varphi \land (\texttt{AGoal}_i\varphi)\,\texttt{U}\,(\texttt{Bel}_i\varphi \lor \texttt{Bel}_i\,\texttt{G}\neg\varphi)$$

3. i has the *intention* that φ if i has the persistent goal that φ and believes he can achieve φ by an action of his. This requires to quantify over i's actions by means of the fused operator quantifying over events. Formally:

$$\texttt{Intend}_i\varphi \overset{\text{def}}{=} \texttt{PGoal}_i\varphi \land \texttt{Bel}_i\,\texttt{F}\,\exists\alpha\,\texttt{Happ}_{i:\alpha}\varphi$$

Some valid and invalid principles for intention. Cohen and Levesque's construction guarantees several desirable properties and avoids some that are unwanted. Here are two of them.

First, i's intention that φ logically implies i's belief that $\neg\varphi$. Formally this writes:

$$\texttt{Intend}_i\varphi \to \texttt{Bel}_i\neg\varphi$$

Second, the formula schema $(\texttt{Bel}_i(\varphi \to \psi) \to (\texttt{Intend}_i\varphi \to \texttt{Intend}_i\psi)$ is invalid: i's intention that φ together with i's belief that φ implies ψ does not logically imply i's intention that ψ. This is crucial both for Bratman and for Cohen and Levesque. Here is a famous example illustrating why the principle should not be valid: if I intend to go to the dentist and believe that going to the dentist will cause pain then I do not necessarily intend to have pain.

Comments on the simplifications. We have simplified the definition of a persistent goal. Cohen and Levesque's original definition allows agents to abandon a persistent goal for some other, superior reason. So their definition is

$$\texttt{PGoal}_i\varphi \overset{\text{def}}{=} \texttt{AGoal}_i\varphi \land (\texttt{AGoal}_i\varphi)\,\texttt{U}\,(\texttt{Bel}_i\varphi \lor \texttt{Bel}_i\texttt{G}\neg\varphi \lor \psi)$$

where ψ is an unspecified condition accounting for that other reason. Such a general condition makes it however difficult to go beyond specific cases.

Variants and extensions

We now overview an extension of the basic logic, together with several alternatives to Cohen and Levesque's incremental definition of intention.

Introspection of intention. Cohen and Levesque did not assume principles of positive and negative introspection of preference. However, they seem natural given that preference is supposed to be realistic and were been added by other authors since [22, 23, 34]. They take the form

- $\mathtt{Pref}_i\varphi \to \mathtt{Bel}_i\mathtt{Pref}_i\varphi$
- $\neg\mathtt{Pref}_i\varphi \to \mathtt{Bel}_i\neg\mathtt{Pref}_i\varphi$

They correspond to the following constraints on the accessibility relations for preference and belief:

- if $\langle w,u\rangle \in B_i$ and $\langle u,v\rangle \in P_i$ then $\langle w,v\rangle \in P_i$;
- if $\langle w,u\rangle \in B_i$ and $\langle w,v\rangle \in P_i$ then $\langle u,v\rangle \in P_i$.

This allows to prove principles of positive and negative introspection of goals, achievement goals, persistent goals, and intention. For instance, they validate $\mathtt{Intend}_i\varphi \to \mathtt{Bel}_i\mathtt{Intend}_i\varphi$ and $\neg\mathtt{Intend}_i\varphi \to \mathtt{Bel}_i\neg\mathtt{Intend}_i\varphi$. For example, both when I intend to go to Paris and when I don't intend to go to Paris then I am aware of this.

Weakly realistic preference. Sadek has argued for a slightly different notion of realistic preference [47]. The latter does not demand that all preference-accessible worlds be in the set of belief-accessible worlds, but only requires that they have a non-empty intersection. (This is sometimes called weak realism, as opposed to Cohen and Levesque's strong realism.) In frames that are weakened in this way the somewhat counterintuitive principle $\mathtt{Bel}_i\varphi \to \mathtt{Pref}_i\varphi$ is no longer valid. Instead, the weaker $\mathtt{Bel}_i\varphi \to \neg\mathtt{Pref}_i\neg\varphi$ is valid, which can be reformulated as

$$\neg(\mathtt{Bel}_i\varphi \wedge \mathtt{Pref}_i\neg\varphi)$$

It says that one cannot simultaneously believe that φ and prefer that $\neg\varphi$.

An epistemic version of achievement goals. Herzig and Longin have advocated a different definition of an achievement goal [22]. It is weaker than Cohen and Levesque's in that they only require that i does not believe that φ is currently true (instead of i's belief that φ is currently false). It is stronger in that they replace i's goal that φ will be true by i's goal that φ will be *believed.* This is the following definition:

$$\mathtt{AGoal}^{\mathtt{w}}_i\varphi \overset{\text{def}}{=} \mathtt{Pref}_i\,\mathtt{F}\,\mathtt{Bel}_i\varphi \wedge \neg\mathtt{Bel}_i\varphi$$

They start by arguing for the replacement of $\mathtt{Pref}_i\mathtt{F}\varphi$ by $\mathtt{Pref}_i\mathtt{F}\mathtt{Bel}_i\varphi$. As they point out, it is the *raison d'être* of an intention to be abandoned at some stage, and an agent can only do so if he believes that he has achieved that goal. So the agent's goal cannot just be that φ be true, but should be that he *believes* that φ is true. Using the same reasons they then argue for the replacement of $\mathtt{Bel}_i\neg\varphi$ by $\neg\mathtt{Bel}_i\varphi$: as long as φ is not believed to be true the agent should stick to his achievement goal φ, so $\neg\mathtt{Bel}_i\varphi$ is better in line with this than $\mathtt{Bel}_i\neg\varphi$. They illustrate the first replacement by a variant of the Byzantine generals example. Let r mean that a message of general i has been received by general j. Suppose i initially believes that j has not received the message yet, i.e., $\mathtt{Bel}_i\neg r$. Suppose moreover that i believes that he will actually *never* know whether j received the message or not, i.e., $\mathtt{Bel}_i\mathtt{G}(\neg\mathtt{Bel}_i r \wedge \neg\mathtt{Bel}_i\neg r)$. (This differs from the original Byzantine generals example, where it is possible that the messengers get through and where it is just possible for i that he will never know.) If we express i's achievement goal that r as $\mathtt{Pref}_i\,\mathtt{F}\,r$ then Cohen and Levesque make us conclude that $\mathtt{AGoal}_i r$, i.e., i has the achievement goal that φ although he believes that he will never be able to abandon that goal.

In contrast, if we express i's achievement goal that r as $\mathtt{Pref}_i\,\mathtt{F}\,\mathtt{Bel}_i r$ then we have $\neg\mathtt{AGoal}_i^w r$: i cannot have the achievement goal that r.[3]

Weaker link between action and goal. Sadek and Bretier have pointed out that the definition of intention is too strong in particular in cooperative situations where agent i's action need not directly achieve his goal φ: it is enough that i triggers a subsequent action of another agent j which will achieve i's goal [8, 48]. Their modification can be formulated as follows:

$$\mathtt{Intend}_i\varphi \overset{\text{def}}{=} \mathtt{PGoal}_i\varphi \wedge \mathtt{Pref}_i\mathtt{F}(\exists\alpha\mathtt{Happ}_{i:\alpha}\mathtt{F}\varphi)$$

Stronger commitment. Sadek and Bretier have discussed a stronger definition of intention where the agent is committed to do all he can to achieve his goal [8, 48]. They express this by a universal quantification over events.[4] We formulate their definition as follows:

$$\mathtt{Intend}_i\varphi \overset{\text{def}}{=} \mathtt{PGoal}_i\varphi \wedge \mathtt{Pref}_i\,\forall\alpha(\mathtt{Bel}_i\,\mathtt{Happ}_{i:\alpha}\,\mathtt{F}\varphi \rightarrow \mathtt{Pref}_i\,\mathtt{F}\,\mathtt{Happ}_{i:\alpha}\top)$$

That definition was criticised in the literature as being too strong [22]. Indeed, it postulates that agents want to achieve their intentions by all possible means, including illegal actions and actions with a huge cost for them. For example, it might commit me to steal a car if this is the only means to go to Paris on that spring weekend (say because there is a train strike).

Attempts. Lorini and Herzig complemented Cohen and Levesque's approach by integrating the concept of an *attempt* to perform an action [35]. The motivation is that intentions typically make an agent *try* to perform an action, while the successful performance of that action is not guaranteed. The central principle there is "can and attempts implies does" : if i intends to (attempt to) perform α and α is feasible then α will indeed take place. This requires a logic with both modal operators of possible action $\mathtt{Poss}_\pi$ and modal operators of actual action $\mathtt{Happ}_\pi$.

Cohen and Levesque succeeded in providing a fine-grained analysis of intention by relating that concept to action, belief and realistic preference. A central point in Bratman's theory their logic does not account for is the refinement of intentions. According to Bratman, an agent starts by forming high-level intentions such as going to Paris in a month, and as time goes by he makes that intention more precise: he first starts to intend to go to Paris by train and not by plane; at a later stage he decomposes the intention to go to Paris by train into the intention to take a taxi to the train station (instead of a bus), then take the TGV to Paris, and then take the metro. It is probably an interesting direction of future research to integrate intention refinement mechanisms e.g. by resorting to dynamic epistemic logics.

2.2 Rao & Georgeff's BDI logic

As mentioned earlier, besides Cohen and Levesque, also Rao and Georgeff, more or less at the same time, published a formalisation of the ground-breaking work of Bratman [6] on the philosophy of intelligent (human) agents. As we have seen, Bratman made a case for the notion of *intention* besides belief and desire, to describe the behaviour of rational agents. Intentions force the agent to commit to certain desires and to really 'go for them'. So focus of attention is an important aspect here, which also enables the agent to monitor how s/he is doing and

[3] Our hypothesis $\mathtt{Bel}_i\neg\varphi$ implies the second condition $\neg\mathtt{Bel}_i\varphi$ because the logic of belief contains the $\mathtt{D}$ axiom, and $\mathtt{Bel}_i\mathtt{G}(\neg\mathtt{Bel}_i r \wedge \neg\mathtt{Bel}_i\neg r)$ implies $\neg\mathtt{Pref}_i\mathtt{F}\mathtt{Bel}_i r$, which is the negation of the first condition.

[4] This is therefore not a fused operator. In order to save space we do not give the the details of the semantics of that quantifier and leave the readers with their intuitions about it.

take measures if things go wrong. Rao & Georgeff stress that in the case of resource-bounded agents it is imperative to focus on desires / goals and make choices. This was also observed by Cohen & Levesque [13], who tried to formalize the notion of intention in a linear-time temporal logic (or, as we have seen in the previous section, a linear version of dynamic logic) in terms of the notion of a (persistent) goal. Here we treat Rao & Georgeff's approach [45] who use a branching-time temporal logic framework (CTL* [19]) to give a formal-logical account of BDI theory. Like Cohen & Levesque's approach, BDI logic has influenced many researchers (including Rao & Georgeff themselves) to think about architectures of agent-based systems in order to realize these systems (cf. [54]). Rao & Georgeff's BDI logic is more liberal than that of Cohen & Levesque in the sense that they *a priori* regard each of the three attitudes of belief, desire and intention as primitive: they introduce separate modal operators for belief, desire and intention, and then study possible relations between them.

(The language of) BDI logic is constructed as follows. Two types of formulas are distinguished: state formulas and path formulas. We assume some given first-order signature. Furthermore, we assume a set E of event types with typical element e. The operators BEL, $GOAL$, $INTEND$ have as obvious intended reading the belief, goal and intention of an agent, respectively, while $\mathsf{U}, \diamond, \mathsf{O}$ are the usual temporal operators, viz. until, eventually and next, respectively.

