BOID*: Autonomous Goal Deliberation through Abduction

Stipe Pandžić
Utrecht University
Utrecht, The Netherlands
s.pandzic@uu.nl

Jan Broersen
Utrecht University
Utrecht, The Netherlands
j.m.broersen@uu.nl

Henk Aarts
Utrecht University
Utrecht, The Netherlands
h.aarts@uu.nl

ABSTRACT

The original BOID [5] is a cognitive architecture that unifies Belief, Obligation, Intention and Desire rules to calculate which actions should an agent undertake next. In the current paper, we adapt the original BOID with an aim to model *autonomous* agency. The new BOID* architecture is able to capture *anticipation* that we believe to be one of the hallmarks of autonomous agency. We focus on developing algorithms for anticipatory reasoning through a new BOID* goal deliberation component. The key method that BOID* introduces is abductive reasoning as a way to represent motivational attitudes, such as desires and obligations. As a result of deliberation via abduction, BOID* specifies intention revision procedures that connect motivational and informational attitudes. The BOID* is a part of the project to build autonomous AI models that make explicit the reasoning behind adopting future goals, prioritizing selected goals and forming intentions.

KEYWORDS

Agent Architecture; Anticipation; Intention; Goal Deliberation; Autonomous Agents; Default Reasoning; Abductive Reasoning

ACM Reference Format:
Stipe Pandžić, Jan Broersen, and Henk Aarts. 2022. BOID*: Autonomous Goal Deliberation through Abduction. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 9 pages.

1 INTRODUCTION

The BOID architecture [5, 6] was developed to provide an algorithm for building 'cognitive' agents that reason with their beliefs, desires, obligations and intentions to arrive at propositional goals that can be pursued using standard planning mechanisms. The core challenge of the BOID was to resolve conflicts between the modeled mental attitudes by way of a defeasible, yet simple and implementable reasoning mechanism. In the current paper, we extend the original BOID with an aim to model *autonomous* agency. The most important step toward autonomous agency is adding an *anticipatory* component for *telic* reasoning to the current algorithm for goal generation. The underlying intuition is that anticipation and purpose-based action are hallmarks of autonomous agency.

One way to substantiate the components of the BOID architecture is to use default logic rules [16], which are close to the rules used in the original BOID papers. While default logic is a suitable basis for a theory of reasons [12], it does not provide a full account of reasoning with motivational attitudes. The general form of default rules reads as follows: "if a and if it is not inconsistent to derive b, an agent believes, is obliged, intends, or desires that also b", the antecedent a is considered to be a reason for b. The BOID allows a to be a factual statement such that b is conditioned on a, but it also allows a to be a statement about a possible anticipated state of affairs conditioned on b taking place. The BOID rules thus represent both *factum* and *faciendum* as antecedents of future-directed defaults. In this paper, we think of defaults as rules only appropriate to represent reactive behavior that starts from observations, which trigger actions based on observations, but not for anticipatory behavior.

We suggest that the type of reasoning involving motivational attitudes, such as desires and obligations, is distinctively anticipatory. This requires the following shift of information flow enabling telic reasoning about goals: agents need to start from the anticipated purpose of their action, instead of starting from observed facts, as mandated by the original BOID default rules. Informally, new rules will read as "*in order to* a and if it is not inconsistent to derive b, an agent needs to believe, is obliged, intends, or desires that also b". Instead of using 'reactive' default rules, the latter type of rules will be formalized as abductive reasoning. Enriching the BOID algorithm with a teleological component opens up a possibility to factor in behavior that does not immediately start from observable facts, but rather directly lines up with projections of anticipated states of affairs. Our inspiration to build a logic for anticipatory goal generation comes from the research on biological systems that emphasizes anticipation as inextricably linked to the level of autonomy in self-regulating living systems [9].

2 ABDUCTION AND MENTAL ATTITUDES

In a nutshell, the system described in this paper adapts deductive, abductive and inductive reasoning patterns [14] to allow forward (default) and backward (abductive) 'jumping to conclusions'. To illustrate the three patterns, consider the following statements:

Birds are flying animals (*Rule*),
This animal is a bird (*Case*) and
This animal flies (*Result*).

Inductive reasoning is what establishes a rule that is taken to hold 'by default': from a large but finite number of cases, a generalization is formulated in terms of a rule. In particular, after we attest to a sufficient number of birds that fly, the case and the result instances support the rule. In deductive reasoning, we simply apply the rule to the case that some animal is a bird to infer the result that the animal flies. Since our BOID rules will only be required to hold "other things being equal", as in "Normally, birds are flying animals", the resulting conclusion will not be deductively valid, but only defeasibly. Finally, in abductive reasoning, we start from the result, which is in our example the statement that some animal flies and, using the rule at hand, hypothesize that this animal is a bird. Notice

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

that both defeasible inferences and abductive inferences use the rules that are inductively confirmed, but, in abductive inferences, we use the rule in the reversed premise-conclusion order.

