Resolving Conflicts Between Beliefs, Obligations, Intentions, and Desires

Jan Broersen, Mehdi Dastani, and Leendert van der Torre

Department of Artificial Intelligence
Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The Nederlands
{broersen,mehdi,torre}@cs.vu.nl

http://www.cs.vu.nl/~boid/

Abstract. This paper provides a logical analysis of conflicts between informational, motivational and deliberative attitudes such as beliefs, obligations, intentions, and desires. The contributions are twofold. First, conflict resolutions are classified based on agent types, and formalized in an extension of Reiter's normal default logic. Second, several desiderata for conflict resolutions are introduced, discussed and tested on the logic. The results suggest that Reiter's default logic is too strong, in the sense that a weaker notion of extension is needed to satisfy the desiderata.

1 Introduction

Various competing agent decision models have been proposed, and it is still unclear which type of model should be used in which type of application. For example, some decision models are based on goal-based planning or on variants of decision theory like qualitative decision theory [13, 1], other models are based on cognitive models like belief-desire-intention models [5, 14], and yet other models are based on social concepts like obligations and norms [6, 20, 19], as in deontic action programs [8]. Typically, the decision model is based on an attempt to reach goals, satisfy desires, or fulfill obligations. In the Belief-Obligation-Intention-Desire or BOID architecture [4] decision models are considered in which the main problem is not finding out how to reach goals, satisfy desires or fulfill obligations, but in which the main problem is to resolve conflicts between them.

The BOID logic discussed in this paper is an abstraction of the BOID architecture. For conflicts so-called extensions are constructed and one extension is selected, an idea adopted from Thomason's BDP logic [18], which is in turn based on Reiter's default logic [15]. In particular, BDP logic is based on conflict resolution for conditional beliefs and desires, which is extended in the BOID logic with conditional obligations and intentions borrowed from respectively deontic action programs [8] and BDI logic [5, 14]. The BOID logic is an *abstraction* from the BOID architecture, in the sense that in the latter the components may not

contain rules or be based on propositional logic, and in case of limited resources the extensions may not be fixpoints.

The contributions of this paper are twofold:

1. We give a classification of conflict resolutions between *conditional* beliefs, obligations, intentions, and desires. Extending the BDP logic with obligations and intentions increases the number of possible conflicts dramatically. In all realistic conflict resolutions beliefs override obligations, intentions, and desires; in stable conflict resolutions intentions override desires and obligations; in unstable conflict resolutions desires and obligations override intentions; in selfish conflict resolutions desires override obligations; and in social conflict resolutions obligations override desires.
2. We propose several desired and undesired properties to analyze this overriding encoded in the BOID logic. As our running example we show how beliefs override desires to block wishful thinking. For example, assume that you believe that you get wet irrespective of your desire to stay dry. This would, according to Thomason, imply that the belief to get wet overrides the desire to stay dry, in the sense that in your planning you will assume that you will get wet.

The layout of this paper is as follows. In Section 2 different types of conflicts are introduced and a classification of conflict resolution types is discussed. In Section 3 the BOID logic and its extension calculation scheme are introduced. In Section 4 properties for wishful thinking are analyzed, and in Section 5 we discuss the properties of extensions provided by the BOID calculation scheme.

2 Beliefs, obligations, intentions and desires

Reasoning about beliefs, obligations, intentions and desires has been discussed in practical reasoning in philosophy [21, 2], and its formalization to build intelligent autonomous agents has more recently been discussed in qualitative decision making in artificial intelligence [7, 8, 14, 18]. On closer inspection each of these four concepts consists of related (though often quite distinct) concepts, for example respectively knowledge and defaults, prohibitions and permissions, commitments and plans, wishes and wants. All these concepts are grouped into these four classes due to their role in the decision making process: beliefs are informational states – how the world is expected to be – obligations and desires are the external and internal motivational states, and intentions are the deliberative states.

2.1 Conflict resolutions

A conflict resolution type is an order of overruling. Given four attitudes, there are twenty-four possible total orders of overruling, and many more partial orders in which for example desires and obligations are equivalent. In this paper, we only consider those orders according to which beliefs overrule any other attitude. This

reduces the number of possible total overruling orders to six. Some examples of conflict resolution are given below.