Definition 2.1 *(State and path formulas.)*

1. *The set of* state formulas *is the smallest closed under:*
   - *any first-order formula w.r.t. the given signature is a state formula*
   - *if* φ_1 *and* φ_2 *are state formulas then also* $\neg\varphi_1, \varphi_1 \vee \varphi_2, \exists x \varphi_1(x)$ *are state formulas*
   - *if* e *is an event type, then* $succeeded(e), failed(e)$ *are state formulas*
   - *if* φ *is a state formula, then* $BEL(\varphi), GOAL(\varphi), INTEND(\varphi)$ *are state formulas*
   - *if* ψ *is a* path formula*, then* $optional(\psi)$ *is a state formula*
2. *The set of* path formulas *is the smallest set closed under:*
   - *any state formula is a path formula*
   - *if* ψ_1, ψ_2 *are path formulas, then* $\neg\psi_1, \psi_1 \vee \psi_2, \psi_1 \mathsf{U} \psi_2, \diamond\psi_1, \mathsf{O}\psi_1$ *are path formulas*

State formulas are interpreted over a state, that is a (state of the) world at a particular point in time, while path formulas are interpreted over a path of a time tree (representing the evolution of a world). In the sequel we will see how this will be done formally. Here we just give the informal readings of the operators.

The operators *succeeded* and $failed$ are used to express that events have (just) succeeded and failed, respectively.

Next there are the modal operators for belief, goal and intend. (In the original version of BDI theory [45], desires are represented by goals, or rather a GOAL operator. In a later paper [46] the $GOAL$ operator was replaced by DES for desire.) The optional operator states that there is a future (represented by a path) where the argument of the operator holds. Finally, there are the familiar (linear-time) temporal operators, such as the 'until', 'eventually' and 'nexttime', which are to be interpreted along a linear time path.

Furthermore, the following abbreviations are defined:

Definition 2.2

1. $\Box\psi = \neg \diamond \neg\psi$ *(always)*
2. $inevitable(\psi) = \neg optional(\neg\psi)$
3. $done(e) = succeeded(e) \vee failed(e)$
4. $succeeds(e) = inevitable \mathsf{O}(succeeded(e))$

5. $fails(e) = inevitable \mathsf{O}(failed(e))$
6. $does(e) = inevitable \mathsf{O}(done(e))$

The 'always' operator is the familiar one from (linear-time) temporal logic. The 'inevitability' operator expresses that its argument holds along all possible futures (paths from the current time). The 'done' operator states that an event occurs (action is done) no matter whether it is succeeding or not. The final three operators state that an event succeeds, fails, or is done iff it is inevitable (i.e. in any possible future) it is the case that at the next instance the event has succeeded, failed, or has been done, respectively. (so, this means that an event, succeeding or failing, is supposed to take one unit of time!)

Definition 2.3 *(Semantics.)*
The semantics is given w.r.t. models of the form $\mathcal{M} = \langle W, E, T, \prec, \mathcal{U}, B, G, I, \Phi\rangle$, *where*

- W *is a set of possible worlds*
- E *is a set of primitive event types*
- T *is a set of time points*
- $\prec$ *is a binary relation on time points, which is serial, transitive and back-wards linear*
- $\mathcal{U}$ *is the universe of discourse*
- Φ *is a mapping of first-order entities to* $\mathcal{U}$, *for any world and time point*
- $B, G, I \subseteq W \times T \times W$ *are accessibility relations for* $BEL, GOAL, INTEND$, *respectively*

The semantics of BDI logic, Rao & Georgeff-style, is rather complicated. Of course, we have possible worlds again, but as we will see below, these are not just unstructured elements, but they are each time trees, describing possible flows of time. So, we also need time points and an ordering on them. As BDI logic is based on branching time, the ordering need not be linear in the sense that all time points are related in this ordering. However, it is stipulated that the time ordering is serial (every time point has a successor in the time ordering), the ordering is transitive and backwards-linear, which means that every time point has only one direct predecessor. The accessibility relations for the 'BDI'-modalities are standard apart from the fact that they are also time-related, that is to say that worlds are (belief/goal/intend-)accessible with respect to a time point. Another way of viewing this is that – for all three modalities – for every time point there is a distinct accessibility relation between worlds.

In order to obtain reasonable properties for beliefs, desires and intentions, a number of constraints on the accessibility relations are stipulated. First of all, a *world / time point compatibilty* requirement ([53]) is assumed for all of the B, G, I accessibility relations: for $R = B, G, I$:

$$\textit{If } w' \in R(w,t) \textit{ then } t \in w \textit{ and } t \in w'$$

where $R(w,t) = \{w' \mid R(w,t,w')\}$ for $R = B, G, I$. This requirement is needed for the semantic clauses for the BEL, GOAL and INTEND modalities that we will give below to work. And next there are the usual requirements of the B accessibility relation to satisfy seriality, transitivity and euclidicity in order to obtain the familiar KD45 properties of belief: beliefs are consistent, and positive and negative introspection. As to the G and I accessibility relations we require seriality in order to obtain the well-known KD property of consistent goals and intentions.

Next we elaborate on the structure of the possible worlds.

Definition 2.4 *(Possible worlds.)*
Possible worlds in W *are assumed to be* time trees*: an element* $w \in W$ *has the form* $w = \langle T_w, A_w, S_w, F_w\rangle$ *where*

- $T_w \subseteq T$ *is the set of time points in world* w

- *A_w is the restriction of the relation $\prec$ to T_w*
- *$S_w : T_w \times T_w \to E$ maps adjacent time points to (successful) events*
- *$F_w : T_w \times T_w \to E$ maps adjacent time points to (failing) events*
- *the domains of the functions S_w and F_w are disjoint*

As announced before, a possible world itself is a time tree, a temporal structure representing possible flows of time. The definition above is just a technical one stating that the time relation within a possible world derives naturally from the *a priori* given relation on time points. Furthermore it is indicated by means of the functions S_w and F_w how events are associated with adjacent time points.

Now we come to the formal interpretation of formulas on the above models. Naturally we distinguish state formulas and path formulas, since the former should be interpreted on states whereas the latter are interpreted on paths. In the sequel we use the notion of a *fullpath*: a fullpath in a world w is an *infinite* sequence of time points such that, for all i, $(t_i, t_{i+1}) \in A_w$. We denote a fullpath in w by $(w_{t0}, w_{t1}, \ldots)$, and define $fullpaths(w)$ as the set of all fullpaths occurring in world w (i.e. all fullpaths that start somewhere in the time tree w).

Definition 2.5 *(Interpretation of formulas.) The interpretation of formulas w.r.t. a model $\mathcal{M} = \langle W, E, T, \prec, \mathcal{U}, B, G, I, \Phi \rangle$ is now given by:*

1. *(state formulas)*
   - $\mathcal{M}, v, w_t \models q(y_1, \ldots, y_n) \leftrightarrow (v(y_1), \ldots, v(y_n)) \in \Phi(q, w, t)$
   - $\mathcal{M}, v, w_t \models \neg\varphi \leftrightarrow \mathcal{M}, v, w_t \not\models \varphi$
   - $\mathcal{M}, v, w_t \models \varphi_1 \vee \varphi_2 \leftrightarrow \mathcal{M}, v, w_t \models \varphi_1$ *or* $\mathcal{M}, v, w_t \models \varphi_2$
   - $\mathcal{M}, v, w_t \models \exists x \varphi \leftrightarrow \mathcal{M}, v\{d/x\}, w_t \models \varphi$ *for some* $d \in \mathcal{U}$
   - $\mathcal{M}, v, w_{t0} \models optional(\psi) \leftrightarrow$ *exists fullpath* $(w_{t0}, w_{t1}, \ldots)$ *such that* $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \psi$
   - $\mathcal{M}, v, w_t \models BEL(\varphi) \leftrightarrow$ *for all* $w' \in B(w, t) : \mathcal{M}, v, w'_t \models \varphi$
   - $\mathcal{M}, v, w_t \models GOAL(\varphi) \leftrightarrow$ *for all* $w' \in G(w, t) : \mathcal{M}, v, w'_t \models \varphi$
   - $\mathcal{M}, v, w_t \models INTEND(\varphi) \leftrightarrow$ *for all* $w' \in I(w, t) : \mathcal{M}, v, w'_t \models \varphi$
   - $\mathcal{M}, v, w_t \models succeeded(e) \leftrightarrow$ *exists* $t0$ *such that* $S_w(t0, t) = e$
   - $\mathcal{M}, v, w_t \models failed(e) \leftrightarrow$ *exists* $t0$ *such that* $F_w(t0, t) = e$

   where $v\{d/x\}$ denotes the function v modified such that $v(x) = d$. (Note that clauses for BEL, GOAL and INTEND are well-defined due to the world / time point compatibility requirement that we have assumed to hold.)

2. *(path formulas)*
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \varphi \leftrightarrow \mathcal{M}, v, w_{t0} \models \varphi$, *for φ state formula*
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \mathsf{O}\varphi \leftrightarrow \mathcal{M}, v, (w_{t1}, w_{t2}, \ldots) \models \varphi$
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \diamond\varphi \leftrightarrow \mathcal{M}, v, (w_{tk}, \ldots) \models \varphi$ *for some* $k \geq 0$
   - $\mathcal{M}, v, (w_{t0}, w_{t1}, \ldots) \models \psi_1 \mathsf{U} \psi_2 \leftrightarrow$
     either there exists $k \geq 0$ *such that* $\mathcal{M}, v, (w_{tk}, \ldots) \models \psi_2$ *and for all* $0 \leq j < k :$ $\mathcal{M}, v, (w_{tj}, \ldots) \models \psi_1$, *or*
     for all $j \geq 0 :$ $\mathcal{M}, v, (w_{tj}, \ldots) \models \psi_1$

Most of the above clauses should be clear, including those concerning the modal operators for belief, goal and intention. The clause for the ‘optional’ operator expresses exactly that optionally ψ is true if ψ is true in one of the possible futures represented by fullpaths starting at the present time point. The interpretation of the temporal operators is as usual.

Rao & Georgeff now discuss a number of properties that may be desirable to have as axioms. In the following we use α to denote so-called *O-formulas*, which are formulas that contain no positive occurrences of the '*inevitable*' operator (or negative occurrences of '*optional*") outside the scope of the modal operators $BEL, GOAL$ and $INTEND$.

1. $GOAL(\alpha) \to BEL(\alpha)$
2. $INTEND(\alpha) \to GOAL(\alpha)$
3. $INTEND(does(e)) \to does(e)$
4. $INTEND(\varphi) \to BEL(INTEND(\varphi))$
5. $GOAL(\varphi) \to BEL(GOAL(\varphi))$
6. $INTEND(\varphi) \to GOAL(INTEND(\varphi))$
7. $done(e) \to BEL(done(e))$
8. $INTEND(\varphi) \to inevitable \diamond (\neg INTEND(\varphi))$

In order to render these formulas validities further constraints should be put on the models, since in the general setting above these are not yet valid.

For reasons of space we only consider the first two. (More can be found in [45, 46, 53].) In order to define constraints on the models such that these two become valid, we introduce the relation $\lhd$ on worlds, as follows:

$w'' \lhd w' \Leftrightarrow fullpaths(w'') \subseteq fullpaths(w')$. So $w'' \lhd w'$ means that there the world (time tree) w'' represents less choices than w'.

Now we define the *B-G condition* as the property that the following holds:

$$\forall w' \in B(w,t) \exists w'' \in G(w,t) : w'' \lhd w'$$

Informally, this condition says that for any belief accessible world there is a goal accessible world that contains less choices. It is now easy to show the following proposition.