Abductive reasoning is notoriously unsafe, even if the rule categorically claims that exactly *all* birds are flying animals. When the rule is a mere default assumption, as in our example, the hypothesis is only little more than a guess. This uncertainty will be reflected in the way we formalize the BOID* motivational attitudes with abduction. For example, there will possibly be multiple competing extensions generated from motivational attitudes. Moreover, abductive inferences that extend the original BOID concern future-oriented reasoning where agents start not from an observation or what is 'given', but from an anticipated state of affairs that they desire to, ought to or intend to attain.

In this paper, we will talk about 'abductive rules', by which we simply understand abductive uses of an inductive generalization. We think that this abuse of terminology is on a par with the use of the term 'default rule', which can be simply thought of as signaling the default use of an inductive generalization [19]. In both cases, the terms are used as a shorthand for patterns of reasoning underlain by inductively obtained rules.

There are two main roles of the abductive reasoning part in BOID* theories. The first role is that of (re)aligning an agent's goals with some values that the agent accepts. The second role is to specify practical requirements that are necessary to follow up on an intention. The latter role is very close to the role of abductive rules in planning, as discussed in the original BOID setting [5, p. 443].

We will give two examples to illustrate the two roles abductive inferences have in BOID*. We first discuss an example of deontic reasoning, adapted from Horty [11, p. 562].[1] Suppose that an agent has promised to meet a friend, but on its way to the meeting place, the agent encounters a drowning child. The agent is facing two conflicting imperatives, namely, to meet a friend for lunch, given the promise to do so, and to help a drowning child, given the fact that the child is drowning. In the original BOID, we would use two default rules to formalize the two obligations:

$$\mathit{promise} \overset{\mathrm{O}}{\hookrightarrow} \mathit{meet} \text{ and } \mathit{drowning\ child} \overset{\mathrm{O}}{\hookrightarrow} \mathit{help}.$$

We also assume that the two actions cannot both be carried out or that $\neg(\mathit{meet} \wedge \mathit{help})$ holds.[2] We would expect that the BOID takes only the second rule to generate the outcome goals, since this rule is intuitively more important. This can be done in BOID by assigning a higher priority to the second rule. Although the outcome meets our expectations, BOID, as well as most systems for deontic reasoning, leave implicit the reasoning behind this rule prioritization.

This is one of the examples where we use abductive rules to make explicit the reasoning behind the agent's priority function that leads to goal adoption. In moral reasoning about the conflicting obligations to help or not to help the child, prioritization is not explained by the facts that an agent made a promise or that an agent encountered a drowning child. We would expect that a truly autonomous agent relies on its own judgement of the values before choosing the right behavior. More specifically, the agent should consider the anticipated value of saving human life as overriding the anticipated value of not breaking a given promise as competing purposes that pertain to the example. To model this type of 'top-down' obligations, an autonomous agent considers the following two rules

$$\mathit{meet} \overset{\mathrm{O}}{\hookleftarrow} \neg \mathit{break\ promises} \text{ and } \mathit{help} \overset{\mathrm{O}}{\hookleftarrow} \mathit{save\ life},$$

saying that "in order not to break promises, you ought to meet a friend for lunch" and "in order to save a human life, you ought to help the drowning child", respectively. These two rules are examples of the teleological component in goal generation that has been absent from most practical reasoning formalizations. The new component increases the explanatory power of the original BOID algorithm.

The next example, borrowed from Thomason [20, p. 707], shows how abduction helps in detecting whether an agent's desire may be pursued as a realistic goal. Agents often find that their desires are not supported by their beliefs or even that their desires conflict their beliefs. Agents should exclude any unrealistic desires from their practical arguments and abductive rules might help in tracking such desires. Imagine that an agent wants to have decaf coffee and that it believes that it can only have decaf coffee if it is available. On Thomason's original formalization [20, p. 707], the rule "I can only have decaf coffee if decaf coffee is available" is a default $\mathit{have\ decaf} \overset{\mathrm{B}}{\hookrightarrow} \mathit{decaf\ available}$. This rule can be applied after the rule $\top \overset{\mathrm{D}}{\hookrightarrow} \mathit{have\ decaf}$ has been applied and, thereby, lead to the BOID extension $\{\mathit{have\ decaf}, \mathit{decaf\ available}\}$. This type of fallacious reasoning is known as 'wishful thinking', since the agent has no prior beliefs about the availability of coffee.[3]

The problem with the original formalization is that each rule is formalized as a default, which blurs the distinction between the states that trigger an action and anticipated states that are conditioned on the success of the action. On our view, the informal description gives the following two abductive rules

$$\mathit{have\ decaf} \overset{\mathrm{D}}{\hookleftarrow} \top \text{ and } \mathit{decaf\ available} \overset{\mathrm{B}}{\hookleftarrow} \mathit{have\ decaf}.$$

Based on these two rules, we would intuitively reject the plausibility of the conclusion that decaf coffee is available. After all, the two rules only say that the agent wants decaf coffee and that, in order to fulfill the desire, it needs to believe that decaf is available. To think that it, therefore, is available would mean that the agent resorted to the wishful thinking type of reasoning. What is needed is that the agent's reasoning is supported either by facts or beliefs that would 'connect' the desire to have decaf and a condition for realizing the desire, namely, the availability of decaf.