- A conflict between a belief and a prior intention means that an intended action can no longer be executed due to the changing environment. Beliefs therefore overrule the prior intention, which is retracted. Any derived consequences of this prior intention are retracted too. Of course, one may allow prior intentions to overrule beliefs, but this results in unrealistic behavior.
- A conflict between a belief and an obligation or desire means that a violation has occurred. As observed by Thomason [18], the beliefs must override the desires or otherwise there is wishful thinking; the same argument applies to obligations.
- A conflict between a prior intention and an obligation or desire means that you now should or want to do something else than you intended before. Here prior intentions override the latter because it is exactly this property for which intentions have been introduced: to bring stability. However, in cases of intention reconsideration such conflicts may be resolved otherwise. For example, if I intend to go to the cinema but I am obliged to visit my mother, then I go to the cinema unless I reconsider my intentions.

2.2 Detecting versus resolving conflicts

Further specifying and implementing the conflict types leads to several complications. It may seem that we can use one of the many approaches to conflict resolution developed in other areas of artificial intelligence like for example diagnosis [16], default reasoning or fusion of knowledge and databases. However, in these approaches a conflict is defined as a *minimal* set, in the sense that if two sets are conflict sets then one of the sets cannot be a strict subset of the other one. Whereas minimal sets may be useful to detect conflicts, it is not sufficient to resolve them.

An example has been given by Dignum *et. al.* [7], who discuss an extension of the BDI logic with obligations. In this example, there is a guy called Al who has an obligation to perform a task for Bob and another incompatible obligation to perform a task for Chris. Moreover, Al has the norm that he should tell Bob if he does not intend to meet this obligation. The problem discussed in the paper is that the existence of the norm should affect Al's decision on whether to intend to fulfill his obligation:

> "Consider Al's obligation above, until he actually commits to not meeting his obligation to Bob, the need to tell Bob does not exist, yet the *potential* for it may have a significant impact on his decision on whether to do the task for Bob. For example, imagine that the task is trivial (i.e., the direct consequences of not doing the task are small), but the social consequences of not informing Bob are very high (i.e., Al is perceived as unreliable)." [7, p.115]

The point is thus that to resolve the conflict we cannot restrict ourselves to the minimal set (the two obligations), but we have to consider the whole set. In general, agents should consider the effects of actions before committing to it. This is the reason why in the BOID logic complete extensions are constructed before one is selected, instead of solving a conflict as one is encountered.

3 BOID Logic

In this section we discuss the BOID logic. First, we consider Reiter's normal default logic and Thomason's BD logic.

3.1 Reiter's normal default logic

Reiter defined extensions of normal default theories as follows, where we write $\alpha \hookrightarrow w$ for $(\alpha : Mw/w)$ and we write $\langle W, D\rangle$ instead of $\langle D, W\rangle$.

Definition 1. *[15, Def. 1] Let $\Delta = \langle W, D\rangle$ be a closed default theory, so that every default of D has the form $\alpha \hookrightarrow w$ where α and w are both closed wffs of a (first-order) language L, and let $Th_L(S)$ be the consequence set of S in L. For any set of closed wffs $S \subseteq L$ let $T(S)$ be the smallest set of closed formulas from L satisfying the following three properties:*

1. $W \subseteq T(S)$
2. $Th_L(T(S)) = T(S)$
3. *If $\alpha \hookrightarrow w \in D$, $\alpha \in T(S)$ and $\neg w \notin S$, then $w \in T(S)$.*

A set of closed wffs $E \subseteq L$ is an extension *for Δ iff $T(E) = E$, i.e. iff E is a fixed point of the operator T.*

A well-known theorem of Reiter's paper is the following more intuitive characterization of extensions.

Theorem 1. *[15, Th. 2.1.] Let $E \subseteq L$ be a set of closed wffs, and let $\Delta = \langle W, D\rangle$ be a closed default theory. Define*
$E_0 = W$
and for $i \geq 0$
$E_{i+1} = Th_L(E_i) \cup \{w \mid \alpha \hookrightarrow w \in D$ *where* $\alpha \in E_i$ *and* $\neg w \notin E\}$
Then E is an extension for Δ iff
$E = \cup_{i=0}^{\infty} E_i$.