Proposition 2.6 *Let $\mathcal{BG}$ be the class of models of the above form that satisfy the B-G condition. Then: $\mathcal{BG} \models GOAL(\alpha) \to BEL(\alpha)$ for O-formulas α.*

Similarly one can define the *G-I condition* as

$$\forall w' \in G(w,t) \exists w'' \in I(w,t) : w'' \lhd w'$$

and obtain:

Proposition 2.7 *Let $\mathcal{GI}$ be the class of models of the above form that satisfy the G-I condition. Then: $\mathcal{GI} \models INTEND(\alpha) \to GOAL(\alpha)$ for O-formulas α.*

Let us now consider the properties deemed desirable by Rao & Georgeff again. The first formula describes Rao & Georgeff's notion of 'strong realism' and constitutes a kind of belief-goal compatibilty: it says that the agent believes he can optionally achieve his goals. There is some controversy on this. Interestingly, but confusingly, Cohen & Levesque [13] adhere to a form of realism that renders more or less the converse formula $BELp \to GOALp$. But we should be careful and realize that Cohen & Levesque have a different logic in which one cannot express options as in the branching-time framework of Rao & Georgeff. Furthermore, it seems that in the two frameworks there is a different understanding of goals (and beliefs) due to the very difference in ontologies of time employed: Cohen & Levesque's notion of time could be called 'epistemically nondeterministic' or 'epistemically branching', while 'real' time is linear: the agents envisage several future courses of time, each of them being a linear history, while in

Rao & Georgeff's approach also 'real' time is branching, representing options that are available to the agent.

The second formula is a similar one to the first. This one is called goal-intention compatibilty, and is defended by Rao & Georgeff by stating that if an optionality is intended it should also be wished (a goal in their terms). So, Rao & Georgeff have a kind of selection filter in mind: intentions (or rather intended options) are filtered / selected goals (or rather goal (wished) options), and goal options are selected believed options. If one views it this way, it looks rather close to Cohen & Levesque's Intention is choice (chosen / selected wishes) with commitment, or loosely, wishes that are committed to. Here the commitment acts as a filter.

The third one says that the agent really does the primitive actions that s/he intends to do. This means that if one adopts this as an axiom the agent is not allowed to do something else (first). (In our opinion this is rather strict on the agent, since it may well be that postponing its intention for a while is also an option.) On the other hand, as Rao & Georgeff say, the agent may also do things that are not intended since the converse does not hold. And also nothing is said about the intention to do complex actions.

The fourth, fifth and seventh express that the agent is conscious of its intentions, goals and what primitive action he has done in the sense that he believes what he intends, has as a goal and what primitive action he has just done.

The sixth one says something like that intentions are really wished for: if something is an intention then it is a goal that it is an intention.

The eighth formula states that intentions will inevitably (in every possible future) be dropped eventually, so there is no infinite deferral of its intentions. This leaves open, whether the intention will be fulfilled eventually, or will be given up for other reasons. Below we will discuss several possibilities of giving up intentions according to different types of commitment an agent may have.

BDI-logical expressions can be used to characterize different types of agents. Rao & Georgeff mention the following possibilities:

1. (blindly committed agent) $INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow$ $\mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi) \mathsf{U} BEL(\varphi))$
2. (single-minded committed agent) $INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow$ $\mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi) \mathsf{U} (BEL(\varphi) \vee \neg BEL(\mathit{optional} \diamond \varphi)))$
3. (open minded committed agent) $INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow$ $\mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi) \mathsf{U} (BEL(\varphi) \vee \neg GOAL(\mathit{optional} \diamond \varphi)))$

A blindly committed agent maintains his intentions to inevitably obtaining eventually something until he actually believes that that something has been fulfilled. A single-minded committed agent is somewhat more flexible: he maintains his intention until he believes he has achieved it *or he does not believe that it can be reached (i.e. that it is still an option in some future) anymore*. Finally, the open minded committed agent is even more flexible: he can also drop his intention if it is not a goal (desire) anymore.

Rao & Georgeff are then able to obtain results under which conditions the various types of committed agents will reach their intentions. For example, for a blindly committed agent it holds that under the assumption of the axioms we have discussed earlier plus an axiom that expresses no infinite deferral of intentions:

$$INTEND(\varphi) \rightarrow \mathit{inevitable} \diamond \neg INTEND(\varphi)$$

that

$$INTEND(\mathit{inevitable}(\diamond \varphi)) \rightarrow \mathit{inevitable}(\diamond BEL(\varphi))$$

expressing that if the agent intends to eventually obtain φ it will inevitably eventually believe that it has succeeded in achieving φ.

The branching-time setup of the approach as opposed to a linear-time one is much more expressive and is shown to solve problems such as the *Little Nell problem* [37, 13, 45]. This is about a girl, Little Nell, that is in mortal peril, and a rescue agent that reasons like this: I intend to rescue Little Nell, and therefore I believe (because I'm confident that my actions will succeed) that she will be safe, but then I can drop my intention to rescue her just because she will be safe...! In a linear-time approach – if one is not very careful – this scenario results in a contradictory (or unintuitive) representation (basically because there is only one future in which apparently Little Nell will be safe), while in a branching-time approach such as Rao and Georgeff's this presents no problem at all (cf. [13, 45]. In fact in CTL_{BDI} the scenario comes down to something like (here φ stands for "Little Nell is safe")

$$INTEND(\mathit{inevitable} \diamond \varphi) \rightarrow \mathit{inevitable}(INTEND(\mathit{inevitable} \diamond \varphi) \mathsf{U} BEL(\mathit{optional} \diamond \varphi))$$

informally saying that since the agent believes that there is a way (by performing its plan) to eventually reaching the goal φ, it may drop its intention to perform the plan to achieve eventually φ, which is definitely not valid in CTL_{BDI}! Intuitively, this is the case, because there may be other branches along which Little Nell will not be safe, so that there is no reason to give up the intention to rescue her.

In the next section we will look at yet another approach, based on (non-linear) dynamic logic, which may perhaps be viewed as an amalgam of those of Cohen & Levesque (using dynamic logic) but allowing for non-linear, i.e. branching, structures.

3 KARO Logic

In this section we review the KARO formalism, in which *action*, together with knowledge / belief, is the primary concept, on which other agent notions are built. The KARO framework has been developed in a number of papers (e.g. [32, 33, 27, 42]) as well as the thesis of Van Linder ([31]). Historically, the KARO approach was the first approach truly based on dynamic logic, although as we have seen, in retrospect, we may view Cohen & Levesque's approach as being based on a linear variant of PDL (Propositional Dynamic Logic). There are differences, though. We will see that in KARO the fact that it is based on a logic of action is even more employed than in Cohen & Levesque: besides BDI-like notions such as knowledge, belief, desires, and goals that are operators that take formulas as arguments, and are quite similar in nature as the notions that are in Cohen & Levesque's approach, in KARO there are also operators taking actions as arguments such as ability and commitment, and operators that take both actions and formulas as arguments, such as a Can operator and a (possible) intention operator. All these operators are used to describe the mental state of the agent. But even more importantly, in the KARO framework (dedicated) actions are used to *change* the mental state of the agent. So there are revise, commit and uncommit actions to revise beliefs and update the agenda (the commitments) of the agent. In this sense KARO is related to dynamic epistemic logic [18], treated elsewhere in this handbook.

KARO logic for rational agents

The KARO formalism is an amalgam of dynamic logic and epistemic / doxastic logic [41], augmented with several additional (modal) operators in order to deal with the motivational aspects of agents. So, besides operators for knowledge (**K**), belief (**B**) and action ($[\alpha]$, "after performance of α it holds that"), there are additional operators for ability (**A**) and desires (**D**).

Assume a set $\mathcal{A}$ of atomic actions and a set $\mathcal{P}$ of atomic propositions.

Definition 3.1 *(Language.) The language $\mathcal{L}_{KARO}$ of KARO-formulas is given by the BNF grammar:*

$$\begin{array}{rcl}\varphi & ::= & p(\in \mathcal{P}) \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \ \ldots \\ & & \mathbf{K}\varphi \mid \mathbf{B}\varphi \mid \mathbf{D}\varphi \mid [\alpha]\varphi \mid \mathbf{A}\alpha\end{array}$$

$$\alpha \quad ::= \quad a(\in \mathcal{A}) \mid \varphi? \mid \alpha_1 ; \alpha_2 \mid \alpha_1 + \alpha_2 \mid \alpha^*$$

Here the formulas generated by the second (α) part are referred to as actions (or rather action expressions). We use the abbreviations $\mathtt{tt} \equiv p \vee \neg p$ (for some fixed $p \in \mathcal{P}$) and $\mathtt{ff} \equiv \neg\mathtt{tt}$. Conditional and while-action are introduced by the usual abbreviations: $\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi} \equiv (\varphi?;\alpha_1) + (\neg\varphi?;\alpha_2)$ and $\mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od} \equiv (\varphi?;\alpha)^*;\neg\varphi?$.

Thus formulas are built by means of the familiar propositional connectives and the modal operators for knowledge, belief, desire, action and ability. Actions are the familiar ones from imperative programming: atomic ones, tests, sequential composition, (nondeterministic) choice and repetition.

Definition 3.2 *(KARO models.)*

1. *The semantics of the knowledge, belief and desires operators is given by means of Kripke structures of the following form: $\mathcal{M} = \langle W, \vartheta, R_K, R_B, R_D \rangle$, where*
   - *W is a non-empty set of states (or worlds)*
   - *ϑ is a truth assignment function per state*
   - *R_K, R_B, R_D are accessibility relations for interpreting the modal operators $\mathbf{K}, \mathbf{B}, \mathbf{D}$. The relation R_K is assumed to be an equivalence relation, while the relation R_B is assumed to be euclidean, transitive and serial. Furthermore we assume that $R_B \subseteq R_K$. No special constraints are assumed for the relations R_D.*
2. *The semantics of actions is given by means of structures of type $\langle \Sigma, \{R_a \mid a \in \mathcal{A}\}, \mathcal{C}, Ag \rangle$, where*
   - *Σ is the set of possible model/state pairs (i.e. models of the above form, together with a state appearing in that model)*
   - *R_a $(a \in \mathcal{A})$ are relations on Σ encoding the behaviour of atomic actions*
   - *$\mathcal{C}$ is a function that gives the set of actions that the agent is able to do per model/state pair*
   - *Ag is a function that yields the set of actions that the agent is committed to (the agent's 'agenda') per model/state pair.*

We have elements in the structures for interpreting the operators for knowledge, belief, and desire. Actions are modelled as model/state pair transformers to emphasize their influence on the mental state (that is, the complex of knowledge, belief and desires) of the agent rather than just the state of the world. Both (cap)abilities and commitments are given by functions that yield the relevant information per model / state pair.

Definition 3.3 *(Interpretation of formulas.) In order to determine whether a formula $\varphi \in \mathcal{L}$ is true in a model/state pair (M, w) (if so, we write $(M, w) \models \varphi$), we stipulate:*

- *$\mathcal{M}, w \models p$ iff $\vartheta(w)(p) = \text{true}$, for $p \in \mathcal{P}$*
- *The logical connectives are interpreted as usual.*
- *$\mathcal{M}, w \models \mathbf{K}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all w' with $R_K(w, w')$*

- $\mathcal{M}, w \models \mathbf{B}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R_B(w, w')$
- $\mathcal{M}, w \models \mathbf{D}\varphi$ *iff* $\mathcal{M}, w' \models \varphi$ *for all* w' *with* $R_D(w, w')$
- $\mathcal{M}, w \models [\alpha]\varphi$ *iff* $\mathcal{M}', w' \models \varphi$ *for all* M', w' *with* $R_\alpha((\mathcal{M}, w), (\mathcal{M}', w'))$
- $\mathcal{M}, w \models \mathbf{A}\alpha$ *iff* $\alpha \in \mathcal{C}(\mathcal{M}, w)$[5]
- $\mathcal{M}, w \models \mathbf{Com}(\alpha)$ *iff* $\alpha \in Ag(\mathcal{M}, w)$[6]

Here R_α is defined as usual in dynamic logic by induction from the basic case R_a (cf. e.g. [21, 31, 27], but now on model/state pairs rather than just states). So, e.g. $R_{\alpha_1+\alpha_2} = R_{\alpha_1} \cup R_{\alpha_2}$, $R_{\alpha^*} = R^*_\alpha$, the reflective transitive closure of R_α, and $R_{\alpha_1;\alpha_2}$ is the relational product of R_{α_1} and R_{α_2}. Likewise the function $\mathcal{C}$ is lifted to complex actions ([31, 27]). We call an action α *deterministic* if $\#\{w' \mid R_\alpha(w, w')\} \leqslant 1$ for any $w \in W$, and *strongly deterministic* if $\#\{w' \mid R_\alpha(w, w')\} \leqslant 1$. (Here # stands for cardinality.)