Thomason's example illustrates the role that the interaction of abductive rules and default rules has in planning and executing goals. Were it the case that the agent had appropriate beliefs about the availability of coffee, such as the rules $\top \overset{\mathrm{B}}{\hookrightarrow} \mathit{coffee\ available}$ and $\mathit{coffe\ available} \overset{\mathrm{B}}{\hookrightarrow} \mathit{decaf\ available}$, the agent would be able to meet the desire to have decaf [20, p. 704]. The rules of the type

[1] Horty's example is itself a variation of the famous example from [18, p. 231].

[2] For the sake of presentation, we henceforth assume that such 'material' inconsistencies are encoded into the background language L, and we leave the formulas of the type $\neg(\mathit{meet} \wedge \mathit{help})$ out of the formalized examples.

[3] The 'converse' of the wishful thinking problem is 'the side effect problem'. That is, motivational attitudes such as desires are not closed under beliefs: if we desire p, we do not necessarily desire the things we believe are consequences of p. Both problems will be solved by resorting to abductive reasoning.

decaf available $\overset{B}{\hookleftarrow}$ *have decaf* will be used to specify goal feasibility in the revised BOID algorithm and to integrate the goal generation and planning components of the BOID. In natural language, such rules are often phrased as "only if" conditions as, for example, in "I can have decaf only if decaf coffee is available".

The distinction between abductive rules and default rules in BOID* roughly corresponds to the distinction between the 'belief' direction of fit and the 'desire' direction of fit. The idea behind the 'direction of fit' concept is derived from Anscombe's examples [1, p. 56]. According to it, propositions can have two different roles corresponding to two different directions of fit: either *world-to-proposition*, or *proposition-to-world*. When reasoning about motivational content, the direction of fit is world-to-proposition: the reasoning is about how to make the world fit the propositions representing what an agent desires, intends or is obliged to do. But when reasoning about informational content (beliefs), the direction of fit is proposition-to-world: the reasoning is about how to make our propositional beliefs fit the world they aim to describe. This distinction is absent in the original BOID calculation scheme. In this paper, it will be used in a generalized sense that takes the notion of agents and agency as a primitive notion in relation to the world. This means that we will generalize the distinction to that between the 'agent-to-world' and 'world-to-agent' direction, where the primitive notion of 'proposition' is replaced with that of 'agent' and 'agency'.[4]

As the examples above show, we syntactically build this distinction into the rule direction for abductive reasoning and default reasoning. In general, the default rule direction ('$\hookrightarrow$') specifies an action in the consequent of a default rule such that the action is conditioned on the antecedent of the rule. The abductive rule direction ('$\hookleftarrow$') specifies an action in the consequent of an abductive rule such that the action conditions the antecedent of the rule. Figure 1 illustrates the two types of rules as two different approaches to interpreting why an agent takes some action α. (The parenthetic

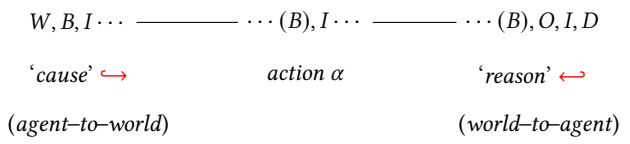

Figure 1: Default and abductive rule directions

'(B)' occurring above '*action* α' indicates that there is a belief aspect to intentions, while its occurrence above '*reason*' indicates the role of beliefs in anticipating necessary conditions to follow up on motivational attitudes.)

From the perspective of α, we can talk about its past triggers and its future purposes. When α is a subject of a mental attitude that is triggered by a past state represented with the formula *cause*, we use default reasoning and the formula *cause* is a part of a knowledge base. This type of action triggering is close to causation by environment or, at least, fitting an action to the agent's environment. It, therefore, modeled as having an 'agent-to-world' direction. Otherwise, when believing, intending, wanting or being obliged to do α is triggered by an anticipated state, we use abductive reasoning. Each chain of rules with an anticipated outcome ends with a 'projective', which is a consequent of an abductive rule with a tautological antecedent. For example, in Figure 1, the formula *reason* could be a consequent of one such rule *reason* $\hookleftarrow \top$.