3.2 Thomason's BD logic

Thomason [18] proposes a so-called BDP-logic for beliefs, desires and planning which is capable of modeling a wide range of common-sense practical arguments, and which can serve as a more general and flexible model for the decision making process. Thomason first discusses the BD formalism and focuses on the interaction between beliefs and desires. The basic idea is to model beliefs and desires

both as Reiter defaults [15], *without modalities for belief or desire*, such that the extensions contain all the derived atoms. That is, a BD-basis is a tuple $\langle Obs, NB, ND \rangle$ with Obs a set of formulas, NB a set of B-defaults 'if a then I believe x' written as $a \overset{B}{\hookrightarrow} x$, and ND a set of D-defaults 'if a then I desire x' written as $a \overset{D}{\hookrightarrow} x$. Extensions are built iteratively by applying default rules without distinguishing between beliefs and desires, so for example, the BD-basis $\langle \{a\}, \{a \overset{B}{\hookrightarrow} b\}, \{b \overset{D}{\hookrightarrow} c\} \rangle$ has as an extension $Th_L(\{a, b, c\})$. But then, there are two types of conflicts:

- Conflicts between a belief and a desire lead to overriding of desire by belief to block wishful thinking.
- Other conflicts, for instance, one between two desires or between two beliefs lead to multiple extensions.

Central in Thomason's iterative calculation of extensions is that belief and desire defaults are treated equally, except for the situations where a desire default conflicts with a subset of the belief defaults applied to the formulas derived in the sequence so far. In such a conflicting situation, the belief defaults are applied preferably.

3.3 BOID logic

The BOID logic extends Thomason's idea with obligations and intentions (like [7]) resulting in the BOID logic. This logic consists of four sets of propositional logical formulae that represent the four attitudes *Beliefs*, *Obligations*, *Intentions*, and *Desires*. One reason for this extension is to incorporate elements of the social level, i.e. social commitments, to formalize for example social agents and social rationality. The BOID logic is parameterized in order to resolve conflicts between attitudes according to a complete conflict resolution type. This input parameter constrains the order in which derivation steps for different sets are undertaken and characterizes the type of conflict resolution.

The iterative procedure of the BOID calculation scheme is given as an extension of Reiter's more intuitive characterization of extensions in Theorem 1. As in [12] we assume that there is an order on the rules, which we represent by ρ. In order to define this calculation scheme, we first define an ordering function ρ that represents the conflict resolution type. In case of multiple applicable rules, one with the lowest ρ value is applied.

Definition 2. *Let L be a propositional language and S be a set of ordered pairs of L written as $\alpha \hookrightarrow w$ and called rules. An agent type is a set of functions ρ from S to the integers.*

The agent type is usually expressed as a constraint. For example, if S is the union of beliefs B and desires D, then the agent type 'realistic' is expressed by the constraint that for all $r_b \in B$ and $r_d \in D$ we have $\rho(r_b) < \rho(r_d)$. Given a specific agent type, the calculation scheme for building extensions is defined as follows.

Definition 3 (BOID Calculation Scheme). *Let L be a propositional language, let a tuple $\Delta = \langle W, B, O, I, D \rangle$ be a BOID theory with W a subset of L and B, O, I and D sets of ordered pairs of L written as $\alpha \hookrightarrow w$, let ρ be a function that assigns to each rule in $B \cup O \cup I \cup D$ a unique integer, and S a subset of L. Moreover, let*

$\rho_{\min}(BOID, S) = \min\{\rho(\alpha \hookrightarrow w) \mid \alpha \hookrightarrow w \in B \cup O \cup I \cup D,\ \alpha \in S, \neg w \notin S\}$

$\min(BOID, S) = w$ *s.t.* $\alpha \hookrightarrow w \in B \cup O \cup I \cup D$, $\rho(\alpha \hookrightarrow w) = \rho_{\min}(BOID, S)$

Define

$E_0 = W$

and for $i \geq 0$

$E_{i+1} = Th_L(E_i \cup \{\min(BOID, S)\})$ *if such a minimal element exists,*

$E_{i+1} = E_i$ *otherwise.*

Then $E \subseteq L$ is an extension for Δ of agent type A iff $\exists \rho \in A$ s.t. $E = \cup_{i=0}^{\infty} E_i$.