We have clauses for knowledge, belief and desire. The action modality gets a similar interpretation: something (necessarily) holds after the performance / execution of action α if it holds in all the situations that are accessible from the current one by doing the action α. The only thing which is slightly nonstandard is that, as stated above, a situation is characterised here as a model / state pair. The interpretations of the ability and commitment operators are rather trivial in this setting (but see the footnotes): an action is enabled (or rather: the agent is able to do the action) if it is indicated so by the function C, and, likewise, an agent is committed to an action α if it is recorded so in the agent's agenda.

Furthermore, we will make use of the following syntactic abbreviations serving as auxiliary operators:

Definition 3.4

- *(dual)* $\langle\alpha\rangle\varphi = \neg[\alpha]\neg\varphi$*, expressing that the agent has the opportunity to perform* α *resulting in a state where* φ *holds.*
- *(opportunity)* $\mathbf{O}\alpha = \langle\alpha\rangle\mathtt{tt}$*, i.e., an agent has the opportunity to do an action iff there is a successor state w.r.t. the* R_α*-relation;*
- *(practical possibility)* $\mathbf{P}(\alpha, \varphi) = \mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$*, i.e., an agent has the practical possibility to do an action with result* φ *iff it is both able and has the opportunity to do that action and the result of actually doing that action leads to a state where* φ *holds;*
- *(can)* $\mathbf{Can}(\alpha, \varphi) = \mathbf{KP}(\alpha, \varphi)$*, i.e., an agent can do an action with a certain result iff it knows it has the practical possibilty to do so;*
- *(realisability)* $\Diamond\varphi = \exists a_1, \ldots, a_n \mathbf{P}(a_1; \ldots; a_n, \varphi)$[7]*, i.e., a state property* φ *is realisable iff there is a finite sequence of atomic actions of which the agent has the practical possibility to perform it with the result* φ*;*
- *(goal)* $\mathbf{G}\varphi = \neg\varphi \wedge \mathbf{D}\varphi \wedge \Diamond\varphi$*, i.e., a goal is a formula that is not (yet) satisfied, but desired and realisable.*[8]

[5] In [28] we have shown that the ability operator can alternatively defined by means of a second accessibility relation for actions, in a way analogous to the opportunity operator below.

[6] The agenda is assumed to be closed under certain conditions such as taking 'prefixes' of actions (representing initial computations). Details are omitted here, but can be found in [42].

[7] We abuse our language here slightly, since strictly speaking we do not have quantification in our object language. See [42] for a proper definition.

[8] In fact, here we simplify matters slightly. In [42] we also stipulate that a goal should be explicitly selected somehow from the desires it has, which is modelled in that paper by means of an additional modal operator. Here we leave this out for simplicity's sake.

- *(possible intend)* $\mathbf{I}(\alpha,\varphi) = \mathbf{Can}(\alpha,\varphi) \wedge \mathbf{KG}\varphi$, *i.e., an agent (possibly) intends an action with a certain result iff the agent can do the action with that result and it moreover knows that this result is one of its goals.*

Remark 3.5

- *The dual of the (box-type) action modality expresses that there is at least a resulting state where a formula* φ *holds. It is important to note that in the context of* deterministic *actions, i.e. actions that have at most one successor state, this means that the* only *state satisfies* φ*, and is thus in this particular case a stronger assertion than its dual formula* $[\alpha]\varphi$*, which merely states that if there are any successor states they will (all) satisfy* φ*. Note also that if atomic actions are assumed to be deterministic all actions including the complex ones will be deterministic.*
- *Opportunity to do an action is modelled by having at least one successor state according to the accessibility relation associated with the action.*
- *Practical possibility to to an action with a certain result is modelled as having both ability and opportunity to do the action with the appropriate result. Note that* $\mathbf{O}\alpha$ *in the formula* $\mathbf{A}\alpha \wedge \mathbf{O}\alpha \wedge \langle\alpha\rangle\varphi$ *is actually redundant since it already follows from* $\langle\alpha\rangle\varphi$*. However, to stress the opportunity aspect it is added.*
- *The Can predicate applied to an action and formula expresses that the agent is 'conscious' of its practical possibility to do the action resulting in a state where the formula holds.*
- *A formula* φ *is realisable if there is a 'plan' consisting of (a sequence of) atomic actions of which the agent has the practical possibility to do them with* φ *as a result.*
- *A formula* φ *is a goal in the KARO framework if it is not true yet, but desired and realisable in the above meaning, that is, there is a plan of which the agent has the practical possibility to realise it with* φ *as a result.*
- *An agent is said to (possibly) intend an action* α *with result* φ *if it 'Can' do this (knows that it has the practical possibility to do so), and, moreover, knows that* φ *is a goal.*

In order to manipulate both knowledge / belief and motivational matters special actions `revise`, `commit` and `uncommit` are added to the language. (We assume that we cannot nest these operators. So, e.g., `commit(uncommit`α`)` is not a well-formed action expression. For a proper definition of the language the reader is referred to [42].) The semantics of these are again given as model/state transformers (We only do this here in a very abstract manner, viewing the accessibility relations associated with these actions as functions. For further details we refer to e.g. [31, 27, 42]):

Definition 3.6 *(Accessibility of revise, commit and uncommit actions.)*

1. $R_{\mathtt{revise}\varphi}(\mathcal{M},w) = \mathit{update\_belief}(\varphi,(\mathcal{M},w))$.
2. $R_{\mathtt{commit}\alpha}(\mathcal{M},w) = \mathit{update\_agenda}^{+}(\alpha,(\mathcal{M},w))$*, if* $\mathcal{M},w \models \mathbf{I}(\alpha,\varphi)$ *for some* φ*, otherwise* $R_{\mathtt{commit}\alpha}(\mathcal{M},w) = \emptyset$ *(indicating failure of the commit action).*
3. $R_{\mathtt{uncommit}\alpha}(\mathcal{M},w) = \mathit{update\_agenda}^{-}(\alpha,(\mathcal{M},w))$*, if* $\mathcal{M},w \models \mathbf{Com}(\alpha)$*, otherwise* $R_{\mathtt{uncommit}\alpha}(\mathcal{M},w) = \emptyset$ *(indicating failure of the uncommit action);*
4. $\mathtt{uncommit}\alpha \in \mathcal{C}(\mathcal{M},w)$ *iff* $\mathcal{M},w \models \neg\mathbf{I}(\alpha,\varphi)$ *for all formulas* φ*, that is, an agent is able to uncommit to an action if it is not intended to do it (any longer) for any purpose.*

Here $\mathit{update\_belief}$, $\mathit{update\_agenda}^{+}$ and $\mathit{update\_agenda}^{-}$ are functions that update the agent's belief and agenda (by adding or removing an action), respectively. Details are omitted here, but essentially these actions are model/state transformers again, representing a change of the mental state of the agent (regarding beliefs and commitements, respectively). The *update_belief*$(\varphi, (\mathcal{M}, w))$ function changes the model $\mathcal{M}$ in such a way that the agent's belief is updated with the formula φ, while *update_agenda*$^{+}(\alpha, (\mathcal{M}, w))$ changes the model $\mathcal{M}$ such that α is added to the agenda, and likewise for the *update_agenda*$^{-}$ function, but now with respect to removing an action from the agenda. The formal definitions can be found in [32, 33] and [42]. The `revise` operator can be used to cater for revisions due to observations and communication with other agents, which we will not go into further here (see [33]).

The interpretation of formulas containing revise and (un)commit actions is now done using the accessibility relations above. One can now define validity as usual with respect to the KARO-models. One then obtains the following validities (of course, in order to be able to verify these one should use the proper model and not the abstraction / simplification we have presented here.) Typical properties of this framework, called the KARO logic, include (cf. [32, 42]):

Proposition 3.7

1. $\models \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$, *for* $\Box \in \{\mathbf{K}, \mathbf{B}, \mathbf{D}, [\alpha]\}$
2. $\models \langle\alpha\rangle\varphi \to [\alpha]\varphi$, *for deterministic* α
3. $\models \Box\varphi \to \Box\Box\varphi$, *for* $\Box \in \{\mathbf{K}, \mathbf{B}\}$
4. $\models \neg\Box\varphi \to \Box\neg\Box\varphi$, *for* $\Box \in \{\mathbf{K}, \mathbf{B}\}$
5. $\models \mathbf{K}\varphi \to \varphi$
6. $\models \neg\mathbf{B}\mathtt{ff}$
7. $\models \mathbf{O}(\alpha;\beta) \leftrightarrow \langle\alpha\rangle\mathbf{O}\beta$
8. $\models \mathbf{Can}(\alpha;\beta,\varphi) \leftrightarrow \mathbf{Can}(\alpha, \mathbf{P}(\beta,\varphi))$
9. $\models \mathbf{I}(\alpha,\varphi) \to \mathbf{K}\langle\alpha\rangle\varphi$
10. $\models \mathbf{I}(\alpha,\varphi) \to \langle\mathtt{commit}\alpha\rangle\mathbf{Com}(\alpha)$
11. $\models \mathbf{I}(\alpha,\varphi) \to \neg\mathbf{A}\mathtt{uncommit}(\alpha)$
12. $\models \mathbf{Com}(\alpha) \to \langle\mathtt{uncommit}(\alpha)\rangle\neg\mathbf{Com}(\alpha)$
13. $\models \mathbf{Com}(\alpha) \wedge \neg\mathbf{Can}(\alpha,\top) \to \mathbf{Can}(\mathtt{uncommit}(\alpha), \neg\mathbf{Com}(\alpha))$
14. $\models \mathbf{Com}(\alpha) \to \mathbf{K}\mathbf{Com}(\alpha)$
15. $\models \mathbf{Com}(\alpha_1;\alpha_2) \to \mathbf{Com}(\alpha_1) \wedge \mathbf{K}[\alpha_1]\mathbf{Com}(\alpha_2)$
16. $\models \mathbf{Com}(\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}) \wedge \mathbf{K}\varphi \to \mathbf{Com}(\varphi?;\alpha_1)$
17. $\models \mathbf{Com}(\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}) \wedge \mathbf{K}\neg\varphi \to \mathbf{Com}(\neg\varphi?;\alpha_2)$
18. $\models \mathbf{Com}(\mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}) \wedge \mathbf{K}\varphi \to \mathbf{Com}((\varphi?;\alpha);\mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od})$

The first of these properties says that all the modalities mentioned are 'normal' in the sense that they are closed under implication. The second states that the dual operator $\langle\alpha\rangle$ is stronger than the operator $[\alpha]$ in case the action α is deterministic: if there is at most one successor state after performing α and we know that there is at least one successor state satisfying φ then *all* successor states satisfy φ. The third and fourth properties are the so-called introspection properties for knowledge and belief. The fifth property says that knowledge is true, while the sixth states that belief (may not be true but) is not inconsistent. The seventh property states that having the opportunity to do a sequential composition of two actions amounts to having the opportunity of doing the first action first and then having the opportunity to do the second. The eighth states that an agent that *can* do a sequential composition of two actions with result φ iff the agent can do the first actions resulting in a state where it has the practical possibility to do the second with φ as result. The ninth states that if one possibly intends to do α

with result φ then one knows that there is a possibility of performing α resulting in a state where φ holds. The tenth asserts that if an agent possibly intends to do α with some result φ, it has the opportunity to commit to α with result that it is committed to α (i.e. α is put into its agenda). The eleventh says that if an agent intends to do α with a certain purpose, then it is unable to uncommit to it (so, if it is committed to α it has to persevere with it). This is the way persistence of commitment is represented in KARO. Note that this is much more 'concrete' (also in the sense of computability) than the persistence notions in the other approaches we have seen, where temporal operators pertaining to a possibly infinite future were employed to capture them...! We think it is no coincidence that in [25] an almost perfect match in the sense of a correspondence was found between the agent programming language GOAL and Cohen & Levesque's logic of intention, the main difference being the inability of GOAL to express the persistence properties of intentions in this logic...!) In KARO we have the advantage of having dedicated actions in the action language dealing with the change of commitment that can be used to express persistence without referring to the (infinite) future, rendering the notion of persistence much 'more computable'. The twelfth property says that if an agent is committed to an action and it has the opportunity to uncommit to it with as result then indeed the commitment is removed. The thirteenth says that whenever an agent is committed to an action that is no longer known to be practically possible, it knows that it can undo this impossible commitment. The fourteenth property states that commitments are known to the agent. The last four properties have to do with commitments to complex actions. For instance, the fifteenth says that if an agent is committed to a sequential composition of two actions then it is committed to the first one, and it knows that after doing the first action it will be committed to the second action.

