# BDI logics for BDI architectures: old problems, new perspectives

Andreas Herzig, Emiliano Lorini, Laurent Perrussel, Zhanhao Xiao

# BDI Logics for BDI Architectures: Old Problems, New Perspectives

**Andreas Herzig**[1] · **Emiliano Lorini**[1] · **Laurent Perrussel**[1] · **Zhanhao Xiao**[1,2]

**Abstract** The mental attitudes of belief, desire, and intention play a central role in the design and implementation of autonomous agents. In 1987, Bratman proposed their integration into a belief–desire–intention (BDI) theory that was seminal in AI. Since then numerous approaches were built on the BDI paradigm, both practical (BDI architectures and BDI agents) and formal (BDI logics). The logical approaches that were most influential are due to Cohen and Levesque and to Rao and Georgeff. However, three fundamental problems remain up to now. First, the practical and the formal approaches evolved separately and neither fertilised the other. Second, only few formal approaches addressed some important issues such as the revision of intentions or the fundamentally paraconsistent nature of desires, and it seems fair to say that there is currently no consensical, comprehensive logical account of intentions. Finally, only few publications study the interaction between intention and other concepts that are naturally connected to intention, such as actions, planning, and the revision of beliefs and intentions. Our paper summarizes the state of the art, discusses the main open problems, and sketches how they can be addressed. We argue in particular that research on intention should be better connected to fields such as reasoning about actions, automated planning, and belief revision and update.

**Keywords** Belief · Desire · Intention · Goal · BDI logic · BDI architecture

✉ Andreas Herzig
herzig@irit.fr;
http://www.irit.fr/∼Andreas.Herzig

Emiliano Lorini
http://www.irit.fr/∼Emiliano.Lorini

Laurent Perrussel
http://www.irit.fr/∼Laurent.Perrussel

Zhanhao Xiao
zhanhaoxiao@gmail.com;
http://www.irit.fr/∼Andreas.Herzig

[1] University of Toulouse, IRIT, Toulouse, France

[2] Department of CS, WSU, Penrith, Australia

## 1 Introduction

The concepts of belief and goal play a central role in the design and implementation of autonomous agents. These concepts do not originate in the AI and multi-agent systems literature but rather stem from philosophy of mind. There, they are considered to be fundamental mental attitudes of agents: beliefs have a 'mind-to-world' direction of fit (agents try to adapt their beliefs to the truths of the world), while intentions have a 'world-to-mind' direction of fit: agents try to make the world match their goals.

In his seminal 1987 book, Bratman proposed a richer, more fine-grained analysis where goals are replaced by *desires* and *intentions* [12]. His integrated account is called belief–desire–intention model, BDI model for short. It was well received in AI: numerous approaches adopted the BDI paradigm, either from an implementation perspective—so-called BDI agent languages and BDI software agents—or from a purely formal perspective: so-called BDI logics, with Cohen and Levesque's [18] and Rao and Georgeff's [48] being most influential.

Our aim in this paper is to reexamine BDI logics and their relation to BDI architectures. We provide an overview of the state of the art, stress the main open problems, and discuss how they can be addressed. Our main message is that right from the start, the field suffers three major

shortcomings. First, the practical and logical approaches evolved separately and neither was fruitful for the other. Second, none of the logical approaches addresses several important issues that were not addressed in the original Cohen and Levesque and Rao and Georgeff papers; in particular, the instrumentality relation between intentions is not accounted for in the logics, and in consequence there is no appropriate account of intention refinement, which is a fundamental concept in Bratman's model. Third, the field has always been poorly connected to other fields it should naturally interact with, most importantly: automated planning, epistemic logic, paraconsistent logic, belief revision and belief update, and action theory and reasoning about actions. As we are going to explain, several promising research avenues may take advantage of (mostly recent) developments in areas such as revision theory and Hierarchical Task Networks (HTN).

Throughout the paper agents ('individuals') are noted $i$, $j$,..., actions are noted $a$, $b$,..., and propositional symbols are noted $p$, $q$,...

The paper is organized as follows. In Sect. , we recall Bratman's BDI model, starting with its individual dimension and pursuing with a discussion of the collective aspect of intentions. In Sect. , we review existing BDI architectures and highlight their shortcomings. In Sect. , we discuss three logical renderings of Bratman's BDI model: Cohen and Levesque's, Rao and Georgeff's, and Shoham's. In Sect. we formulate challenges for future research. Section concludes.

## 2 Bratman's BDI Model

In his seminal book [12], Bratman highlighted the fundamental role of an agent's future-directed intentions: they are *high-level plans* to which the agent is committed and that she *refines* step by step, finally leading to intentional actions. Intentions therefore play a role that is intermediate between goals, plans, and actions. In this section, we first detail Bratman's perspective on individual intentions and next remind how collective or joint intention is linked to individual intentions.

### 2.1 Individual Intentions

Being commitments, intentions are *stable* mental attitudes. Indeed, according to Bratman there are only two possible reasons to abandon an intention:

- either it turns out to be impossible to satisfy;
- or it is only instrumental for another, higher-level intention the agent is about to abandon.

Here is an example involving both processes: suppose I intend to take out a loan in order to buy a house and learn that it has already been sold. Learning that I will not be able to buy the house should make me drop both intentions: I first abandon my high-level intention to buy the house (because I learned that it cannot be achieved any more); and then my instrumental intention to take a loan (because would be useless to do so).

Intentions being high-level plans, they cannot be executed directly: they have to be *refined* as time goes by, resulting in more and more elaborate plans. At the end of the refinement process there are *basic actions*, which are the actions the agent can directly execute. For example, my high-level plan to submit a paper to *KI Zeitschrift* is refined into writing a paper and uploading it to a paper management system; further down the line, the second intention is refined into logging into the system, entering information about the paper (authors, title, etc.) and uploading the PDF file; all these are again high-level actions that have to be further refined, down to basic intentions of typing words or characters on my keyboard.