3.4 Discussion

Space does not permit us to compare the BOID logic in any detail with classical approaches to specification and verification of agent systems, based on for example modal and temporal logics like BDICTL [14, 17]. We just make the following remarks:

- The analysis of conflicts in BDICTL is limited, in the sense that for example two conflicting desires cannot be represented in a consistent way.
- The representation of conditionals in BDICTL is not straightforward, whereas this is a central issue in BOID logics.
- To compare BDICTL and BOID logic the propositional base language of BOID logic must be replaced by BDICTL.[1]
- Each state in the BOID logic has the same logic, i.e. normal default logic, but it can be further developed such that for example for obligations and desires we do not have that inputs are included in the extensions, see [10, 11].

A second and more interesting issue is the comparison of BOID logic with extensions of default logic such as preferred answer sets [3]. One of the results obtained here is that a greedy approach as used in the BOID logic (always try to apply the rule with the highest priority) may lead to globally suboptimal results (e.g. by first applying a rule of priority 3 instead of one of priority 2 we can thereafter apply a rule of priority 1 - by convention the highest priority). The greedy approach is justified by the fact that the BOID logic is only an idealization. In reality fixpoints may never be reached due to limited resources.

[1] This extension is not as interesting as it may seem at first sight, because the extensions are used in the agent's planning and to plan to achieve goal p it is irrelevant whether there is an intention, desire or obligation to see to it that p. Note that it is important in the implementation [4]. There have also been convincing philosophical arguments to do without modal operators, see [9]. Advantages of this extension are the formalization of more complex notions like permissions and ignorance.

4 No wishful thinking

Thomason [18] argues that beliefs override desires with the following example. If you think it is going to rain and you believe that if it rains, you will get wet, and you would not like to get wet, then you have to conclude that you get wet. Beliefs therefore prevail in conflicts with desires.

How can we formulate this intuition as a property of extensions? In this section we consider three properties that guarantee that beliefs override desires. These properties are not restricted to one particular approach, but can be applied to any extension-based approach. To facilitate the definitions of the properties in this section we use the following definition.

Definition 4. *Let $\Delta = \langle W, B, D\rangle$ be a BD theory, where W is a set of propositional sentences and B and D ordered pairs of such sentences. We write $E_{BD}(\Delta)$ for the set of all extensions of a propositional BD theory, and for representational convenience we write $E_{BD}(W, B, D)$ for $E_{BD}(\langle W, B, D\rangle)$.*

4.1 Applied Desire rules

The intuition behind Property 1 of no wishful thinking below is as follows. If in a conflict between a desire and a belief the desire rule is removed, then the extension cannot increase because the belief rule already had priority over the desire rule. In other words, the removal of desires can only decrease the extension, not increase it or remove it.

Property 1 (Applied D rules; first attempt). For each $E' \in E_{BD}(W, B, D')$ and $D \subseteq D'$ there is an $E \in E_{BD}(W, B, D)$ such that $E \subseteq E'$.

The following example illustrates that Property 1 is, unfortunately, too strong.

Example 1. Let $\Delta_1 = \langle \emptyset, \emptyset, \{\top \overset{D}{\hookrightarrow} p\}\rangle$ and $\Delta_2 = \langle \emptyset, \emptyset, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} \neg p\}\rangle$. Intuitively we have $E_{BD}(\Delta_1) = \{Th_L(p)\}$ and $E_{BD}(\Delta_2) = \{Th_L(p), Th_L(\neg p)\}$. But for $E' = Th_L(\neg p) \in E_{BD}(\Delta_2)$, there is no $E \in E_{BD}(\Delta_1)$ such that $E \subseteq E'$. This example contradicts Property 1.

Example 1 also illustrates where our first attempt goes wrong. The problem is that D may contain rules which have not been used to build E' of $E_{BD}(W, B, D')$, but they may be used when building E of $E_{BD}(W, B, D)$. In the example, this rule was $\top \overset{D}{\hookrightarrow} p$. We first introduce a definition to identify an extension with the set of rules which are applied in it (sometimes called its generators).

Definition 5 (Applied rules). *Let $\Delta = \langle W, B, D\rangle$ be a BD theory and let the set E be one of its extensions. The set of applied rules in extension E is $R_B(\Delta, E) = \{\alpha \overset{B}{\hookrightarrow} w \in B \mid \alpha \wedge w \in E\}$, $R_D(\Delta, E) = \{\alpha \overset{D}{\hookrightarrow} w \in D \mid \alpha \wedge w \in E\}$, and $R(\Delta, E) = R_B(\Delta, E) \cup R_D(\Delta, E)$.*

The following Property 2 is a weaker form of the Property 1, because we have $R_D(\langle W, B, D' \rangle, E') \subseteq D'$.