For an illustration, we will use an example of elaborating a plan that an agent worked through by means of a series of intention rules. Intentions will be of special interest in this paper because they deal with both directions of fit. Assume that an agent formed an intention to go to the supermarket, and an intention to buy pasta for tonight's dinner, if it goes to the supermarket. Assume also that the agent formed an intention to have dinner if it buys pasta for the dinner and that all the intentions are based on the desire to have dinner. The following is the default rule direction formalization of the plan:

$$\top \overset{I}{\hookrightarrow} \textit{supermarket}, \textit{supermarket} \overset{I}{\hookrightarrow} \textit{buy pasta},$$
$$\textit{buy pasta} \overset{I}{\hookrightarrow} \textit{dinner} \text{ and } \top \overset{D}{\hookrightarrow} \textit{dinner}.$$

The original BOID has *observed* propositions as an input and *goal* propositions as an output. That means that along the way, in the deliberation, there is a transition of the direction of fit of the propositions involved: the incoming propositions have to fit the world while for the outgoing propositions the agent is going to act in a way that fits the world to the propositions. The desire to have dinner is motivating the agent to form the intention to go to the supermarket, and this intention is already a result of 'transitioning' to the 'agent-to-world' direction.

The change in the direction of fit is what will be addressed by using abductive reasoning. One way of interpreting this is to say that the realm of intentions cannot be entirely derived from the world of observed facts. In the example above, this implies that the rule $\top \overset{D}{\hookrightarrow} \textit{dinner}$ formally should not be of the same kind as the rule $\top \overset{I}{\hookrightarrow} \textit{supermarket}$. The intention to go to the supermarket is relevant for the agent *because* of the agent's desire to have dinner.[5]

In the BOID* setting, we are able to make explicit the telic dependence between the rules in the following way:

$$\textit{supermarket} \overset{I}{\hookleftarrow} \textit{buy pasta},$$
$$\textit{buy pasta} \overset{D}{\hookleftarrow} \textit{dinner} \text{ and } \textit{dinner} \overset{D}{\hookleftarrow} \top.$$

Notice that this does not mean that there is something wrong with the BOID default intention rules such as *supermarket* $\overset{I}{\hookrightarrow}$ *buy pasta*. After all, intention rules may be said to have both directions of fit. The added value of abductive rules is that they can be used to explain how the transition of the direction of fit came about as a result of deliberation, which is then encoded with the intention ('$\overset{I}{\hookrightarrow}$') rules. The '$\hookrightarrow$' direction rules can be said to implement the plan to have pasta for dinner that the agent had already adopted. Moreover,

[4] We will see in Section 3.3 that this generalization is important because beliefs are not the only type of mental attitudes that are patterned after the world. More importantly for modeling agency, intentions have to be patterned after the world, if they are to be executed in an agent's environment.

[5] In Anscombe's seminal work on intention, we find the distinction between *intention to* act and *intention with which* one acts [1, p. 17]. In the current formalism, this distinction can be mapped onto the directionality of rules. That is, prior intentions are intentions to do something, whereas intentions with which one acts are reasons formalized as consequents of abductive rules with a tautological antecedent.

abductive rules rectify the direction of fit for motivational attitudes such as obligation and desire rules, as illustrated by *dinner* $\overset{D}{\hookleftarrow} \top$.

In the next section, abductive reasoning will be the key element of the added deliberation component. Abduction addresses two important issues of autonomus reasoning about goals, namely, that of assessing the viability of previously adopted goals and that of adopting new goals that are not exclusively based on an agent's immediate observations. Our method to implement the deliberation phase is based on adding and removing intention rules. This is where we need to get precise on the procedure of transitioning from the 'world-to-agent' to the 'agent-to-world' direction of fit.

3 BOID* ARCHITECTURE

This section details the BOID architecture with the added mechanism of abductive reasoning for autonomous goal deliberation. The original BOID, at its core, is a computational scheme that deals with conflicts between beliefs, obligations, intentions and desires. The original BOID can already deal with very intricate ways in which conflicts between different mental attitudes arise, and it offers a class of simple strategies to solve them. The new BOID* starts from the background logical theory of propositional logic, extended with default rules that provide defeasible information about beliefs and intentions. In addition to default rules, BOID* includes abductive inferential information about motivational mental attitudes.

Definition 3.1 (BOID* Theory). *For a propositional language L and $p, q \in L$, a BOID* theory is defined as a tuple $\Delta = \langle W, B, O, I, D, \rho \rangle$ with $W \subset L$ as a finite set of observations, B and I as sets of belief and (prior) intention of the form $p \hookrightarrow q$ and $p \hookleftarrow q$, O and D as sets of obligation and desire rules of the form $p \hookleftarrow q$ and ρ a function from $B \cup O \cup I \cup D$ to the integers representing an agent's type.*

The function ρ can be used to distinguish between, for instance, selfish agents that prioritize the application of desire rules over obligation rules, and social agents, that do this the other way around. The function ρ gives ample opportunity to define all kinds of agent types [5, pp. 437-440], also ones that are, for instance, more specific in their selection of obligations rules such that certain subsets of obligation rules weigh stronger than other subsets. In the rest of the paper, we will assume the agent type ρ: $B > \{O, I\} > D$, where desires have the lowest priority and beliefs the highest.