While intentions have to be refined in order to obtain executable actions, this should not be done too early, for two reasons. First, an agent's memory and computational power is limited and she is not able to store fully elaborate plans for the far ahead future. Second, even if resources were unlimited, the agent only has imperfect beliefs about the future that may turn out to be wrong: fully worked-out plans would force her to re-plan much more frequently than more abstract, high-level plans would (the issue was also highlighted in [14]). So when and how to refine an intention is a fundamental issue in an agent's management of her intentions.

Forming future-directed intentions enables agents to extend the influence of their deliberations beyond the present moment. This is important given the limited cognitive capacities and time for deliberation of human agents. Specifically, it may be the case that at time $t$ an agent will have less time to deliberate and think through the options, or she may be distracted. For example, I may decide on Sunday what to do during the next weekend since I know that I will have a busy week at work and will have no time to make my plan for the weekend. Another reason why future-directed intentions are useful is that agents may be sensitive to temptations negatively biasing their choice. For example, a heavy smoker may decide to stop smoking at a certain point in his life: he may decide that he will not light up a cigarette at the later time when he will desire to smoke it. By forming this future-directed intention, he commits himself to do something later in order to contrast the opposite force of his future temptation. The idea that intentions imply some kind of commitment is explicit in Bratman's theory. It is this peculiarity which qualifies intention for a functional role that mere desires do not play. Once an agent has deliberated in favour of an action and

has formed the corresponding intention, he is "locked into" the project that he has decided to pursue and, in the absence of relevant new information, the intention to do the action will resist further reconsideration. Consequently, in being the product of deliberation and having associated a kind of commitment, intentions are characterized by an intrinsic form of persistence which makes them more resistant to temptations than desires.

Bratman's theory is qualified as a planning theory of intention and traditionally opposed to so-called cognitivist theories of intention [28, 65]. While according to Bratman's theory, intention has certain distinctive functional properties which cannot be adequately characterized by conceiving it as a combination of a desire to do a certain action *plus* the belief that one will do the action (or the belief that one will possibly do the action), the cognitivist view defends the idea that intention basically consists in the belief that one will act in a certain way (or, will try to act in a certain way). Thus, according to this view, an agent's intention involves a sort of self-referential aspect: the belief that an intention to perform a certain action $a$ in the future will be responsible for the future occurrence of action $a$ (or the future attempt to do the action $a$). A formalization of this self-referential aspect of intention is given in [39].

Before discussing the main challenges raised by Bratman's model, we detail how individual intentions are the fundamental bricks of joint intentions.

### 2.2 From Individual to Collective Intentions

Collective attitudes such as common goal and joint intention are traditionally studied in in the philosophical area to account for the concept of collaborative activity [13, 51, 61]. Notable examples of collaborative activity are painting a house together, dancing together a tango, or moving a heavy object together. Two or more agents acting together in a collaborative way need to have a common goal and need to form a joint intention aimed at achieving the common goal. In order to make collaboration effective, each agent has to commit to her part in the shared plan and form the corresponding intention to perform her part of the plan. Moreover, she has to monitor the behaviors of the others and, eventually, to reconsider her plan and adapt her behavior to new circumstances.

The concept of joint intention has been considered by logicians and AI practitioners to account for the concept of collaborative activity in multi-agent systems (cf. [21, 26]). However, much work has to be done in order to develop comprehensive formal theories of joint intention. The interesting aspect of joint intention is the conditional nature of the individual intentions composing it. Specifically, an agent in a group has the intention to do her part in the shared plan *conditional* on the fact that the other agents in the group also intend to do their part. In this sense and as Bratman emphasizes [13], the individual intentions composing a joint intention form an *interlocking web* of individual intentions. From this perspective, joint intention refinement and revision are interdependent as: (1) the refinement of an individual plan by an agent in the group may lead to the refinement of an individual plan by another agent in the group, and (2) the reconsideration of an individual intention by an agent in the group may trigger the reconsideration of an individual intention by another agent in the group. For example, suppose two agents Mary and Bob have the joint intention to paint a house together. Two options are available: the house can be painted either in blue or in green. Mary refines her individual plan by deciding to paint the house in blue. Consequently, Bob has to refine his individual plan in the same way by deciding to the paint the house in blue. Now, suppose Mary reconsiders her individual intention to paint the house in blue and chooses to paint the house in green. In order to coordinate with Mary effectively, Bob too should change his plan and decide to paint the house in green.

To sum it up, joint intention cannot be considered before individual intention is clearly characterized. Hereafter, we only discuss issues and challenges raise by individual intentions.

## 3 BDI Implementations and Their Shortcomings

Soon after Bratman's and Cohen and Levesque's papers, the BDI paradigm inspired a multitude of models and platforms aiming at the implementation of software agents. Examples are KARO [42], 3APL [20], dMars [22], AgentSpeak-Jason [6] and GOAL [34]. All these software platforms are made up of a 'B', a 'D', and an 'I' component that are interfaced appropriately. Such architectures are inspired by the Intelligent Resource-bounded Machine Architecture proposed by Bratman et al. [8]. Figure 1

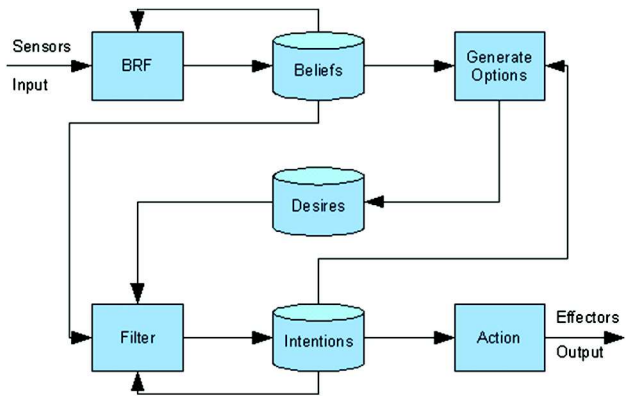


**Fig. 1** A typical BDI Implementation

contains a typical example taken from [7].[1] As it can be seen from the figure, sensor input leads to the modification of beliefs (BRF stands for the belief revision function), intentions are produced from desires by filtering with beliefs, and intentions lead to actions.

In the rest of the section we discuss the shortcomings of these models and platforms.