Property 2 (Applied D rules, second attempt). For each $E' \in E_{BD}(W, B, D')$ and $D \subseteq R_D(\langle W, B, D' \rangle, E')$ there is an $E \in E_{BD}(W, B, D)$ such that $E \subseteq E'$.

The following example reconsiders Example 1 and illustrates that Property 2 does not have the undesirable behavior.

Example 2. $\Delta_1 = \langle \emptyset, \emptyset, \{\top \overset{D}{\hookrightarrow} p\} \rangle$, $\Delta_2 = \langle \emptyset, \emptyset, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} \neg p\} \rangle$. As mentioned in Example 1, $E_{BD}(\Delta_1) = \{Th_L(p)\}$ and $E_{BD}(\Delta_2) = \{Th_L(p), Th_L(\neg p)\}$ contradict Property 1. However, it does not contradict Property 2, because for $E' = Th_L(\neg p)$ we have $R_D(\langle \emptyset, \emptyset, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} \neg p\} \rangle, E') = \{\top \overset{D}{\hookrightarrow} \neg p\}$, and this set is not a superset of the desire rules in Δ_1.

The following simple examples further illustrate Property 2.

Example 3. $\Delta_1 = \langle \emptyset, \{\top \overset{B}{\hookrightarrow} p, q \overset{B}{\hookrightarrow} \neg p\}, \emptyset \rangle$, $\Delta_2 = \langle \emptyset, \{\top \overset{B}{\hookrightarrow} p, q \overset{B}{\hookrightarrow} \neg p\}, \{\top \overset{D}{\hookrightarrow} q\} \rangle$. If $E_{BD}(\Delta_1) = \{Th_L(p)\}$, then each element of $E_{BD}(\Delta_2)$ has to contain $Th_L(p)$, and $E_{BD}(\Delta_2)$ thus cannot contain for example $Th_L(q \wedge \neg p)$.

Example 4. $\Delta_1 = \langle \emptyset, \{\top \overset{B}{\hookrightarrow} p\}, \emptyset \rangle$, $\Delta_2 = \langle \emptyset, \{\top \overset{B}{\hookrightarrow} p\}, \{\top \overset{D}{\hookrightarrow} q, p \overset{D}{\hookrightarrow} \neg q\} \rangle$. If $E_{BD}(\Delta_1) = \{Th_L(p)\}$, then each element of $E_{BD}(\Delta_2)$ has to contain $Th_L(p)$, but $E_{BD}(\Delta_2)$ still can contain for example $Th_L(p, q)$ and $Th_L(p, \neg q)$.

Example 5. $\Delta_1 = \langle \emptyset, \{p \overset{B}{\hookrightarrow} \neg q\}, \{\top \overset{D}{\hookrightarrow} p\} \rangle$, $\Delta_2 = \langle \emptyset, \{p \overset{B}{\hookrightarrow} \neg q\}, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} q\} \rangle$. If $E_{BD}(\Delta_1) = \{Th_L(p, \neg q)\}$ then generalized no-wishful thinking based on applied desire rules implies $Th_L(p, q) \notin E_{BD}(\Delta_2)$. However, note that $Th_L(q)$ may be in $E_{BD}(\Delta_2)$ (verification left to the reader).

Example 6. $\Delta_1 = \langle \emptyset, \{p \overset{B}{\hookrightarrow} \neg q, r \overset{B}{\hookrightarrow} q\}, \{\top \overset{D}{\hookrightarrow} p\} \rangle$, $\Delta_2 = \langle \emptyset, \{p \overset{B}{\hookrightarrow} \neg q, r \overset{B}{\hookrightarrow} q\}$ $, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} r\} \rangle$. If $E_{BD}(\Delta_1) = \{Th_L(p, \neg q)\}$ then we have that the sets $Th_L(p, r, \neg q), Th_L(p, r, q) \notin E_{BD}(\Delta_2)$ but $Th_L(p, \neg q)$ and $Th_L(r, q)$ may be in $E_{BD}(\Delta_2)$ (analogous to the previous example, verification left to the reader).