BOID* theories are interpreted by two connected components, namely goal generation and goal deliberation. The two components are based on calculating two different types of BOID* extensions. Output extensions of the goal generation component inform extensions calculated in the goal deliberation component. Goal generation builds on the default reasoning direction of BOID* mental attitudes. This BOID* extension is calculated to answer which candidate goals are the ones that an agent is committing to, albeit only provisionally. The goal deliberation component includes goal selection and planning. Deliberation produces extensions by the application of abductive rules. Finally, the deliberation component extension is used to define intention reconsideration procedures.

3.1 Goal Generation

In the original BOID, calculating extensions is simply a process of applying default rules to a set of observations W. The BOID* calculation scheme inherits this procedure, but the extensions calculation scheme initially takes only (prior) intention and belief rules to generate goals. The intuition is that, at this stage, an agent relies on the goals that had been derived from motivational attitudes before the goal generation step started.

To be able to consider chosen goals and their effects, extensions will be defined as deductively closed sets of formulas. Deductive closure is defined using the function Th_L that takes a set of propositional formulas to give its logical consequences. For a set of propositional formulas $S \subseteq L$, we say that the set $Th_L(S)$ is its deductive closure". Default rule applicability is defined as follows.

Definition 3.2 (Default Rule Applicability). *We say that a default rule $p \hookrightarrow q$ is* applicable *to a deductively closed subset $D \subseteq L$, iff $p \in D$ and $\neg q \notin D$.*

The goal generation component can now be calculated as a default extension or multiple default extensions of a BOID* theory. Producing default extensions always starts with observations from W. The set of observations is extended by the available defaults from the belief set B and the prior intention set I, that is, those intentions that resulted from previous deliberations.

Definition 3.3 (BOID* Goal Generation Extensions). *Let $\Delta = \langle W, B, O, I, D, \rho \rangle$ be a BOID* theory. Define*

$S_0 = \{W\}$, *and for* $i \geq 0$

$S_{i+1} = \{\ Th_L(E^Y \cup \{q\}) \mid E^Y \in S_i,$
$\exists (p \hookrightarrow q)$ *such that*
$(p \hookrightarrow q) \in B \cup I$ *and*
$(p \hookrightarrow q)$ *is applicable to* E^Y *and*
$\nexists (v \hookrightarrow w) \in B \cup I$ *applicable to* E^Y
such that $\rho(v \hookrightarrow w) < \rho(p \hookrightarrow q)$,
or else $q = \top$ $\}$.

Then $E^Y \subseteq L$ is a goal generation extension *for Δ iff $S = \cup_{i=0}^{\infty} S_i$ and E^Y is a maximal element of S.*

Notice that the motivational attitudes from O and D do not yet contribute to the current extension calculation. Nevertheless, motivational attitudes do indirectly influence goal generation. Namely, those obligations and desires that had earlier met some fulfillment criteria are now realized through prior intentions in the view of previous deliberation.

Building on goal generation, the goal deliberation component will further specify how to compute which of the possibly conflicting motivations encoded into '$\hookleftarrow$' rules prevail as those that an agent commits to. In evaluating the viability of motivational attitudes, we consider the sets of beliefs:

Definition 3.4 (BOID* Belief Extension). *Let $\Delta = \langle W, B, O, I, D, \rho \rangle$ be a BOID* theory. Define*

$S_0 = \{W\}$, *and for* $i \geq 0$

$S_{i+1} = \{\ Th_L(E^\beta \cup \{q\}) \mid E^\beta \in S_i,$
$\exists (p \hookrightarrow q)$ *such that*
$(p \hookrightarrow q) \in B$ *and*
$(p \hookrightarrow q)$ *is applicable to* E^β *and*
$\nexists (v \hookrightarrow w) \in B$ *applicable to* E^β
such that $\rho(v \hookrightarrow w) < \rho(p \hookrightarrow q)$,
or else $q = \top$ $\}$.

Then $E^\beta \subseteq L$ is a belief extension *for Δ iff $S = \cup_{i=0}^{\infty} S_i$ and E^β is a maximal element of S.*

Formally, the belief set is defined analogously to the extension sets for the standard Reiter's default logic.

We will be interested in the sets of *strict beliefs*:

$$\Gamma = \bigcap \{Th_L(E^\beta) \mid E^\beta \text{ is a belief extension}\}.$$

The set contains beliefs included in each candidate belief extension, that is, those beliefs that are 'skeptically entailed' by Δ on the standard default logic consequence relation [2, p. 166]. The next section describes the deliberation phase in which an agent makes plans and adopts goals, within the bounds of what the agent's strict beliefs in Γ preclude as impossible.[6]

3.2 Goal Deliberation

The key component in the new BOID* architecture is the goal deliberation process guided by abductive reasoning. In this paper, deliberation is understood as a procedure that provides criteria for intention reconsideration. Informally, the procedure defines how to use BOID* to abductively reason from a projective or a purpose, as a premise, to an intended action, as a conclusion.