KARO logic for emotional agents

In this subsection we will look at a nice, recent application of BDI logic that deals with agent behaviour that is strictly beyond the scope of the original aim of BDI logic, viz. describing the behaviour of rational agents. We will sketch how the KARO framework can be used for describing emotional agents. Although it is perhaps a bit paradoxical to describe emotions and emotional behaviour with logic, one should bear in mind that we are dealing with behaviour here, and this can be described in logic, especially a logic that deals with actions such as the KARO framework. Furthermore, as we shall see, emotional behaviour will turn out to be complimentary rather than opposed to rational behaviour of agents, something that is also acknowledged by recent work in cognitive science [14]. We have actually looked at two psychological theories: that of Oatley & Jenkins [43] and that of OCC [44]. Since the latter is much more involved (treating 22 emotions, while the former only treats 4 basic emotions), we here mainly follow the ideas of Oatley & Jenkins, but say a few words on modelling OCC later.

According to [43] the 4 basic emotions, happiness, sadness, anger and fear, have the following characteristics:

- Happiness results when in the process of trying to achieve a goal, things go 'right', as expected, i.e. subgoals are achieved thus far.
- Sadness results when in the process of trying to achieve a goal, things go 'wrong', i.e. not as expected, i.e. subgoals are not being achieved.
- Anger is the result of frustration about not being able to execute the current plan, and makes the agent try harder to execute the plan.
- Fear results when a 'maintenance goal' is threatened, so that the agent will make sure that this maintenance goal is restored before going on with other activities.

It is directly obvious from these descriptions that these emotions are BDI-related notions! So it is not so strange to use a BDI-logic to describe them. KARO was chosen for this in [38, 39].

Let's take sadness as an example. For simplicity, assume that plans consist of sequences of atomic actions. In KARO we can then express the trigger condition for sadness as follows:

$$\mathbf{I}(\pi, \varphi) \wedge \mathbf{Com}(\pi) \wedge \mathbf{B}([\alpha]\psi) \rightarrow$$

$$[\alpha]((\mathbf{B}\neg\psi \wedge \mathbf{Com}(\pi\backslash\alpha)) \rightarrow \mathit{sad}(\pi\backslash\alpha, \varphi))$$

where α is a prefix of plan π. Inttuitively, this says that if the agent has the (possible) intention to perform plan π with goal φ, it is committed to π (so it has a true intention to do π), and it believes that after doing the initial fragment α of the plan π it holds that ψ, then after doing α if it believes that ψ does not hold while it is still committed to the rest of the plan, it is sad (with respect to the rest of the plan and goal φ). In a similar way the trigger conditions of the other emotions can be formalised [38, 39]. Later we turned to the much more complex framework of OCC [50, 51]. Particularly in the latter publication we show how to formalise the (trigger conditions of) emotions in OCC in three steps: first we present a more general logical structure of the emtions, which are later refined in terms of doxastic logic and finally in the full-blown BDI logic KARO again. The way emtions get a semantics based on BDI models is quite intricate and beyond the scope of the present paper, but one of the properties that can be proven valid in this approach is the following, using KARO's (possible) intend operator (here parametrized by an agent):

$$\mathbf{I}_i(\alpha, \pi) \rightarrow [i:\alpha](\mathit{Pride}^T_j(i:\alpha) \wedge \mathit{Joy}^T_i(\varphi) \wedge \mathit{Gratification}^T_i(i:\alpha, \varphi))$$

(Here $i:\alpha$ in the dynamic logic box refers to the action of i performing α, and the superscript ^{T} placed at the emotion operators pertains to the idea that we are considering triggering / elicitation forms of the emotions concerned.) Informally this reads that if the agent i has the possble intention to do α with goal φ, then if he has performed α he is proud (triggered pride) of his action, has (triggered) joy about the achievement of the goal φ and has (triggered) gratification with respect to action $i:\alpha$ and goal φ, which makes sense, and particularly so in the light of the OCC theory.

Finally, let us also mention here the strongly related work by Adam et al. [1, 2]. Also this work is devoted to a formalisation of OCC emotions in BDI terms. There are differences with the work of Steunebrink et al., though. For instance, in [1] Adam simply defines joy as a conjunction of belief and desire: $\mathit{Joy}\ \varphi =_{def} \mathit{Bel}\ \varphi \wedge \mathit{Des}\ \varphi$. This seems to express a 'state of joy' (experience) rather than a trigger for joy. This raises a confusion of emotion elicitation (triggering) and experience, which is kept separately in the approach of Steunebrink. This confusion also appears at other places in the work of Adam, e.g. where she defines gratification as the conjunction of pride (which pertains to triggering) and joy (which is about experience as we saw earlier). In [2] this issue is improved upon and it is explained that the above definition of Joy is solely about the triggering of of joy, not the experience. However, the confusion of triggering versus experience is still not resolved completely since in that paper it is still present in the introspective properties of emotional awareness $\mathit{Emotion}\varphi \leftrightarrow \mathit{BelEmotion}\varphi$ and $\neg\mathit{Emotion}\varphi \leftrightarrow \mathit{Bel}\neg\mathit{Emotion}\varphi$, which hold in that framework for any Emotion (Theorem 13 of [2]). This is counterintuitive if Emotion should capture the triggering of the associated emotion, since an agent may not be aware of this triggering.

4 BDI-modalities in *stit* logic

In the 90s, the standard knowledge-based agent paradigm from the 50s saying that agents need to know relevant aspects of their environments in order to be able to behave intelligently, was refined to the BDI-paradigm, which pictures the logic of agentive decision making as an interplay between an agent's Beliefs, Desires and Intentions. The principles of BDI logics reflect rationality postulates for agent modalities. In particular, these principles model how B, D and I modalities interact with each other over time (well known are the so called 'commitment strategies' [13] stating under which belief and desire conditions intentions have to be dropped). BDI logics are not meant for knowledge representation but for agent specification: ideally concretely built agents will some day be verified against the logic principles of BDI-logics (how exactly this could ever be done is a question we set aside here).

An essential component of any BDI logic is then its dynamic part. Traditionally, either the dynamic part is formed by a dynamic logic fragment (Cohen & Levesque system [13], KARO [27, 42]) or a temporal fragment (Rao & Georgeff [45]). Recently a third alternative has been considered: *stit* (seeing to it that) logic. *Stit* logics can be said to be in between dynamic logic and temporal logic. Where dynamic logic sees actions as the steps of a program, and temporal logic leaves actions entirely out of the picture, *stit* logic sees action as a relation between agents and the effects they can see to. *Stit* logic achieves this by generalizing temporal structures to choice structures. The most distinguishing feature of *stit* logic is that truth of formulas often expresses information about the dynamics of the world. For instance, the *stit* logic formula $[ag\ \mathtt{stit}]X(at\_station))$ says that agent *ag* sees to it that next it is at the station. It does not say that the agent *can* see to it that next it is at the station (this is however a logical consequence), which would be a truth concerning a static condition.

In the present section we will discuss how in recent years several authors have aimed to combine *stit* logic and BDI notions. There are two parts. In the first part, section 4.1, we focus on classical instantaneous *stit* logics and the BDI extensions that have been suggested for them. In the second part, section 4.2, we consider dynamic variants of the BDI modalities and discuss the notion of 'knowingly doing' within a version of *stit* where effects of actions take effect in next states: XSTIT.

4.1 BDI modalities in instantaneous stit

Traditionally *stit* logics encompass operators for agency that assume that an agentive choice exertion is something that takes no time. So, an instantaneous stit operator $[ag\ \mathtt{stit}]\varphi$ typically obeys the success axiom $[ag\ \mathtt{stit}]\varphi \rightarrow \varphi$ to capture the intuition concerning instantaneity saying that if *ag now* sees to it that φ holds, then φ must be true *now*. Before putting forward an alternative to this view, where the central agency operator has a built-in step to a next moment in time, we give the formal definition of standard (Chellas) instantaneous *stit* logic and discuss its logic properties. We will here use a slightly different syntax and semantics than used by Chellas himself [12] and by Horty [30], but, the logic is the same.

CSTIT

Definition 4.1 *Given a countable set of propositions P and $p \in P$, and given a finite set Ags of agent names, and $ag \in Ags$, the formal language $\mathcal{L}_{\mathsf{CSTIT}}$ is:*

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [ag\ \mathtt{Cstit}]\varphi$$

Besides the usual propositional connectives, the syntax of CSTIT comprises two modal operators. The operator $\Box\varphi$ expresses 'historical necessity', and plays the same role as the well-known path quantifiers in logics such as CTL and CTL* [19]. Another way of talking about this operator is to say that it expresses that φ is 'settled'. We abbreviate $\neg\Box\neg\varphi$ by $\Diamond\varphi$. The operator $[ag\ \mathtt{Cstit}]\varphi$ stands for 'agent *ag* sees to it that φ' (the 'c' referring to Chellas). $\langle ag\ \mathtt{Cstit}\rangle\varphi$ abbreviates $\neg[ag\ \mathtt{Cstit}]\neg\varphi$.

Definition 4.2 *A CSTIT-frame is a tuple $\langle S, H, R_{ag}\rangle$ such that*[9]*:*

1. *S is a non-empty set of static states. Elements of S are denoted s, s', etc.*
2. *H is a non-empty set of possible system histories $\ldots s_{-2}, s_{-1}, s_0, s_1, s_2, \ldots$ with $s_x \in S$ for $x \in \mathbb{Z}$. Elements of H are denoted h, h', etc.*
3. *Dynamic states are tuples $\langle s, h\rangle$, where $s \in S$, $h \in H$ and s appears on h. Now the relations R_{ag} are 'effectivity' equivalence classes over dynamic states such that $\langle s, h\rangle R_{ag}\langle s', h'\rangle$ only if $s = s'$. For any state s and agent ag, the relation R_{ag} defines a partition of the dynamic states built with s. The partition models the possible choices $C^s_{ag}, {C'}^s_{ag}, {C''}^s_{ag}, \ldots$ of ag in s. A choice profile $\langle C^s_{ag_1}, C^s_{ag_2} \ldots C^s_{ag_n}\rangle$ at s is a particular combination of choices $C^s_{ag_i}$ at s, one for each agent ag_i in the system. For any s the intersection of choices in any choice profile is non-empty: $\bigcap\limits_{ag_i \in Ags} C^s_{ag_i} \neq \emptyset$*

In definition 4.2 above, we refer to the states s as 'static states'. This is to distinguish them from 'dynamic states', which are combinations $\langle s, h\rangle$ of static states and histories. Dynamic states function as the elementary units of evaluation of the logic. This means that the basic notion of 'truth' in the semantics of this logic is about dynamic conditions concerning choice exertions.