### 3.1 Lack of Formal Logical Semantics

Most BDI software models and platforms are semi-formal: while they provide a taxonomy of basic concepts and their relationships, the agent programming languages are usually equipped with an operational semantics only and lack a formal logical semantics. Typically, they support the specification of BDI agents with respect to some specific BDI implementation. For instance, language AgentSpeak [7] enables to express what are the initial beliefs, actions and plans available in an AgentSpeak-Jason implementation of a multi-agent system. More generally speaking, there are only few attempts to formally relate BDI implementation and BDI logics. For example, the language of AgentSpeak does not enable reasoning about the consequences of an action. The main exception is Meyer et al. work on the KARO framework [1, 4, 29]. However, it seems to be fair to say that this logic and its mathematical properties are not well understood yet.

A further weak point of BDI architectures (and consequently of BDI agents) is that their associated agent language is often severely restricted: it consists of literals, i.e., propositional symbols or their negations. Typically, the dMars agent language requires that beliefs are only sets of literals. In our view this is a major obstacle to the use of BDI agents, for two reasons. First, it does not allow for second-order beliefs, i.e., beliefs about other agents' beliefs. Such beliefs—and more generally higher-order beliefs—are however central for the reasoning of a socially intelligent agent. Their fundamental role in human intelligence was highlighted in experiments such as false belief tasks [11]. In Game Theory, higher-order beliefs are at the heart of the definition of notion of equilibrium as each agent has to assume that the other agents are rational [41, 59, 71].

Second, while some agent languages do allow for disjunctions (e.g. 3APL), most of them don't (e.g. AgentSpeak does not allow to express such kind of belief). This is clearly a disadvantage: goals such as to *know whether* some proposition is true cannot be expressed. This is highly problematic if one wants to employ BDI agents as conversational agents, where agent $i$'s yes-no question whether $\varphi$ is conditioned by $i$'s goal to know whether $\varphi$ is true and where $i$'s speech act of informing $j$ that $\varphi$ is conditioned by $i$'s belief that $j$ does not know whether $\varphi$. This situation is quite common for example in game playing.

### 3.2 Lack of Intention Refinement

As we have said, operations of refinement of intentions are fundamental in the BDI model. As pointed by Rao and Geogeff in [48], "the potential of non-primitive events for decomposition into primitive events can be used to model hierarchical plan development". One would therefore expect refinement to be a central ingredient of any model of autonomous agents. However, most of the papers in the literature on BDI logics and BDI agents remain silent about this concept. Indeed, from [35] to recent work [70], mainstream implementations of BDI-agents have adopted plan libraries: functions associating to each intention the set of plans that can achieve it. Such approaches therefore do not give any structure to an agent's intentions: no means-end relation between intentions is considered and intentions are achieved in an isolated way by finding a solution for each of them independently of the others. This is also the case even when the focus is on the dynamics of the intention base [60]. We believe that this is a major shortcoming of such approaches.

Notable exceptions are [23, 50] which import ideas from HTN planning and [24] which describes their concrete implementation framework.

A Hierarchical Task Networks (HTN) is made up of a hierarchy of actions ('tasks') that are either basic ('primitive') or high-level ('non-primitive') [25]. Contrasting with the classical planning approach, HTN-based plan generation decomposes high-level actions step-by-step into lower-level actions. Actions fall into two categories: STRIPS-like basic actions that can be executed directly and high-level actions that cannot. An action network is a couple $d = [T, \varphi]$ consisting of a set of actions $T$ and a boolean formula $\varphi$. It is achieved if the set of actions $T$ are achieved and the boolean formula $\varphi$ imposing restrictions on the temporal occurrence of action instances and on their pre- and post-conditions of actions holds. A decomposition method $(a, \psi, d)$ specifies that when formula $\psi$ holds, high-level action $a$ can be decomposed into action network $d$: $a$ is going to be achieved once $d$ is achieved. For example, the method for the high-level action of submitting a paper to *KI Zeitschrift* is conditioned by $\psi$ = "the Easychair website is available", and when that $\psi$ holds then submitting a paper can be decomposed into an action network $d = [T, \varphi]$ where $T$ consists of the two actions of writing a paper and uploading it and constraint $\varphi$ expresses that the writing action has to be performed before the uploading action. The solution for an HTN planning problem $\mathcal{P} = \langle d, B_0, \mathcal{D} \rangle$

[1] https://commons.wikimedia.org/wiki/File:Bdi-agent-architecture. Viewed July 1st, 2016.

is a plan: a sequence of basic actions such that intended (high-level) action network $d$ will be achieved by decomposing them iteratively via predefined decomposition methods in $\mathcal{D}$, starting from the initial state $B_0$.

As far as we are aware, there are few contributions relating HTN concepts with BDI agents. In [23], de Silva and Padgham show through experiments that BDI systems are more suitable when facing highly dynamic environments, while HTN solutions are more efficient in a static context. In [50], Sardina et al. integrate a BDI agent system with an HTN offline planner as a "lookahead" component and develop a BDI agent language CANPLAN. In their architecture, an intention is a program consisting of primitive actions and operations on these actions. The intention is considered to be successfully executed if its corresponding HTN network task is accomplished. Later in [24], the authors propose a notion of 'ideal' (precisely, minimal non-redundant maximally-abstract) plan and compute a suboptimal 'ideal' plan, which is non-redundant and preserves abstraction as much as possible, based on the hierarchical decomposition generated by HTN planning. The above approaches inherently restrict intentions to be handled by an underlying predefined set of decomposition methods in a static way. However, defining all possible decompositions in the beginning may be a challenge for a modeler.

To sum it up, intention refinement is absent from almost all existing BDI implementations and we believe that it is fundamental to incorporate means-end relations between intentions into the picture, building on existing work in the HTN and hybrid planning literature.

## 4 BDI Logics and Their Shortcomings

We now turn to BDI logics. We start with Cohen and Levesque's approach and the similar approach due to Rao and Georgeff. As the reader will see, these logics are fairly complicated. This leads us to Shoham et al. simpler database approach which, we argue, provides an interesting, simple alternative that however still lacks an account of refinement.