In Section 2, we already suggested that the application of abductive rules in BOID* ultimately does not start with an observation. Instead, abductive reasoning is triggered by an anticipated state of events, not the environment. We called such outcomes 'projectives' and projectives are, technically, abductive rules with a tautological antecedent. For example, the antecedent *save life* of the rule *help* $\overset{O}{\hookleftarrow}$ *save life* might be one of such projectives, assuming that an agent follows the value of saving lives without needing any further explanation for such decision. In that case, the agent would have the rule *save life* $\overset{O}{\hookleftarrow} \top$ in the set of obligation rules O. The desire rule *have decaf* $\overset{D}{\hookleftarrow} \top$ from Thomason's example is another projective.

When an agent reasons abductively from projectives to reach new intentions and plans, it engages in a distinctive use of rules in which the agent chooses to do an action because it leads to a foreseeable outcome of that action. To use the example from Section 2, an agent's intention *supermarket* $\overset{I}{\hookleftarrow}$ *buy pasta* is a result of the agent's expectation that going to the supermarket will lead, other things being equal, to a side effect of buying pasta. However, in reasoning to the conclusion that the agent intends to do the action *supermarket*, the underlying regularity that it is the supermarket where the agent normally buys pasta is applied in reverse.

In reasoning abductively, agents often deal with multiple necessary conditions that need to be satisfied or multiple actions that need to be undertaken to bring about some projective that motivated their reasoning. For example, the agent first reasoned that in order to fulfill the desire to have dinner, it needs to undertake the action of buying pasta. Then it reasoned to the conclusion that in order to buy pasta, it will go to the supermarket, which is typical for the means-end type of reasoning. The agent might also think about necessary conditions such as that the supermarket needs to be open in order to go to the supermarket. The pattern that can be obtained from abductive reasoning applied to the 'projectives' is that each inference to a new action or condition can be seen as a means to obtain the outcome of a previous abductive inference step. Abductive reasoning thus creates a characteristic pattern of 'chains', that is, possible paths to reach a projective. This intuition is captured in the following definition:

Definition 3.5 (Chain of Abductive Rules). *For a BOID* theory $\Delta = \langle W, B, O, I, D, \rho\rangle$, a* chain of abductive rules *is a sequence of rules $(p_n \overset{X}{\hookleftarrow} p_{n-1}), \ldots, (p_1 \overset{X}{\hookleftarrow} p_0)$, where $X \in \{B, O, I, D\}$ and $p_0 = \top$, such that*

- *for $n > i > 0$, the antecedent p_i of the rule $p_{i+1} \overset{X}{\hookleftarrow} p_i$ is the consequent of the rule $p_i \overset{X}{\hookleftarrow} p_{i-1}$,*
- *for $p_1 \overset{X}{\hookleftarrow} p_0, X \in \{O, I, D\}$ and*
- *for $n \geq k > 0$ such that $p_{k+1} \overset{B}{\hookleftarrow} p_k$, it holds that $p_{m+1} \overset{B}{\hookleftarrow} p_m$, for each $m, n \geq m > k$.*

We refer to the first and last rule in a chain of abductive rules as its 'top' and 'leaf' rule, respectively. Each chain of abductive rules has a motivational attitude as its top rule, and it might potentially have one or more belief rules at its leaf rule end.

Out of all the possible abductive chains, we are interested in those chains that do not omit any available information encoded in abductive rules, that are grounded in conditions that do not contradict strict beliefs and that respect the priority constraints given in ρ. That is, given a chain of abductive rules $(p_n \overset{X}{\hookleftarrow} p_{n-1}), \ldots, (p_1 \overset{X}{\hookleftarrow} p_0)$ of Δ, we say that it is *maximal*, *grounded* and *prioritized* if

- $\nexists X(q \overset{X}{\hookleftarrow} p_n)$ such that $X \in \{B \cup O \cup I \cup D\}$ **(maximality)**
- $\nexists Y(p_{k+1} \overset{Y}{\hookleftarrow} p_k)$, for $n > i > 0$, such that $Y = B$ and $\neg p_{k+i} \in \Gamma$ **(groundedness)** and
- $\nexists Z(q \overset{Z}{\hookleftarrow} p_i)$, for $n > i \geq 0$, such that $Z \in \{O \cup I \cup D\}$ and $\rho(q \overset{Z}{\hookleftarrow} p_i) > \rho(p_{i+1} \overset{X}{\hookleftarrow} p_i)$ **(priority)**.

Notice that there may be several chains of abductive rules connected to a single projective. This corresponds to the idea that an agent anticipates several possible paths that bring about the same outcome.