We now define models by adding a valuation of propositional atoms to the frames of definition 4.2.

Definition 4.3 *A frame $\mathcal{F} = \langle S, H, R_{ag}\rangle$ is extended to a model $\mathcal{M} = \langle S, H, R_{ag}, V\rangle$ by adding a valuation V of atomic propositions:*

- *V is a valuation function $V : P \longrightarrow 2^{S \times H}$ assigning to each atomic proposition the set of state history pairs relative to which they are true.*

We evaluate truth with respect to dynamic states.

Definition 4.4 *Relative to a model $\mathcal{M} = \langle S, H, R_{ag}, V\rangle$, truth $\langle s, h\rangle \models \varphi$ of a formula φ in a dynamic state $\langle s, h\rangle$, with $s \in h$, is defined as:*

$$\begin{array}{lcl}
\langle s,h\rangle \models p & \Leftrightarrow & \langle s,h\rangle \in V(p) \\
\langle s,h\rangle \models \neg\varphi & \Leftrightarrow & \textit{not } \langle s,h\rangle \models \varphi \\
\langle s,h\rangle \models \varphi \wedge \psi & \Leftrightarrow & \langle s,h\rangle \models \varphi \textit{ and } \langle s,h\rangle \models \psi \\
\langle s,h\rangle \models \Box\varphi & \Leftrightarrow & \forall h' : \textit{if } s \in h' \textit{ then } \langle s,h'\rangle \models \varphi \\
\langle s,h\rangle \models [ag\ \texttt{Cstit}]\varphi & \Leftrightarrow & \forall h' : \textit{if } \langle s,h\rangle R_{ag} \langle s,h'\rangle \textit{ then } \langle s,h'\rangle \models \varphi
\end{array}$$

Satisfiability, validity on a frame and general validity are defined as usual.

Now we proceed with the axiomatization.

Theorem 4.5 (Xu [56]) *The following axiom schemas, in combination with a standard axiomatization for propositional logic, and the standard rules (like necessitation) for the normal modal operators, define a complete Hilbert system for CSTIT:*

$$\begin{array}{ll}
 & \textit{The S5 axioms for } \Box \\
 & \textit{For each ag the S5 axioms for } [ag\ \texttt{Cstit}] \\
\text{(SettC)} & \Box\varphi \to [ag\ \texttt{Cstit}]\varphi \\
\text{(Indep)} & \Diamond[ag_1\ \texttt{Cstit}]\varphi \wedge \ldots \wedge \Diamond[ag_n\ \texttt{Cstit}]\psi \to \\
 & \Diamond([ag_1\ \texttt{Cstit}]\varphi \wedge \ldots \wedge [ag_n\ \texttt{Cstit}]\psi) \\
 & \textit{for } Ags = \{ag_1, \ldots, ag_n\}
\end{array}$$

[9] In the meta-language we use the same symbols both as constant names and as variable names, and we assume universal quantification of unbound meta-variables.

Balbiani et.al [4] propose an alternative axiomatization and a semantics whose units of evaluation are not two dimensional pairs $\langle s, h\rangle$ but one dimensional worlds w. Here we have chosen to give a two-dimensional semantics to emphasize the relation with the XSTIT semantics in section 4.2.

BDI-stit

Semmling and Wansing [49] add BDI modalities to a basic Chellas stit logic as the one just defined. Their BDI-stit formalism extends the syntax as follows (we take the liberty of using our own notation for the BDI operators and to define an alternative but equivalent semantics).

Definition 4.6 *Given a countable set of propositions P and $p \in P$, and given a finite set Ags of agent names, and $ag \in Ags$, the formal language $\mathcal{L}_{\textsf{bdi-stit}}$ is:*

$$\begin{aligned} \varphi \quad := \quad & p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [ag\ \texttt{Cstit}]\varphi \mid \langle[ag\ \texttt{bel}]\rangle\varphi \mid \langle[ag\ \texttt{des}]\rangle\varphi \mid \\ & \langle[ag\ \texttt{int}]\rangle\varphi \end{aligned}$$

To emphasize their weak modal character, we denote the introduced belief, desire, intention and possibility operators with a combination of sharp and square brackets. This alludes to the combination of first order existential and universal quantifications that is present in any first order simulation of a weak modal operator. The reading of the operators speaks for itself; they express belief, desire and intention concerning a proposition φ.

Definition 4.7 *A* ***bdi-stit****-frame is a tuple $\langle S, H, R_{ag}, N_b, N_d, N_i\rangle$ such that:*

1. *$\langle S, H, R_{ag}\rangle$ is a* ***CSTIT****-frame*
2. *N_b N_d and N_i are neighborhood functions of the form $N : S \times H \times Ags \mapsto 2^{2^{S\times H}}$ mapping any combination of a dynamic state $\langle s, h\rangle$ and an agent ag to a set of neighborhoods of $\langle s, h\rangle$. [49] imposes constraints on neighborhood frames that are equivalent to:*
   - **a.** *All three functions N_b, N_d and N_i obey $\emptyset \notin N(s, h, ag)$*
   - **b.** *All three functions N_b, N_d and N_i obey that if $N \in N(s, h, ag)$ and $N \subset N'$ then $N' \in N(s, h, ag)$*
   - **c.** *$N \in N_i(s, h, ag)$ and $N' \in N_i(s, h, ag)$ implies $N \cap N' \neq \emptyset$*

The intuition underlying neighborhood functions is the following. $N_b(s, h, ag)$ gives for agent ag in situation $\langle s, h\rangle$ the clusters of possible worlds (situations / dynamic states) the joint possibility of which it believes in. Since clusters and propositions correspond to each other one-to-one (modulo logic equivalence of the propositions), it will also be convenient to look at the clusters or neighborhoods as propositions and to say that if $N \in N_b(s, h, ag)$ the agent ag beliefs the proposition (modulo logical equivalence) corresponding to N, that $N \in N_d(s, h, ag)$ holds if ag desires the proposition and that $N \in N_i(s, h, ag)$ holds if ag intends the proposition.

Now **a.** says that there is no belief, desire or intention for impossible states of affairs. **b.** says that belief, desire and intention are closed under weakening of the propositions believed, desired or intended. **c.** says that intentions are consistent in the sense that it is not possible to hold at the same time an intention for a proposition and for its negation.

Definition 4.8 (Truth conditions BDI operators) *Relative to a model $\langle S, H, R_{ag}, N_b, N_d, N_i, V\rangle$, truth of belief, desire and intention operators is defined as ($[\![\varphi]\!]$ is the truth set of φ, that is, the subset of all dynamic elements in $S \times H$ satisfying φ):*

$$\begin{gathered} \langle s, h\rangle \models \langle[ag\ \texttt{bel}]\rangle\varphi \Leftrightarrow [\![\varphi]\!] \in N_b(s, h, ag) \\ \langle s, h\rangle \models \langle[ag\ \texttt{des}]\rangle\varphi \Leftrightarrow [\![\varphi]\!] \in N_d(s, h, ag) \\ \langle s, h\rangle \models \langle[ag\ \texttt{int}]\rangle\varphi \Leftrightarrow [\![\varphi]\!] \in N_i(s, h, ag) \end{gathered}$$

An axiomatization of the probabilistic epistemic logic is obtained by formulating axioms corresponding to the conditions on neighborhood functions.

Theorem 4.9 (Hilbert system BDI operators) *Relative to the semantics following from definitions 4.7 and 4.8 we define the following Hilbert system. We assume the standard derivation rules for the weak modalities, like closure under logical equivalence.*

$$\begin{array}{llll}
\text{(BelPos)} & \neg\langle[ag\ \texttt{bel}]\rangle\bot & \text{(BelWk)} & \langle[ag\ \texttt{bel}]\rangle\varphi \to \langle[ag\ \texttt{bel}]\rangle(\varphi \vee \psi) \\
\text{(DesPos)} & \neg\langle[ag\ \texttt{des}]\rangle\bot & \text{(DesWk)} & \langle[ag\ \texttt{des}]\rangle\varphi \to \langle[ag\ \texttt{des}]\rangle(\varphi \vee \psi) \\
\text{(IntPos)} & \neg\langle[ag\ \texttt{int}]\rangle\bot & \text{(IntWk)} & \langle[ag\ \texttt{int}]\rangle\varphi \to \langle[ag\ \texttt{int}]\rangle(\varphi \vee \psi) \\
 & & \text{(IntD)} & \langle[ag\ \texttt{int}]\rangle\varphi \to \neg\langle[ag\ \texttt{int}]\rangle\neg\varphi
\end{array}$$

Within the definitions of their own semantics, Semmling and Wansing prove completeness of their logic. Here the completeness of the axiomatization relative to the frames of definition 4.7 follows from general results in neighborhood semantics and monotonic modal logic [20]. We can check that the conditions on the frames correspond one-to-one with the axioms in the axiomatization.

As can be seen from the axioms and conditions we have shown above, Semmling and Wansing chose to make their BDI-stit logic rather weak, trying to commit only to a minimum of logical properties. But even with this minimalistic approach there is room for debate. For instance, the condition on intentions only looks at pairwise consistency of intentions, but conflicts are still possible in case there is a combination of three intentions: $\{\langle[ag\ \texttt{int}]\rangle\varphi, \langle[ag\ \texttt{int}]\rangle(\varphi \to \psi), \langle[ag\ \texttt{int}]\rangle\neg\psi\}$ is satisfiable. If we do not want that, it is straightforward to adapt the condition on neighborhood functions (demand that any finite number of neighborhoods has a state in common), but it is unclear how to axiomatize it.

Even though the *stit* framework's notion of truth refers to the dynamics of a system of agents at given states (which, for this reason, we call 'dynamic states'), Semmling and Wansing do not focus on dynamic interpretations of the BDI attitudes. A formula like $\langle[ag\ \texttt{bel}]\rangle[ag\ \texttt{Cstit}]\varphi$ must express something like "ag believes that it sees to it that φ", but the inherent dynamic aspect of this notion is not analyzed. In particular, no interactions between *stit* and BDI modalities are studied. This is likely to be due to the fact that explicit dynamic temporal operators are absent in the logic and because agency is instantaneous. In section 4.2 we will report on the study of the inherent dynamic aspect of such combinations of operators.

BDI, *stit* and regret

In [36] Lorini and Schwarzentruber use *stit* logic as the basis for investigations into what they call counterfactual emotions[10]. The typical counterfactual emotion is 'regret'. Regret can be described as a discrepancy between what actually occurs and what could have happened. Based on this they argue that for a definition of regret in the *stit* framework, it needs to be extended with modalities for knowledge and desire. Lorini and Schwarzentruber consider three different *stit* formalisms to base their definitions on, but here it will suffice to discuss their ideas using the Chellas *stit* logic given in section 4.1.

Lorini and Schwarzentruber add knowledge to their *stit* framework in the most straightforward way: a normal S5 knowledge operator $[ag_i\ \texttt{kno}]$ is added for every agent ag_i in the system. The interpretation is in terms of equivalence classes over the basic units of evaluation (for the stit language in section 4.1: dynamic states). The second operator we take from their system is an operator for desire, which is defined using propositional constants.