### 4.1 Cohen and Levesque's Linear Time Logic

Cohen and Levesque provided a seminal logical modeling of Bratman's BDI model [18] that was awarded the IFAAMAS most influential paper award in 2006. Their approach accounts for achievement intentions (as opposed to maintenance intentions). It distinguishes intention-to-do and intention-to-be and mainly focuses on the latter. The definition of intention-to-be comes in four steps—chosen goals, achievement goals, persistent goals and intentions—that are couched in a quantified modal logic of linear time, action, and belief.

1. Chosen goals corresponds to future states where the agent would like to be.
2. Achievement goals are chosen goals that are not true yet (more precisely: that the agent believes to be false now).
3. Persistent goals are achievement goals that are only abandoned when they are either achieved, or learned to be unachievable, or 'for some other reason'.
4. Intentions are persistent goals for which the agent is prepared to act; this excludes persistent goals to which the agent cannot contribute anything, such as my persistent goal that there be snow at Christmas.

While Cohen and Levesque's approach is much cited, it is fair to say that it is rather complicated. Some early criticisms of technical details can be found in [54]. In Shoham and Leyton-Brown's textbook the approach is called "the road to hell" [57]. It speaks for itself that its mathematical properties—such as axiomatizability, decidability and complexity of fragments—were never investigated. None of the BDI logics that were introduced subsequently—starting with [48]—adapted Cohen an dLevesque's four steps definition of intention and instead considered intentions to be primitive, the only exceptions being [30, 49]. Cohen and Levesque's approach moreover has three major shortcomings. First, it does not provide a solution to the frame problem:[2] what is true at different time points $t$ and $t'$ may vary wildly and is not determined by the actions occurring between $t$ and $t'$. Second, it does not account for intention refinement. Third, it does not fully account for revision; indeed, while Cohen and Levesque provide some criteria for the abandonment of intentions through the notion of *rational balance* (forbidding to intend something that is true or believed to be impossible to achieve), it does not further analyze the 'other reasons' for which a persistent goal is abandoned. These reasons should mainly cover abandonment of goals that are instrumental for another, higher-level goal that is dropped, and more generally intention reconsideration.

### 4.2 Rao and Georgeff-Based Logics

Contrarily to Cohen and Levesque, Rao and Georgeff [48] embrace a primitive notion of intention. It is based on the branching time logic CTL*.

[2] The frame problem, one of the main and oldest problems in reasoning about actions, concerns the specification of the effects of actions [43]. The main challenge is to characterize these effects without explicitly specifying which conditions are not affected by executing actions.

Just as Cohen and Levesque's approach, Rao and Georgeff's suffers from the shortcomings that we have listed above: intention revision is basically absent from the picture and the frame problem is not solved. Indeed, due to the temporal logic framework agents can perform actions whose effects are not further specified. It is also not described how beliefs are preserved while agents act.

Rao and Georgeff's approach was fleshed out by Winikoff et al. [69] who link intentions[3] to the actions associated to them by means of transition rules that are close to the predefined refinement rules of HTNs. The logical framework they propose, called Conceptual Agent Notation, is defined in terms of a declarative and an operational semantics. Together, they allow to reason about the relations between goals, such as dependence, mutual consistency, and mutual support. Overall, the framework is rather complex and, just as all other existing BDI logics, the frame problem remains unsolved: the framework describes how sub-goals may be inferred (with respect to some library of plans) but does not keep track of these steps. In other words, no instrumentality relation between the ongoing goals can be exhibited and consequently revision cannot be handled in a rational way.

### 4.3 Shoham's Database Perspective

Shoham recently argued for a simpler approach that he baptized the *database perspective* [52]. His aim is to define a framework that is simpler than Cohen and Levesque's and Rao and Georgeff's and that thereby provides a more suitable basis for the design and implementation of BDI agents. Shoham abandons Cohen and Levesque's idea to express achievement goals by means of the temporal 'eventually' modality. His central idea is that beliefs and intentions-to-do are organized in two temporal databases. A belief database $B$ is a set of pairs made up of time points $t$ in the set of non-negative integers $\mathbb{N}^0$ and literals $p$.[4] They are written $p_t$ and read "$p$ is true at $t$". Similarly, an intention database $I$ is a set of pairs made up of time points $t$ and (basic) actions $a$. They are noted $a_t$ and read "the agent intends to do action $a$ at time $t$".

Shoham supposes that each action $a$ has pre- and post-conditions. They are described by functions *pre* and *post* mapping each action $a$ to special atomic formulas $pre(a)$ and $post(a)$.

Letting $B_t$ be the set of $t$-indexed literals of $B$ and $I_t$ the set of $t$-indexed actions of $I$, Shoham requires the following *coherence* constraints:

1. Every $B_t$ is consistent;
2. Every $I_t$ is either empty or a singleton;
3. If $a \in I_t$ then $B_t \not\models \neg pre(a)$;
4. If $a \in I_t$ then $B_{t+1} \models post(a)$.

Icard et al. [36] provide a semantics and an axiomatization for such belief-intention databases in terms of sets of paths. A *path* $\pi$ associates to every non-negative integer $t$ a set of propositional symbols and an action: the propositional symbols that are true at $t$ and the action that is going to be performed by the agent at $t$. A set of paths $\Pi$ is *appropriate* if (1) on each path, the postcondition of each action at time $t+1$ is true and (2) once the precondition of each action is satisfied at time $t$ on $\pi$ then it must be performed on some path that is identical to $\pi$ up to time $t-1$.[5] Intuitively, $B$ and $I$ are coherent if the agent considers it possible to do all actions she intends with respect to some appropriate set of paths. Based on this formalization, Icard et al. propose AGM-like postulates for the joint revision of beliefs and intentions and provide a representation theorem.