The possibility of following multiple available paths also means that there could be multiple goal deliberation extensions that are not necessarily inconsistent.[7] The definition of goal deliberation extensions for BOID* is based on the definition of chains of abductive rules:

Definition 3.6 (BOID* Goal Deliberation Extension). *For a BOID* theory $\Delta = \langle W, B, O, I, D, \rho\rangle$, a set of formulas*

$$E^\delta = Th_L(\{p_0, \ldots, p_n\})$$

is a goal deliberation extension *iff $(p_n \overset{X}{\hookleftarrow} p_{n-1}), \ldots, (p_1 \overset{X}{\hookleftarrow} p_0)$ is a maximal and grounded chain of abductive rules of Δ such that no other maximal and grounded chain of Δ is prioritized over it.*

The goal deliberation extension does not respect the flow of information from the original BOID. The original BOID extensions are built according to the flow of information from *observations* to goal sets, whereas BOID* includes an additional component with the flow of information from *anticipations* to goal sets (Figure 2).

[6] However, in Section 3.2, deliberation will not be restricted by prior intentions. We could say that BOID* thus implicitly sides with philosophers such as Broome [8] who claim that intentions on their own do not count as reasons in rethinking goals.

[7] For example, one can bring decaf in a vacuum flask or order decaf to have decaf, but doing both 'overdetermines' the desired outcome.

3.3 Intention Reconsideration

The effects of the deliberation process are modeled through altering agent's intentions. The procedure of reconsidering intention will be informed by goal deliberation in such a way that those sequences of abductive rules that generate deliberate extensions become a criteria for adopting new or discarding old intention rules. This section outlines the core process of *intention reconsideration* in the new BOID* architecture. Intention reconsideration amounts to revising the set of intention rules I or their priorities for a BOID* theory Δ. Intentions are reconsidered in the following ways: an agent may add a new intention (not) to do an action or it can discard or disregard an intention that the deliberation process revealed to be infeasible or overridden.

Both adding and removing intention rules will depend on what formulas does a chain of abductive rules specify as conditions to bring about some state of affairs. The most straightforward way to add a new intention rule results from those extension generating chains of abductive rules whose leaf rule is a motivational attitude. For example, the leaf rule $supermarket \overset{\text{I}}{\hookleftarrow} buy\ pasta$ results in a new intention rule $\top \overset{\text{I}}{\hookrightarrow} supermarket$. This type of intention rules is close to what Bratman calls 'simple intentions' [3].

In contrast to the simple intention rules, an agent may adopt conditional intentions, contingent on whether a certain condition holds. The rule $supermarket \overset{\text{I}}{\hookrightarrow} buy\ pasta$ is one such intention rule saying that, if an agent goes to the supermarket, it intends to buy pasta. Another conditional intention is $decaf\ available \overset{\text{I}}{\hookrightarrow} have\ decaf$, which would follow from Thomason's scenario, were it the case that the enabling condition *decaf available* is satisfied.[8]

Both simple and conditional intentions discussed above might also be 'negative' in the sense that an agent forms an intention not to do something, that is, an intention to refrain itself from doing an action [10, p. 120]. We will explicitly define a procedure for those negative conditional intentions for which it holds that an agent can ascertain that a necessary precondition for an action cannot be fulfilled. For example, if the rule $decaf\ available \overset{\text{B}}{\hookleftarrow} have\ decaf$ is the leaf rule of each chain of abductive rules starting from the rule $have\ decaf \overset{\text{D}}{\hookleftarrow} \top$, and if an agent believes that *decaf available* is not true, then it adds the following intention rule: $\neg decaf\ available \overset{\text{I}}{\hookrightarrow} \neg have\ decaf$ or "if the decaf is not available, then I intend not to have decaf".

The following definition formalizes the described intention expansion procedures.

Definition 3.7 (Intention Expansion). *For a prioritized grounded and maximal chain of abductive rules* $(p_n \overset{\text{X}}{\hookleftarrow} p_{n-1}), \ldots, (p_1 \overset{\text{X}}{\hookleftarrow} p_0)$*, a deliberate extension* E^δ*, and a set of strict beliefs* Γ *of a BOID* theory* $\Delta = \langle W, B, O, I, D, \rho \rangle$*, the theory* $\Delta^+ = \langle W, B, O, I^+, D, \rho \rangle$ *is the* intention expansion *of* Δ*, where the set of intention rules* I^+ *is closed under the conditions* (a)*-*(d) *defined in such a way that if*

(a) **[simple intention]**

$\exists X(p_n \overset{\text{X}}{\hookleftarrow} p_{n-1})$ *such that* $X \in \{O, I, D\}$,

then $I^+ = I \cup \{\top \overset{\text{I}}{\hookrightarrow} p_n\}$;

(b) **[conditional (enabling) intention]**

1. $\exists X(p_i \overset{\text{X}}{\hookleftarrow} p_{i-1})$, *for* $n \geq i > 0$, *such that* $X \in \{O, I, D\}$,
2. $\forall Y(p_k \overset{\text{Y}}{\hookleftarrow} p_{k-1})$, *for* $n \geq k > i$, $Y = B$,

then $I^+ = I \cup \{p_n \wedge \ldots \wedge p_{i+1} \overset{\text{I}}{\hookrightarrow} p_i\}$;