[10] Note however that it is not the emotions that are counterfactual; the theory is about factual emotions based on beliefs about counterfactual conditions

Definition 4.10 *Let $good_{ag_i}$ denote a propositional constant, one for each agent ag_i in the system, whose truth expresses that a state is good for that agent. Now the modal operators $[ag_i\ \texttt{good}]\varphi$ and $[ag_i\ \texttt{des}]\varphi$ are defined by:*

$$[ag_i\ \texttt{good}]\varphi \equiv_{def} \Box(good_{ag_i} \rightarrow \varphi)$$
$$[ag_i\ \texttt{des}]\varphi \equiv_{def} [ag_i\ \texttt{kno}][ag_i\ \texttt{good}]\varphi$$

The counterfactual aspect is introduced by the definition of the notion of "could have prevented" (CHP). For definitions of such counterfactual properties, the *stit* framework is more suited than dynamic logic or situation calculus frameworks, since in *stit* we reason about actual performances of actions which also makes it possible to reason about choices that are not (f)actual. Lorini and Schwarzentruber define their notion of CHP in a group stit framework. Here we have only defined individual agency. Therefore we slightly adapt Lorini and Schwarzentruber's definition to the present setting. The intuition behind Lorini and Schwarzentruber's definition of CHP is that something could have been prevented if and only if (1) it is true, and (2) some other agent does not see to it that it is true[11].

Definition 4.11 $\langle[ag_i\ \texttt{CHP}]\rangle\varphi \equiv_{def} \varphi \wedge \neg \bigvee_{ag_j \in Ags \setminus ag_i} [ag_j\ \texttt{Cstit}]\varphi$

Note that CHP is not a normal modality (which is why we use the combination of sharp and square brackets, as explained before). Lorini and Schwarzentruber show that it obeys agglomeration[12], but not weakening.

With the right concepts in place, finally we are able to define the notion of regret for a proposition φ as the desire for $\neg\varphi$ in conjunction with knowledge about the fact that φ could have been prevented.

Definition 4.12 $\langle[ag_i\ \texttt{rgt}]\rangle\varphi \equiv_{def} [ag_i\ \texttt{des}]\neg\varphi \wedge [ag_i\ \texttt{kno}]\langle[ag_i\ \texttt{CHP}]\rangle\varphi$

4.2 BDI modalities in XSTIT: dynamic attitudes

In the two systems discussed in sections 4.1 and 4.1 BDI notions were combined with *stit* operators. However, in both approaches there was no special attention for a dynamic interpretation of the combination of BDI and stit operators. Yet this interpretation strongly suggests itself. If an operator like $[ag\ \texttt{Cstit}]\varphi$ expresses that agent ag now exercises his choice to ensure that φ, and a knowledge operator like $[ag\ \texttt{kno}]\varphi$ also has a dynamic reading (note that the truth condition of knowledge is not with respect to static states s but with respect to dynamic states $\langle s, h\rangle$), then a natural interpretation of a combination like $[ag\ \texttt{kno}][ag\ \texttt{Cstit}]\varphi$ is that agent ag "knowingly sees to it that φ". We suspect that this dynamic reading was not suggested by the authors of the discussed systems because these systems do not contain temporal modalities and because the used version of *stit* has instantaneous effects. In this section we report on work studying the notions of knowingly and intentionally doing that is based on a version of *stit* where agency inherently involves a move to some next state: XSTIT.

[11] In our opinion these are necessary but not sufficient conditions for defining CHP: the fact that agents other than ag_i do not ensure that φ, does not imply that ag_i can ensure that $\neg\varphi$, which seems necessary for preventing φ. Also, counterfactual truth is about alternative *possibilities*, which suggests using a $\Diamond$ operator in the definition.

[12] Actually agglomeration might be considered a counter intuitive property for CHP: if ag could have prevented φ (by some means) and if ag could have prevented ψ (by some other means), it should not follow that ag could have prevented $\varphi \wedge \psi$ (by which of the two means? Why would there be a third means preventing both?).

XSTIT

We give here the basic definitions for XSTIT. XSTIT enriches the CSTIT language of section 4.1 with a temporal next operator and replaces the operator $[ag\ \texttt{Cstit}]\varphi$ by the operator $[ag\ \texttt{xstit}]\varphi$ where the effect φ is not instantaneous but occurs in a next state. Further explanations and motivations can be found elsewhere [9, 10].

Definition 4.13 *Given a countable set of propositions P and $p \in P$, and given a finite set Ags of agent names, and $ag \in Ags$, the formal language $\mathcal{L}_{\mathsf{XSTIT}}$ is:*

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [ag\ \texttt{xstit}]\varphi \mid X\varphi$$

The operator $[ag\ \texttt{xstit}]\varphi$ stands for 'agent ag sees to it that φ in the next state'. The operator $X\varphi$ is a standard next time operator. $\Box$ is again the operator for historical necessity (settledness).

Definition 4.14 *An XSTIT-frame is a tuple $\langle S, H, E\rangle$ such that:*

1. *S is a non-empty set of static states. Elements of S are denoted s, s', etc.*
2. *H is a non-empty set of possible system histories $\ldots s_{-2}, s_{-1}, s_0, s_1, s_2, \ldots$ with $s_x \in S$ for $x \in \mathbb{Z}$. Elements of H are denoted h, h', etc. We denote that s' succeeds s on the history h by $s' = succ(s, h)$ and by $s = prec(s', h)$. We have the following bundling constraint on the set H:*
   - **a.** *if $s \in h$ and $s' \in h'$ and $s = s'$ then $prec(s, h) = prec(s', h')$*
3. *$E : S \times H \times Ags \mapsto 2^S$ is an h-effectivity function yielding for an agent ag the set of next static states allowed by the choice exerted by the agent relative to a history. We have the following constraints on h-effectivity functions:*
   - **a.** *if $s \notin h$ then $E(s, h, ag) = \emptyset$*
   - **b.** *if $s' \in E(s, h, ag)$ then $\exists h' : s' = succ(s, h')$*
   - **c.** *if $s' = succ(s, h')$ and $s' \in h$ then $s' \in E(s, h, ag)$*
   - **d.** *$E(s, h, ag_1) \cap E(s, h', ag_2) \neq \emptyset$ for $ag_1 \neq ag_2$*

Condition **3.b** ensures that next state effectivity as seen from a current state s does not contain states s' that are not reachable from the current state through some history. Condition **3.c** expresses the *stit* condition of 'no choice between undivided histories'. Condition **3.d** above states that simultaneous choices of different agents never have an empty intersection. This is the central condition of 'independence of agency'. It reflects that a choice exertion of one agent can never have as a consequence that some other agent is limited in the choices it can exercise simultaneously.

Again, we evaluate truth with respect to dynamic states.

Definition 4.15 *Relative to a model $\mathcal{M} = \langle S, H, E, V\rangle$, truth $\langle s, h\rangle \models \varphi$ of a formula φ in a dynamic state $\langle s, h\rangle$, with $s \in h$, is defined as:*

$$\begin{array}{lll}
\langle s,h\rangle \models p & \Leftrightarrow & s \in V(p) \\
\langle s,h\rangle \models \neg\varphi & \Leftrightarrow & \text{not } \langle s,h\rangle \models \varphi \\
\langle s,h\rangle \models \varphi \wedge \psi & \Leftrightarrow & \langle s,h\rangle \models \varphi \text{ and } \langle s,h\rangle \models \psi \\
\langle s,h\rangle \models \Box\varphi & \Leftrightarrow & \forall h' : \text{ if } s \in h' \text{ then } \langle s,h'\rangle \models \varphi \\
\langle s,h\rangle \models X\varphi & \Leftrightarrow & \text{if } s' = succ(s,h) \text{ then } \langle s',h\rangle \models \varphi \\
\langle s,h\rangle \models [ag\ \texttt{xstit}]\varphi & \Leftrightarrow & \forall s',h' : \text{if } s' \in E(s,h,ag) \text{ and} \\
 & & s' \in h' \text{ then } \langle s',h'\rangle \models \varphi
\end{array}$$

Satisfiability, validity on a frame and general validity are defined as usual.

Now we proceed with the axiomatization.

Theorem 4.16 ([10]) *The following axiom schemas, in combination with a standard axiomatization for propositional logic, and the standard rules (like necessitation) for the normal modal operators, define a complete Hilbert system for* XSTIT:

	S5 for $\Box$
	For each ag, KD for $[ag\ \texttt{xstit}]$
(Lin)	$\neg X \neg \varphi \leftrightarrow X\varphi$
(Sett)	$\Box X\varphi \rightarrow [ag\ \texttt{xstit}]\varphi$
(XSett)	$[ag\ \texttt{xstit}]\varphi \rightarrow X\Box\varphi$
(Indep)	$\Diamond[ag_1\ \texttt{xstit}]\varphi \wedge \ldots \wedge \Diamond[ag_n\ \texttt{xstit}]\psi \rightarrow$ $\Diamond([ag_1\ \texttt{xstit}]\varphi \wedge \ldots \wedge [ag_n\ \texttt{xstit}]\psi)$ *for* $Ags = \{ag_1, \ldots, ag_n\}$

Knowingly doing

To study the notion of knowingly doing, like before, in section 4.1, a normal S5 knowledge operator $[ag_i\ \texttt{kno}]\varphi$ is added for every agent ag_i in the system. The equivalence classes, or 'information sets' of these operators contain state-history pairs, which means knowledge concerns information about the dynamics of the system of agents. Now knowingly doing is suitably modeled by the combination of operators $[ag_i\ \texttt{kno}][ag_i\ \texttt{xstit}]\varphi$. For the logic of this notion we can consider several interactions between the contributing modalities. Here we only briefly mention some of the possibilities. For a more elaborate discussion we refer to [10] or [11].

It is a fundamental property of agency that an agent cannot know what other agents choose simultaneously. This is expressed by the following axiom.

Definition 4.17 *The property of ignorance about concurrent choices of others is defined by the axiom:*

(IgnCC) $\quad [ag_1\ \texttt{kno}][ag_2\ \texttt{xstit}]\varphi \rightarrow [ag_1\ \texttt{kno}]\Box[ag_2\ \texttt{xstit}]\varphi$ *for* $ag_1 \neq ag_2$

(IgnCC) expresses that if an agent knows that something results from the choice of another agent, it can only be that the agent knows it is settled that that something results from a choice of the other agent.

Definition 4.18 *The property of knowledge about the next state is defined by the axiom:*

(XK) $\quad [ag\ \texttt{kno}]X\varphi \rightarrow [ag\ \texttt{kno}][ag\ \texttt{xstit}]\varphi$

The (XK) property expresses that the only things an agent can know about the next state are the things it knows to be seeing to it itself.

Definition 4.19 *The property of effect recollection is defined by the axiom:*

(Rec-Eff) $\quad [ag\ \texttt{kno}][ag\ \texttt{xstit}]\varphi \rightarrow [ag\ \texttt{xstit}][ag\ \texttt{kno}]\varphi$

(Rec-Eff) expresses that if agents knowingly see to something, then they know that something is the case in the resulting state.

The above three properties for knowingly doing just exemplify some of the possibilities. More properties have been studied. Also the theory on these dynamic attitudes has been extended to beliefs and to intentions [11]. The case of intentional action is particularly interesting because there is an extensive philosophical literature on this notion. One of the philosophical scenario's discussed in [11] is the side effect problem. Here we can only point to the fact that XSTIT, after addition of the right BDI modalities, seems to be a suitable base logic for the study of such notions.

5 Conclusion

In this paper we have reviewed the use of epistemic logic, extended with other modalities for motivational attitudes such as desires and intentions, for describing (the behaviour of) intelligent agents. What is immediately clear is that although all logical approaches are based on and inspired by Bratman's seminal work on BDI theory for practical reasoning, the formalizations themselves are quite different in nature. They also enjoy different (and sometimes even on first sight contradictory) properties. In our view this means that although Bratman did his uttermost to present a clear philosophy, as is often the case when formalizing these kind of philosophical theories, there is still a lot of ambiguity or, put more positively, freedom to formalise these matters. We have even seen that quite different base logics may be used, such as (branching-time) temporal logic, dynamic logic and also stit logic. This makes the formal logics in themselves hard to compare. We think it depends on the purpose of the formalisation (is it used for better understanding, or does it serve as a basis for computational and executable frameworks) which of the BDI logics will be most appropriate. The latter is especially important for designers and programmers of agent systems. (As an example, the KARO framework played an important role when we were devising the agent programming language 3APL [24], and we have looked at several formal relations between the two in order to get a verification logic for 3APL and its derivatives [26, 40].) As a proper treatment of this aspect of BDI logics goes beyond the scope of this paper we refer to the literature on this issue, e.g. [29, 16, 3, 15, 25].