Van Zee et al. [67] recently criticized that Icard et al. logic is unsound because their axiom which describes the appropriate set of paths is not necessarily valid. They adapted Icard et al. logic by moving to a semantics *à la* Rao and Georgeff in terms of CTL$^*$-like tree structures, plus a language with time-indexed modalities. They also provided a sound and complete axiomatization of their new logic w.r.t. the class of all models. They moreover gave an example showing that Icard et al. coherence constraint (which only considers the precondition of actions) is too weak. They proposed a stronger coherence condition where the pre- and postconditions of actions and beliefs are always jointly consistent. Based on that logic, van Zee et al. focused on the AGM-like revision of beliefs about actions and time [66, 68]. They adapted the AGM semantics of belief revision by adding a condition saying that infinite models with the same finite prefix have the same priority in the revision preorder. They then proved representation theorems in the style of Katsuno–Mendelzon and Darwiche–Pearl.

According to [53], the database perspective is at the heart of the Personal Time Assistant (PTA), which is a next-generation calendar helping people to manage time. His *Timeful* application has intentions as its basic concept and was developed within a start-up company that was acquired by Google in 2015.

[3] They use the term goals.

[4] Shoham mentioned that the belief could be any formula indexed by multiple time values, but does not elaborate this further. Such a generalization should come with more complex notation and new semantical and computational problems.

[5] The time parameter $t-1$ is missing in [36].

## 5 Challenges for Future Research

Let us now list some challenges that result from our discussions in the preceding sections. Underlying all these challenges is a general desideratum: to provide a simpler but nevertheless meaningful logic of intention encompassing the main concepts of Bratman's BDI model, which will hopefully bring about a tighter connection between BDI implementations and BDI logics.

A second general desideratum concerns tractability. The agent programming languages of BDI implementations are typically restricted for the sake of efficiency, so that agents can react on-line to a dynamic environment. One reason explaining the distance between theory and practice is the too high complexity of existing logics. To witness, model checking for BDI Rao and Georgeff logic is already PSPACE, and the satisfiability problem is way beyond. However, tractable fragments of epistemic logics can be isolated. Recent development in the logic of belief alone (and thus not intention) has demonstrated that efficient reasoning with at least some restricted forms of higher-order belief is possible [40, 45]. A similar approach might guide the definition of future BDI logics.

Another possible alternative for addressing tractability are recent approaches based on a modular definition of belief and intention. In [17], Casali et al. show how belief, intention and desire interplay via a logic based on Giunchiglia et al. multi-context systems [27]. They offer an interesting approach showing how to switch between intentions and beliefs in a simple, yet expressive way via bridges rules.

In the rest of the section we offer a list of more detailed challenges. The first is about intention refinement and somewhat includes all others. We however list it separately because it leads us to Shoham's database perspective. It is a promising research avenue, with a simple but still meaningful and non-trivial account of intention.

### 5.1 Design and Integrate Intention Refinement

The refinement of intentions is fundamental and should be a central ingredient of any model of autonomous agents. However and as we have seen, the literature on BDI logics and BDI architectures basically remain silent on this aspect. As far as we know, the only exceptions are the work of Padgham et al. [23, 50] and perhaps the work of Hunsberberger and Ortiz [31].

While refinement is also missing in Shoham's database perspective, we believe the latter to be a good starting point. Its temporal database associates to every time point a set of propositional symbols that are true at that time point and the action the agent intends to perform at that time point (which takes one unit of time). This framework should be extended by *high-level actions* which may require more than one time unit: they are performed within temporal intervals. To keep things simple one might start with STRIPS-like actions or Reiter-style basic action theories. On that basis, Bratman's central relation of *instrumentality* between intentions should be studied. This relation is, so to speak, instrumental in order to maintain an intention database: on the one hand, we refine an agenda by adding lower-level intentions (the means) that are instrumental for some high-level intention (the end) in the intention database; on the other hand, during the revision of an agenda, when we learn the unsatisfiability of a higher-level intention we also drop all those intentions that are instrumental for it, even if they can still be satisfied. Our house buying example in Sect. illustrates the latter.

### 5.2 Integrate a Solution to the Frame Problem

None of the existing BDI logics solves the frame problem: the agent's beliefs at time point $t$ together with her actions at $t$ do not determine her beliefs at $t+1$. Indeed, even in Shoham's database perspective, when action $a$ is executed at time point $t$ then its effect $post(a)$ holds at $t+1$, but there is no guarantee that the propositional symbols that are not affected by $a$ have the same truth value at $t$ and at $t+1$.

This calls for an integration of existing solutions to the frame problem, such as STRIPS-like actions or Reiter's basic action theories with successor state axioms [47], or better its epistemic extension [55]. When one tries to integrate, say, STRIPS-like actions into Shoham's database approach one however faces a new problem: STRIPS as well as Reiter's solution to the frame problem come with the hypothesis that the world evolves exclusively due to the agent's actions and is static otherwise. This forbids to take actions into account that are performed by the environment or by other agents. To witness, although Icard et al. approach has STRIPS-like action theories, it fails to solve the frame problem [36].

One way of solving this problem could be to not only consider the actions of the planning agent under concern, but also the environment's actions. Taking the perspective of the planning agent one might call the latter (external) 'events' and the former just 'actions'. Such events could be equipped with pre- and postconditions, just as actions are. We have undertaken first steps towards this in our [33].

### 5.3 Establish a Link with Revision Theory

Revision theory [3] is mainly about the evolution of an agent's belief when she learns that she was wrong about some proposition $\varphi$. While such revisions naturally also modify the agent's goals, the belief revision literature basically never studied intention revision. In contrast, the

BDI literature contains some papers accounting for this dynamic aspect [2, 4, 31, 36, 60].

As we have already mentioned, the concept of instrumentality should be an important ingredient of a theory of intention revision: when dropping a high-level intention we also drop the lower-level intentions that are instrumental for it. This can be viewed as a coarsening of the agent's intentions. In [58], Shapiro et al. gives some intuition on this instrumentality aspect and its impact on intention revision by considering relations between a predefined library of plans (end) and intentions (means). We believe that this contribution, probably combined with [67, 68], provides a good starting point. In any case, we believe that a successful intention revision theory has to be based on a definition of instrumentality among intentions.