A simple instance of this generalized no-wishful thinking property, which we call *Restricted no-wishful thinking*, is the case where D is the empty set. This property says that every BD extension extends a B extension.

Property 3 (Restricted Applied D rules). For each $E' \in E_{BD}(W, B, D')$ there is an $E \in E_{BD}(W, B, \emptyset)$ such that $E \subseteq E'$.

4.2 Applied Belief rules

The second way to define no wishful thinking we consider is to look for a constraint on just the beliefs. The set of applicable belief rules of one extension cannot be a strict subset of the applicable belief rules of another extension.

Property 4 (Applied B rules, first attempt). For all $E_1, E_2 \in E_{BD}(\Delta)$ we have $R_B(\Delta, E_1) \subseteq R_B(\Delta, E_2)$ implies $R_B(\Delta, E_1) = R_B(\Delta, E_2)$.

Unfortunately, this property does not give intuitive results, as the following example illustrates.

Example 7. Let $\Delta = \langle \emptyset, \{p \overset{B}{\hookrightarrow} q\}, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} \neg p\}\rangle$. Intuitively we have $E_{BD}(\Delta) = \{Th_L(p,q), Th_L(\neg p)\}$, i.e. $R_B(\Delta, Th_L(\neg p)) \subset R_B(\Delta, Th_L(p,q))$. This example contradicts Property 4.

The following property is a variant of Property 1. The removal of desires can only decrease the set of applied belief rules, not increase it or remove it.

Property 5 (Applied B rules, second attempt). For each $E' \in E_{BD}(W, B, D')$ and $D \subseteq D'$ there is an $E \in E_{BD}(W, B, D)$ such that $R_B(\langle W, B, D\rangle, E) \subseteq R_B(\langle W, B, D'\rangle, E')$.

Property 5 gives the desired results for the rule sets in Example 1 and 7. However, Example 8 is a generalization of these two examples that shows why Property 5 has similar problems as Property 1.

Example 8. $\Delta_1 = \langle \emptyset, \{p \overset{B}{\hookrightarrow} q\}, \{\top \overset{D}{\hookrightarrow} p\}\rangle$, $\Delta_2 = \langle \emptyset, \{p \overset{B}{\hookrightarrow} q\}, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} \neg p\}\rangle$. Intuitively we have $E_{BD}(\Delta_1) = \{Th_L(p,q)\}$ and $E_{BD}(\Delta_2) = \{Th_L(p,q), Th_L(\neg p)\}$. However, for $E' = Th_L(\neg p) \in E_{BD}(\Delta_2)$ there is no $E \in E_{BD}(\Delta_1)$ such that $R_B(\Delta_1, E) \subseteq R_B(\Delta_2, E')$.

The following property is analogous to Property 2.

Property 6 (Applied B rules, third attempt). For each $E' \in E_{BD}(W, B, D')$ and $D \subseteq R_D(\langle W, B, D'\rangle, E')$ there is an $E \in E_{BD}(W, B, D)$ such that we have $R_B(\langle W, B, D\rangle, E) \subseteq R_B(\langle W, B, D'\rangle, E')$.

The following example illustrates the distinction between Property 2 and 6.

Example 9. Let $\Delta_1 = \langle \emptyset, \emptyset, \{\top \overset{D}{\hookrightarrow} p\}\rangle$ and $\Delta_2 = \langle \emptyset, \emptyset, \{\top \overset{D}{\hookrightarrow} p, \top \overset{D}{\hookrightarrow} q\}\rangle$. If $E_{BD}(\Delta_1) = Th_L(p)$ then we cannot have $Th_L(\emptyset)$ in $E_{BD}(\Delta_2)$ according to generalized no wishful thinking based on applied desire rules, but it can be according to generalized no wishful thinking based on applied belief rules.

Intuitively we do not want $Th_L(\emptyset)$ in $E_{BD}(\Delta_2)$, but the reason for this is not the blocking of wishful thinking. Property 6 seems therefore a better characterization of no-wishful thinking than Property 2.

Property 7 is analogous to Property 3.

Property 7 (Restricted Applied B rules). For each $E' \in E_{BD}(W, B, D')$ there is an $E \in E_{BD}(W, B, \emptyset)$ such that $R_B(E) \subseteq R_B(E')$.