References

[1] C. Adam, Emotions: from psychological theories to logical formalization and implementation in a BDI agent, PhD Thesis, Institut National Polytechnique de Toulouse, Toulouse, 2007.

[2] C. Adam, A. Herzig & D. Longin, A logical formalization of the OCC theory of emotions, *Synthese* 168(2), 2009, pp. 201-248.

[3] N. Alechina, M. Dastani, B. Logan & J.-J. Ch. Meyer, A Logic of Agent Programs, in: Proc. AAAI-07 (R.C. Holte & A.E. Howe, eds.), Vancouver, Canada, AAAI Press, 2007, pp. 795-800.

[4] P. Balbiani, A. Herzig & N. Troquard. Alternative axiomatics and complexity of deliberative stit theories, *Journal of Philosophical Logic*, 2007.

[5] Blackwell Reference Online: www.blackwellreference.com

[6] M.E. Bratman, *Intentions, Plans, and Practical Reason*, Harvard University Press, Massachusetts, 1987.

[7] M.E. Bratman. What is intention? In P. R. Cohen, J. Morgan & M. E. Pollack (eds.), *Intentions in Communication*, chapter 2, pages 15Ð31. MIT Press, Cambridge, MA, 1990.

[8] P. Bretier. La communication orale coopérative: contribution à la modélisation logique et à la mise en oeuvre d'un agent rationnel dialoguant, PhD thesis, Université Paris Nord, Paris, France, 1995.

[9] J.M. Broersen, A complete stit logic for knowledge and action, and some of its applications, in: M. Baldoni, T. Cao Son, M.B. van Riemsdijk & M. Winikoff, eds., Declarative Agent Languages and Technologies VI (DALT 2008), volume 5397 of *Lecture Notes in Computer Science*, pages 47Ð59, 2009.

[10] J.M. Broersen, Deontic epistemic stit logic distinguishing modes of mens rea, *Journal of Applied Logic* 9(2), 2011, pp. 127–152.

[11] J.M. Broersen. Making a start with the stit logic analysis of intentional action, *Journal of Philosophical Logic* 40, 2011, pp. 399–420.

[12] B. F. Chellas, On bringing it about, *Journal of Philosophical Logic* 24, 1995, pp. 563–571.

[13] P.R. Cohen & H.J. Levesque, Intention is Choice with Commitment, *Artificial Intelligence* 42(3), 1990, pp. 213–261.

[14] A.R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Grosset / Putnam Press, New York, 1994.

[15] M. Dastani, K.V. Hindriks & J.-J. Ch. Meyer (eds.), *Specification and Verification of Multi-Agent Systems*, Springer, New York/Dordrecht/Heidelberg/London, 2010.

[16] M. Dastani, B. van Riemsdijk & J.-J. Ch. Meyer, A Grounded Specification Language for Agent Programs, in Proc. 6th Int. J. Conf. On Autonomous Agents and Multi-Agent Systems (AAMASÕ07) (M. Huhns, O. Shehory, E.H. Durfee & M. Yokoo, eds.), Honolulu, Hawai'i, USA, 2007, pp. 578-585.

[17] D.C. Dennett, *The Intentional Stance*, MIT Press, Cambridge, Mass., 1987.

[18] H. van Ditmarsch, W. van der Hoek & B. Kooi, Dynamic Epistemic Logic, Springer, Dordrecht, 2007.

[19] E.A. Emerson. Temporal and modal logic, in J. van Leeuwen, ed., *Handbook of Theoretical Computer Science, volume B: Formal Models and Semantics*, chapter 14, Elsevier Science, Amsterdam, 1990, pp. 996–1072.

[20] H. H. Hansen, Monotonic modal logics, Master's thesis, ILLC, Amsterdam, 2003.

[21] D. Harel, Dynamic Logic, in: D. Gabbay & F. Guenthner (eds.), *Handbook of Philosophical Logic, Vol. II*, Reidel, Dordrecht/Boston, 1984, pp. 497–604.

[22] A. Herzig & D. Longin. C&L intention revisited, in: Proc. 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning (KR2004) (D. Dubois, C. Welty & M.-A. Williams, eds.), AAAI Press, 2004, pp. 527–535.

[23] A. Herzig, E. Lorini, J.F. Hübner & L. Vercouter. A logic of trust and reputation, *Logic Journal of the IGPL* 18(1, 2010, pp. :214–244.

[24] K.V. Hindriks, F.S. de Boer, W. van der Hoek & J.-J. Ch. Meyer, Agent Programming in 3APL, *Int. J. of Autonomous Agents and Multi-Agent Systems* 2(4), 1999, pp.357–401.

[25] K.V. Hindriks, W. van der Hoek & J.-J. Ch. Meyer, GOAL Agents Instantiate Intention Logic, in: Logic Programs, Norms and Action (Sergot Festschrift) (A. Artikis, R. Craven, N.K. Çiçekli, B. Sadighi & K. Stathis, eds), LNAI 7360, Springer, Heidelberg, 2012, pp. 196-219.

[26] K.V. Hindriks & J.-J. Ch. Meyer, Agent Logics as Program Logics: Grounding KARO, in 29th Annual German Conference on AI, KI 2006 (C. Freksa, M. Kohlhase, & K. Schill, eds.)), Bremen, Germany, June 14-17, 2006, Proceedings, LNAI 4314, Springer, 2007, pp. 404–418

[27] W. van der Hoek, B. van Linder & J.-J. Ch. Meyer, An Integrated Modal Approach to Rational Agents, in: M. Wooldridge & A. Rao (eds.), *Foundations of Rational Agency*, Applied Logic Series 14, Kluwer, Dordrecht, 1998, pp. 133–168.

[28] W. van der Hoek, J.-J. Ch. Meyer & J.W. van Schagen, Formalizing Potential of Agents: The KARO Framework Revisited, in: *Formalizing the Dynamics of Information* (M. Faller, S. Kaufmann & M. Pauly, eds.), CSLI Publications, (CSLI Lect. Notes 91), Stanford, 2000, pp. 51–67.

[29] W. van der Hoek & M. Wooldridge, Towards a Logic of Rational Agency, *Logic J. of the IGPL* 11(2), 2003, pp. 133–157.

[30] J.F. Horty, *Agency and Deontic Logic*, Oxford University Press, 2001.

[31] B. van Linder, Modal Logics for Rational agents, PhD. Thesis, Utrecht University, 1996.

[32] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Actions that Make You Change Your Mind: Belief Revision in an Agent-Oriented Setting, in: *Knowledge and Belief in Philosophy and Artificial Intelligence* (A. Laux & H. Wansing, eds.), Akademie Verlag, Berlin, 1995, pp. 103–146.

[33] B. van Linder, W. van der Hoek & J.-J. Ch. Meyer, Seeing is Believing (And So Are Hearing and Jumping), *Journal of Logic, Language and Information* 6, 1997, pp. 33–61.

[34] E. Lorini & R. Demolombe, Trust and norms in the context of computer security: toward a logical formalization: in: Proc. of the International Workshop on Deontic Logic in Computer Science (DEON 2008) (R. Van der Meyden & L. Van der Torre, eds.), volume 5076 of LNCS, Springer-Verlag, Berlin/Heidelberg, 2008, pp. 50–64.

[35] E. Lorini & A. Herzig, A logic of intention and attempt, *Synthese KRA* 163(1), 2008, pp. 45–77.

[36] E. Lorini & F. Schwarzentruber, A logic for reasoning about counterfactual emotions, *Artificial Intelligence* 175, 2011, pp. 814–847.

[37] D.V. McDermott, A temporal logic for reasoning about processes and plans, *Cognitive Science* 6, 1982, pp. 101–155.

[38] J.-J. Ch. Meyer, Reasoning about Emotional Agents, in Proc.16th European Conf. on Artif. Intell. (ECAI 2004) (R. López de Mántaras & L. Saitta, eds.), IOS Press, 2004, pp. 129-133.

[39] J.-J. Ch. Meyer, Reasoning about Emotional Agents, *Int. J. of Intelligent Systems* 21 (6), 2006, pp. 601-619.

[40] J.-J. Ch. Meyer, Our Quest fot the Holy Grail of Agent Verification, in: Proc. TABLEAUX 2007 (N. Olivetti, ed.), LNAI 4548, Springer, Berlin/Heidelberg, 2007, pp. 2-9.

[41] J.-J. Ch. Meyer & W. van der Hoek, *Epistemic Logic for AI and Computer Science*, Cambridge Tracts in Theoretical Computer Science 41, Cambridge University Press, 1995.

[42] J.-J. Ch. Meyer, W. van der Hoek & B. van Linder, A Logical Approach to the Dynamics of Commitments, *Artificial Intelligence* 113, 1999, 1–40.

[43] K. Oatley & J.M. Jenkins, *Understanding Emotions*, Blackwell Publishing, Malden/Oxford, 1996.

[44] A. Ortony, G.L. Clore & A. Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, 1988.

[45] A.S. Rao & M.P. Georgeff, Modeling rational agents within a BDI-architecture, in *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)* (J. Allen, R. Fikes & E. Sandewall, eds.), Morgan Kaufmann, 1991, pp. 473–484.

[46] A.S. Rao & M.P. Georgeff, Decision Procedures for BDI Logics, *J. of Logic and Computation* 8(3), 1998, pp. 293–344.

[47] M. D. Sadek. A study in the logic of intention, in: Proc. 4th Int. Conf. on Knowledge Representation and Reasoning (KRÕ92) (B. Nebel, C. Rich & W. Swartout, eds.), Cambridge, Massachusetts, Morgan Kaufmann, 1992, pp. 462–473.

[48] M. D. Sadek. Dialogue acts are rational plans, in: The structure of mutimodal dialogue (M.M. Taylor, F. Nel, and D.G. Bouwhuis, eds.), Philadelphia/Amsterdam, 2000. John Benjamins publishing company. From ESCA/ETRW, Workshop on The Structure of Multimodal Dialogue (Venaco II), 1991, pp. 167–188.

[49] C. Semmling & H. Wansing, From BDI and stit to bdi-stit Logic, *Logic and Logical Philosophy* 17(17), 2008, pp. 185–207.

[50] B. Steunebrink, M. Dastani & J.-J. Ch. Meyer, A Logic of Emotions for Intelligent Agents, in Proc. AAAI-07 (R.C. Holte & A.E. Howe, eds.), Vancouver, Canada, AAAI Press, 2007, pp. 142–147.

[51] B.R. Steunebrink, M. Dastani & J.-J. Ch. Meyer, A Formal Model of Emotion Triggers for BDI Agents with Achievement Goals, Synthese/KRA 185 (1), 2012, pp. 83–129 (KRA, pp. 413–459).

[52] M.J. Wooldridge, Intelligent Agents, in: *Multiagent Systems* (G. Weiss, ed.), The MIT Press, Cambridge, MA, 1999, pp. 27–77.

[53] M.J. Wooldridge, *Reasoning about Rational Agents*, The MIT Press, Cambridge, MA, 2000.

[54] M. Wooldridge, *An Introduction to MultiAgent Systems* (2nd edition), John Wiley & Sons, Chichester, UK, 2009.

[55] M.J. Wooldridge & N.R. Jennings (eds.), *Intelligent Agents*, Springer, Berlin, 1995.

[56] M. Xu. Axioms for deliberative stit. *Journal of Philosophical Logic* 27, 1998, pp. 505–552.