### 5.4 Connect with Dynamic Epistemic Logics

Closely related to revision theory, the evolution of an agent's knowledge and belief when some event occurs has been much studied in dynamic epistemic logic (DELs) [64]. There also exist some papers about the evolution of agent preferences, e.g. [15, 63] as well as on belief revision, e.g. [62]. However, this stream of research has not been linked to logics of intention yet. A major shortcoming of existing DELs is that the author of an action does not have a particular status. This in particular makes it difficult to distinguish actions from mere events. A good starting point to relate intentions to DEL updates might be Castelfranchi and Paglieri's goal filtering approach to intention [19].

### 5.5 Integrate Paraconsistent Reasoning About Desires

While having it in the acronym, BDI logics actually say only little about desires (while BDI architectures do). The reason is probably that an agent's desires can be jointly inconsistent, which, it seems, makes a further logical analysis of the concept somewhat difficult. An example is my desire of buying a house and my desire of buying an expensive car, which are inconsistent with my belief that I only have enough money to buy one of them. For that reason, desires do not obey any of the standard logical laws that other mental attitudes such as belief and intention do; in particular, when agent $i$ desires $\varphi$ and desires $\psi$ then he does not necessarily desire $\varphi \wedge \psi$.

There exist logical approaches to inconsistency-tolerant reasoning: the so-called logics of paraconsistency [16]. There seem to be no approaches integrating such logics with BDI logics. One explanation could be that the former give only little consideration to modal logics. It might be of interest to explore whether and how this could be done in a meaningful way. Again, a good starting point might be Castelfranchi and Paglieri's approach where intentions are obtained by filtering of typically inconsistent sets of desires [19]. An alternative potential starting point is the numerical approach proposed by Casali et al. [17] where Belief, Desire and Intention have degrees: using these numerical values, inconsistent intentions can be handling in a natural way as in possibilistic or fuzzy logics. However, going back to the previous challenge, the revision of intentions becomes more challenging as it entails to revise not only intentions but also their associated degrees.

### 5.6 Clarify the Relation with Game Theory

Just as the BDI model, decision theory and game theory are also about the behavior of agents given their goals and their information state. The relationship has however not been clarified up to now. It should be relevant in particular for games in extensive form (as opposed to one-shot strategic games).

The conceptual apparatus of classical decision theory and game theory includes the concepts of action, belief and desire. In particular, the quantitative aspect of beliefs and desires is captured, respectively, by means of Bayesian probabilities and utilities. Thus, these theories can account for the cognitivist view of intention. Indeed and as highlighted in Sect. , the cognitivist view conceives intention as a mere belief about the future performance of an action. In contrast, classical decision theory and game theory cannot account for Bratman's concept of intention, which is not reducible to the more primitive concepts of belief and desire. According to Bratman, intentions play functional roles in mind that cannot be adequately characterized by conceiving it as a combination of beliefs and desires. We believe that extending classical decision theory and game theory with the concept of intention might be relevant when trying to model resource-bounded agents who need to plan their future actions in advance since they have limited cognitive capacities and limited time for deliberation.

### 5.7 Join Forces with the Planning Community

Plans being a central concept in Bratman's model, it is astonishing that—leaving aside some early tentatives such as [8]—no connections with the planning community were established yet. This can be explained by the planning community's 'top level goal' to provide efficient plan generating algorithms. Such algorithms are well-studied by now, with highly competitive solvers running not only on classical planning problems, but also on problems with incomplete knowledge [46] and with temporally extended actions [9]. Consequently, the planning community recently moved towards multiagent planning problems

[37]. This is paralleled by an interest in planning in the DEL community [5].

Some promising contributions aiming at a connection between the planning domain and BDI agents exist. We already mentioned [50], but classical planning has also been considered [10, 44]. All these contribution mix a declarative and an operational semantics (a similar view is also considered in [56]). We believe that this perspective should guide the specification of the next generation of BDI logics. These logics will be successful if they are rooted in these two semantics.

According to this point of view, it seems to us that time has come to reconsider the link between BDI models and plan generation: the integration of HTN planning into BDI logics that we have mentioned above is a promising first step. As mentioned, up to now, decomposition methods bring a too rigid solution for defining instrumentality relations between intentions. A more general perspective, such as the one offered by hybrid planning [38], is to consider that high-level actions also have effects. Characterizing such effect is not trivial, as it raises the question of the main ('primary') effect [38] of an action. We propose a logical approach in [32]. Mixing BDI reasoning and hybrid planning has been proposed in [24]: this work is a promising starting point even if the primary effect of an action is not clearly characterized.

To sum it up, we believe that one of the very first step is to explore how the notion of instrumentality can be examined by relating hybrid planning and intention refinement. To do so, HTN and hybrid planning have to be characterized in a more declarative way in order to implement in agent languages the ability to reason about action effects. Building such bridges between the planning and the BDI field should contribute to push further the definition of innovative BDI agent theories and languages.

## 6 Conclusion

We have provided a concise overview of the 25 years old AI literature on BDI logics, BDI architectures, and BDI agents. We have shown that despite numerous publications, some fundamental theoretical issues were neglected up to now. We believe that the research avenues that we have sketched are potentially fruitful and should lead to progress within the near future. We first advocate that the second generation of BDI logics should be rooted in an effective definition of intention refinement; next we propose to adopt Shoham's database perspective and planning as key components of future BDI agents. It will allow to have a promising and innovative semantics of future BDI logics. Our short term goal is to be part of this adventure. First steps on this research avenue are proposed in [33].

Even if the challenges are numerous, the list is still partial. As mentioned in Sect. , while we focus on the individual dimension of intention, the collective aspect cannot be ignored: future BDI agents will run a in multi-agent environment. It is is clear that the numerous notions we have introduced need to be fully redefined as soon as a society of agents is considered. Such representative example is the notion of instrumentality: instrument may be shared between agents and conflicts may then appear. It is our long term goal to tackle these new challenges.

**Acknowledgments** Our warmest thanks go to the reviewers of the KI Zeitschrift for their thorough reading and thoughtful comments. This work was partially supported by CSC (Chinese Scholarship Council) and by by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02.