4.3 Abnormal Belief rules

The third way to define no wishful thinking is not based on applied rules but on rules which could not be applied, which we call abnormal rules. These abnormal rules are defined analogously to applied rules in Definition 5.

Definition 6 (Abnormal rules). *Let $\Delta = \langle W, B, D\rangle$ be a BD theory and let the set E be one of its extensions. The set of abnormal rules is represented by* $Ab_B(\Delta, E) = \{\alpha \overset{B}{\hookrightarrow} w \in B \mid \alpha \wedge \neg w \in E\}$.

Generalized no wishful thinking based on abnormal belief rules is defined analogous to generalized no wishful thinking property based on applied rules in Property 2 and 6.

Property 8 (Abnormal B rules, first attempt). For each $E' \in E_{BD}(W, B, D')$ and $D \subseteq R_D(\langle W, B, D'\rangle, E')$ there is an $E \in E_{BD}(W, B, D)$ such that we have $Ab_B(\langle W, B, D\rangle, E) \supseteq Ab_B(\langle W, B, D'\rangle, E')$.

The following example illustrates that generalized wishful thinking based on abnormal belief rules is different from generalized wishful thinking based on applied desire or belief rules in Property 2 and 6.

Example 10. Let $\Delta_1 = \langle\{\neg q\}, \{p \overset{B}{\hookrightarrow} q\}, \emptyset\rangle$ and $\Delta_2 = \langle\{\neg q\}, \{p \overset{B}{\hookrightarrow} q\}, \{\top \overset{D}{\hookrightarrow} p\}\rangle$. If $Th_L(\neg q, p) \notin E_{BD}(\Delta_1)$ then according to generalized no wishful thinking based on abnormal belief rules $Th_L(\neg q, p) \notin E_{BD}(\Delta_2)$. However, according to generalized wishful thinking based on applied desire or belief rules, it may be that $Th_L(\neg q, p) \notin E_{BD}(\Delta_1)$ as well as $Th_L(\neg q, p) \in E_{BD}(\Delta_2)$.

5 BOID properties

The first BOID property is called *Existence* and says that there is at least one BD extension, if the facts W are consistent. This is a very desirable and crucial property for decision making agents, because an agent needs an extension to act rationally. Otherwise the agent is stuck or starts to make random movements.

Property 9 (Existence). $E_{BD}(W, B, D) \neq \emptyset$ if $\bot \notin Th_L(W)$.

The second BOID property we discuss here is called BD maximality, and says that if a rule can be applied then it is applied. That is, we go as far as possible. This property implies that the set of BD extensions are a subset of the set of Reiter extensions where the set of rules consists of the union of belief and desire rules. We write $E_R(\Delta)$ for the set of all Reiter extensions of a propositional default theory, and if we consider Reiter extensions of BD theories consisting of $\alpha \overset{B}{\hookrightarrow} w$ and $\alpha \overset{D}{\hookrightarrow} w$, then we ignore the superscript above the arrows, i.e. we interpret $E_R(\langle W, BD\rangle)$ as $E_R(\langle W, \{\alpha \hookrightarrow w \mid \alpha \overset{B}{\hookrightarrow} w \in BD \text{ or } \alpha \overset{D}{\hookrightarrow} w \in BD\rangle)$.

Property 10 (BD maximality). $E_{BD}(W, B, D) \subseteq E_R(W, B \cup D)$

The following example reconsiders Example 10 and questions the BD maximality property.

Example 11. Let $\Delta = \langle\{\neg q\}, \{p \overset{B}{\hookrightarrow} q\}, \{\top \overset{D}{\hookrightarrow} p\}\rangle$ (we can also replace $\neg q$ by $\top \overset{B}{\hookrightarrow} \neg q$). We have $E_R(\Delta) = \{Th_L(\neg q, p)\}$, and thus with the existence property and the BD maximality property we can derive $E_{BD}(\Delta) = \{Th_L(\neg q, p)\}$. However, $\neg q \wedge p$ implies that to fulfill the desire for p we get into a situation in which something happens which we believe will not happen, namely the exception to the belief that p implies q.[2]

The following theorem and its corollary suggest that BD maximality is too strong (see [11] for an alternative notion of extension).