## References

1. Alechina N, Dastani M, Logan B, Meyer JJC (2008) Reasoning about agent deliberation. In: Proceedings of the 11th international conference on principles of knowledge representation and reasoning (KR)
2. Alechina N, Dastani M, Logan B, Meyer JJC (2011) Reasoning about plan revision in BDI agent programs. Theoret Comput Sci 412(44):6115–6134
3. Alchourrón CE, Gärdenfors P, Makinson D (1985) On the logic of theory change: partial meet contraction and revision functions. J Symb Log 50(02):510–530
4. Alechina N, Jago M, Logan B (2008) Preference-based belief revision for rule-based agents. Synthese 165(2):159–177
5. Bolander T, Andersen MB (2011) Epistemic planning for single and multi-agent systems. J Appl Non Class Log 21(1):9–34
6. Bordini RH, Hübner JF (2010) Semantics for the Jason variant of AgentSpeak (plan failure and some internal actions). In: Proceedings of the 19th European conference on artificial intelligence (ECAI), volume 215 of frontiers in artificial intelligence and applications. IOS Press, pp 635–640
7. Bordini RH, Hübner JF, Wooldridge MJ (2007) Programming multi-agent systems in AgentSpeak using Jason, volume 8 of Wiley Series in Agent Technology. Wiley, Oxford
8. Bratman ME, Israel DJ, Pollack ME (1988) Plans and resource-bounded practical reasoning. J Comput Intell 4(3):349–355
9. Bacchus F, Kabanza F (1998) Planning for temporally extended goals. Ann Math Artif Intell 22(1–2):5–27
10. Bauters K, Liu W, Hong J, Sierra C, Godo L (2014) CAN(-PLAN)+: extending the operational semantics of the BDI architecture to deal with uncertain information. In: Proceedings of the 13th conference on uncertainty in artificial intelligence (UAI), pp 52–61
11. Bolander T (2014) Seeing is believing: formalising false-belief tasks in dynamic epistemic logic. In: Proceedings of the European conference on social intelligence (ECSI), volume 1283, CEUR Workshop Proceedings, pp 87–107
12. Bratman ME (1987) Intention, plans, and practical reason. Cambridge: Harvard University Press **(Reedited 1999 with CSLI Publications)**
13. Bratman ME (1992) Shared cooperative activity. Philos Rev 101(2):327–341

14. Bratman ME (2009) Intention, belief, and instrumental rationality. In: Sobel D, Wall S (eds) Reasons for action. Cambridge University Press, Cambridge, pp 13–36
15. Baltag A, Smets S (2006) Conditional doxastic models: a qualitative approach to dynamic belief revision. Electron Notes Theoret Comput Scie 165:5–21
16. Carnielli W, Coniglio ME, Marcos J (2007) Logics of formal inconsistency. Handbook of philosophical logic. Springer, Berlin, pp 1–93
17. Casali A, Godo L, Sierra C (2011) A graded BDI agent model to represent and reason about preferences. Artif Intell 175(7):1468–1478
18. Cohen PR, Levesque HJ (1990) Intention is choice with commitment. Artif Intell 42(2):213–261
19. Castelfranchi C, Paglieri F (2007) The role of beliefs in goal dynamics: prolegomena to a constructive theory of intentions. Synthese 155(2):237–263
20. Dastani M, de Boer F, Dignum F, Meyer JJC (2003) Programming agent deliberation: an approach illustrated using the 3APL language. In: Proceedings of the 2nd international joint conference on autonomous agents and multiagent systems (AAMAS), pp 97–104. ACM
21. Dunin-Keplicz B, Verbrugge R (2010) Teamwork in multi-agent systems: a formal approach. Wiley, Oxford
22. d'Inverno M, Luck M, Georgeff MP, Kinny D, Wooldridge MJ (2004) The dMars architecture: a specification of the distributed multi-agent reasoning system. Auton Agents Multi Agent Syst 9(1–2):5–53
23. De Silva L, Padgham L (2005) A comparison of BDI based real-time reasoning and HTN based planning. AI 2004: advances in artificial intelligence. Springer, Berlin, pp 1167–1173
24. De Silva L, Sardina S, Padgham L (2009) First principles planning in BDI systems. In: Proceedings of the 8th international conference on autonomous agents and multiagent systems (AAMAS), pp 1105–1112
25. Erol K, Hendler JA, Nau DS (1994) HTN planning: complexity and expressivity. In: Proceedings of the 12th national conference on artificial intelligence (AAAI), volume 94, pp 1123–1128
26. Grosz B, Kraus S (1996) Collaborative plans for complex group action. Artif Intell 86(2):269–357
27. Giunchiglia F, Serafini L (1994) Multilanguage hierarchical logics, or: how we can do without modal logics. Artif Intell 65(1):29–70
28. Harman G (1976) Practical reasoning. Rev Metaphys 29(3):431–463
29. Hustadt U, Dixon C, Schmidt RA, Fisher M, Meyer JJC, van der Hoek W (2001) Reasoning about agents in the KARO framework. In: Proceedings of the 8th international symposium on temporal representation and reasoning, (TIME)
30. Herzig A, Longin D (2004) C&L intention revisited. In: Proceedings of the 8th international conference on principles of knowledge representation and reasoning (KR). AAAI Press, pp 527–535
31. Hunsberger L, Ortiz CL (2008) Dynamic intention structures I: a theory of intention representation. Auton Agents Multi Agent Syst 16(3):298–326
32. Herzig A, Perrussel L, Xiao Z (2016) On hierarchical task networks. In: Proceedings of the 15th European conference on logics in artificial intelligence (JELIA). Springer
33. Herzig A, Perrussel L, Xiao Z, Zhang D (2016) Refinement of intentions. In: Proceedings of the 15th European conference on logics in artificial intelligence (JELIA). Springer
34. Hindriks KV, van der Hoek W, Meyer JJC (2012) GOAL agents instantiate intention logic. Logic programs, norms and action, volume 7360 of lecture notes in computer science. Springer, Berlin, pp 196–219
35. Ingrand FF, Georgeff MP, Rao AS (1992) An architecture for real-time reasoning and system control. IEEE Expert 7(6):34–44
36. Icard T, Pacuit E, Shoham Y (2010) Joint revision of belief and intention. In: Proceedings of the 6th international conference on principles of knowledge representation and reasoning (KR), pp 572–574
37. Kominis F, Geffner H (2015) Beliefs in multiagent planning: from one agent to many. In: Proceedings of the 25th international conference on automated planning and scheduling (ICAPS). AAAI Press, pp 147–155
38. Kambhampati S, Mali A, Srivastava B (1998) Hybrid planning for partially hierarchical domains. In: Proceedings of the 15th national conference on artificial intelligence and 10th innovative applications of artificial intelligence conference (AAAII/IAAI), pp 882–888
39. Lorini E, Herzig A (2008) A logic of intention and attempt. Synthese 163(1):45–77
40. Lakemeyer G, Lespérance Y (2012) Efficient reasoning in multiagent epistemic logics. In: Proceedings of the 20th European conference on artificial intelligence (ECAI), pp 498–503
41. Lorini E, Moisan F (2011) An epistemic logic of extensive games. Electron Notes Theoret Comput Sci 278:245–260
42. Meyer JJC, de Boer FS, van Eijk RM, Hindriks KV, van der Hoek W (2001) On programming KARO agents. Log J IGPL 9(2):245–256
43. McCarthy J, Hayes PJ (1969) Some philosophical problems from the standpoint of artificial intelligence. Machine intelligence, 4th edn. Edinburgh University Press, Edinburgh, pp 463–502
44. Ma J, Liu W, Hong J, Godo L, Sierra C (2014) Plan selection for probabilistic BDI agents. In: Proceedings of the 26th IEEE international conference on tools with artificial intelligence (ICTAI), pp 83–90
45. Miller T, Muise CJ (2016) Belief update for proper epistemic knowledge bases. In: Proceedings of the 25th international joint conference on artificial intelligence (IJCAI), pp 1209–1215
46. Petrick RPA, Bacchus F (2004) Extending the knowledge-based approach to planning with incomplete information and sensing. In: Proceedings of the 14th international conference on automated planning and scheduling (ICAPS), pp 2–11
47. Reiter R (2001) Knowledge in action: logical foundations for specifying and implementing dynamical systems. The MIT Press, Cambridge
48. Rao AS, Georgeff MP (1991) Modeling rational agents within a BDI-architecture. In: Proceedings of the 2nd international conference on principles of knowledge representation and reasoning (KR). Morgan Kaufmann, pp 473–484
49. Sadek MD (1992) A study in the logic of intention. In: Proceeedings of the 3rd international conference on principles of knowledge representation and reasoning (KR), pp 462–473
50. Sardina S, de Silva L, Padgham L (2006) Hierarchical planning in BDI agent programming languages: a formal approach. In: Proceedings of the 5th international joint conference on autonomous agents and multiagent systems (AAMAS). ACM, pp 1001–1008
51. Searle JR (1990) Collective intentions and actions. Intentions in communication. MIT Press, Cambridge, pp 401–415
52. Shoham Y (2009) Logical theories of intention and the database perspective. J Philos Log 38(6):633–647
53. Shoham Y (2016) Why knowledge representation matters. Commun ACM 59(1):47–49
54. Singh MP (1992) A critical examination of use cohen-levesque theory of intentions. In: Proceedings of the 10th European conference on artificial intelligence (ECAI), pp 364–368
55. Scherl R, Levesque HJ (1993) The frame problem and knowledge producing actions. In: Proceedings of the 11th national conference on artificial intelligence (AAAI). AAAI Press, pp 689–695