Theorem 2. *No-wishful thinking based on applied desire rules (Property 2), on applied belief rules (Property 6) or on abnormal belief rules (Property 8) conflicts with BD maximality (Property 10) together with Existence (Property 9).*

Proof. *For the applied rules, see Example 5 and 6. For the abnormal rules, see Example 10 and 11.*

Corollary 1. *The BOID logic does not satisfy any of the three notions of no-wishful thinking discussed in this paper.*

6 Concluding remarks

We have discussed possible conflict types that may arise within or among informational and motivational attitudes and explained how these conflicts can be resolved within the BOID calculation scheme. The resolution of conflicts is based on Thomason's idea of prioritization, which is considered in the BOID logic as the order of derivations from different types of attitudes. We have shown that the order of derivations determines the type of conflict resolution method. For example, deriving desire before beliefs produces wishful thinking and deriving obligations before desires produces sociality. We have also introduced some desired and undesired properties, and checked whether some conflict resolution methods satisfied the properties.

Two issues for further research are the generalization of properties for overriding to multiple attitudes, and for other input/output logics [10, 11] than Reiter's normal default logic. Although the properties are defined independent of the logic, both Definition 5 and 6 of applied and abnormal rules must be adapted if we allow for e.g. reasoning by cases (e.g. $E_{BD}(\emptyset, \{\alpha \overset{B}{\hookrightarrow} w, \neg\alpha \overset{B}{\hookrightarrow} w\}, \emptyset) = Th_L(w)$).

[2] Example 11 is not very convincing, because of the following two reasons. First, the behavior in Example 11 seems to be what is expected from *conditional* rules. If you do not like it, then you can formalize the belief rule with $\top \overset{B}{\hookrightarrow} p \rightarrow q$, where $\rightarrow$ is a material implication. Second, in the discussion in Example 11 the rules are used as a kind of causal rules. However, if the conditional $p \overset{B}{\hookrightarrow} q$ represents a causal relation, then the world will change such that $\neg q$ will turn into q.

Acknowledgment

Thanks to Salem Benferhat, Zisheng Huang and Joris Hulstijn for discussions on the issues discussed in this paper.

References

1. C. Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the KR'94*, pages 75–86, 1994.
2. Michael E. Bratman. *Intention, plans, and practical reason.* Harvard University Press, Cambridge Mass, 1987.
3. G. Brewka and T. Eiter. Preferred answer sets for extended logic programs. *Artificial Intelligence*, 109:297–356, 1999.
4. J. Broersen, M. Dastani, Z. Huang, J. Hulstijn, and L. van der Torre. The BOID architecture: Conflicts between beliefs, obligations, intentions, and desires. In *Proceedings of International Conference on Autonomous Agents (AA'01)*, 2001.
5. P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
6. F. Dignum. Autonomous agents and norms. *Artificial Intelligence and Law*, 7:69–79, 1999.
7. F. Dignum, D. Morley, E.A. Sonenberg, and L. Cavedon. Towards socially sophisticated BDI agents. In *Proceedings of the ICMAS 2000*, pages 111–118, 2000.
8. Thomas Eiter, V.S. Subrahmanian, and George Pick. Heterogeneous active agents I: Semantics. *Artificial Intelligence*, 108 (1-2):179–255, 1999.
9. D. Makinson. On a fundamental problem of deontic logic. In *Norms, logics and information systems*, pages 29–53. IOS Press, 1999.
10. D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.
11. D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic*, 30(2):155–185, 2001.
12. V.W. Marek and M. Truszczynski. *Nonmonotonic logic: Context-dependent reasoning.* Springer, Berlin, 1993.
13. J. Pearl. From conditional oughts to qualitative decision theory. In *Proceedings of the UAI'93*, pages 12–20, 1993.
14. A. Rao and M. Georgeff. BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95)*, pages 312–319, 1995.
15. R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.
16. R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.
17. K. Schild. On the relationship between BDI logics and standard logics of concurrency. *Autonomous Agents and Multi Agent systems*, 2000.
18. R. Thomason. Desires and defaults: a framework for planning with inferred goals. In *Proceedings of the KR'2000*, pages 702–713. Morgan Kaufmann, 2000.
19. L. van der Torre and Y. Tan. Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, 27:49–78, 1999.
20. L. van der Torre and Y. Tan. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law*, 7:51–67, 1999.
21. G.H. von Wright. *Norms, truth and logic. Practical Reason.* Blackwell, Oxford, 1983.