56. Sardina S, Lespérance Y (2010) Golog speaks the BDI language. In: Programming multi-agent systems—7th international workshop, ProMAS 2009. Revised selected papers, volume 5919 of lecture notes in computer science. Springer, pp 82–99
57. Shoham Y, Leyton-Brown K (2008) Multiagent systems: algorithmic, game-theoretic, and logical foundations. Cambridge University Press, Cambridge
58. Shapiro S, Sardina S, Thangarajah J, Cavedon L, Padgham L (2012) Revising conflicting intention sets in BDI agents. In: Proceedings of the 11th international conference on autonomous agents and multiagent systems (AAMAS). IFAAMAS, pp 1081–1088
59. Strzalecki T (2014) Depth of reasoning and higher order beliefs. J Econ Behav Org 108:108–122
60. Schut MC, Wooldridge MJ, Parsons S (2004) The theory and practice of intention reconsideration. J Exp Theoret Artif Intell 16(4):261–293
61. Tuomela R, Miller K (1988) We-intentions. J Philos Stud 53:367–389
62. van Benthem J (2007) Dynamic logic for belief revision. J Appli Non Class Log 17(2):129–155
63. van Benthem J, Liu F (2007) Dynamic logic of preference upgrade. J Appl Non Class Log 17(2):157–182
64. van Ditmarsch HP, van der Hoek W, Kooi B (2007) Dynamic epistemic logic. kluwer Academic Publishers, Dordrecht
65. Velleman JD (1989) Practical reflection. Princeton University Press, Princeton
66. van Zee M, Doder D (2016) AGM-style revision of beliefs and intentions. In: Proceedings of the 22nd European conference on artificial intelligence (ECAI), volume 285 of frontiers in artificial intelligence and applications. IOS Press, pp 1511–1519
67. van Zee M, Dastani M, Doder D, van der Torre L (2015) Consistency conditions for beliefs and intentions. In: Proceedings of the 12th international symposium on logical formalizations of commonsense reasoning, pp 152–158
68. van Zee M, Doder D, Dastani M, van der Torre L (2015) AGM revision of beliefs about action and time. In: Proceedings of the 24th international joint conference on artificial intelligence (IJCAI), pp 3250–3256
69. Winikoff M, Padgham L, Harland J, Thangarajah J (2002) Declarative and procedural goals in intelligent agent systems. In: Proceedings of the 8th international conference on principles of knowledge representation and reasoning (KR), pp 470–481
70. Waters M, Padgham L, Sardina S (2015) Improving domain-independent intention selection in BDI systems. Auton Agents Multi Agent Syst 29(4):683–717
71. Weinstein J, Yildiz M (2007) Impact of higher-order uncertainty. Games Econ Behav 60(1):200–212