(c) **[conditional intention]**

1. $\exists X(p_i \overset{\text{X}}{\hookleftarrow} p_{i-1})$, *for* $n \geq i > 1$, *such that* $X \in \{O, I, D\}$,
2. $\neg p_n, \neg p_{n-1}, \ldots, \neg p_{i+1} \notin \Gamma$,

then $I^+ = I \cup \{p_i \overset{\text{I}}{\hookrightarrow} p_{i-1}\}$;

(d) **[conditional negative intention]**

1. $\exists X(p_i \overset{\text{X}}{\hookleftarrow} p_{i-1})$, *for* $n \geq i > 0$, *such that* $X \in \{O, I, D\}$,
2. $\exists p_k$, *for* $n \geq k > i$, *such that* $\neg p_k \in \Gamma$,
3. $\forall E^\delta$ *such that* $p_i \in E^\delta, p_k \in E^\delta$,

then $I^+ = I \cup \{\neg p_k \overset{\text{I}}{\hookrightarrow} \neg p_i\}$;
otherwise, $I^+ = I$.

Definition 3.7 lists four ways in which an agent might form a new intention after deliberation. The first way, defined in (a), is to simply add an unconditioned intention because the agent reasoned directly to the conclusion that it intends to (or is obliged to to or wants to) do p_n. An agent might also consider intentions contingent on whether some condition is fulfilled or not, as defined in (b) and (c). If an agent does not believe that a necessary condition for an action does not hold, the agent adds an intention rule with the condition(s) as the rule's antecedent(s) and the action as its consequent. In (b), an agent adopts intentions that are only executed if an agent believes that some conditions hold, while, in (c), an agent adopts intentions conditioned on its other intentions. Finally, in (d), an agent might have a strict belief that some goals are not feasible and form an intention not to follow up on an obligation, intention or desire.

BOID* agents may also abandon their intentions when they turn out to be unfounded or no longer preferred. In short, an agent might either discard infeasible intentions that go against agent's strict beliefs or decrease the priority of overridden intentions that go against other, more important, goals.

Definition 3.8 (Intention Contraction). *For a maximal chain of abductive rules* $(p_n \overset{\text{X}}{\hookleftarrow} p_{n-1}), \ldots, (p_1 \overset{\text{X}}{\hookleftarrow} p_0)$*, a deliberate extension* E^δ*, and a set of strict beliefs* Γ *of a BOID* theory* $\Delta = \langle W, B, O, I, D, \rho \rangle$*, the theory* $\Delta^- = \langle W, B, O, I^-, D, \rho^- \rangle$ *is the* intention contraction *of* Δ*, where the set of intention rules* I^- *and the priority function* ρ^- *are defined in such a way that if*

(a) **[infeasible intention]**

1. $\exists X(p_i \overset{\text{X}}{\hookleftarrow} p_{i-1})$, *for* $n \geq i > 1$, *such that* $X \in \{B, O, I, D\}$,
2. $\exists p_k$, *for* $n \geq k \geq i$, *such that* $\neg p_k \in \Gamma$,

then $I^- = I \setminus \{p_i \overset{\text{I}}{\hookrightarrow} p_{i-1}, p_n \wedge \ldots \wedge p_i \overset{\text{I}}{\hookrightarrow} p_{i-1}, \top \overset{\text{I}}{\hookrightarrow} p_{i-1}\}$;

(b) **[overridden intention]**

1. $\exists X(p_i \overset{\text{X}}{\hookleftarrow} p_{i-1})$, *for* $n \geq i > 0$, *such that* $X \in \{O, I, D\}$,
2. $a \overset{\text{I}}{\hookrightarrow} p_i \in I$,
3. $\exists E^\delta$ *such that* $\neg p_i \in E^\delta$ & $\nexists E^\delta$ *such that* $p_i \in E^\delta$,

then $\rho^-(a \overset{\text{I}}{\hookrightarrow} p_i) = 0$;
otherwise, $\rho^- = \rho$.

According to (a), if an agent safely believes that a condition to follow through on an intention cannot be fulfilled, the agent gives

[8] According to [4, p. 218], intentions that are merely contingent on 'enabling' conditions do not count as 'genuine' conditional intentions. According to [13], they do.

up on unrealistically pursuing the intention. According to (b), if an agent formed an intention to $\neg p$ via an extension-generating chain of abductive rules and some prior intention to p is not a consequent in any such chain of abductive rules, then the intention to p is overridden.

Intention contraction and intention expansion will be used to define an intention revision operation for BOID* theories. The revised set of intentions and the revised priority function are considered as an output of the goal deliberation stage, which is further used in the goal initiation stage. The definition of intention revision is as follows.

Definition 3.9 (Intention Revision). *For a BOID* theory* $\Delta = \langle W, B, O, I, D, \rho\rangle$*, the theory* $\Delta^* = \langle W, B, O, I^*, D, \rho^*\rangle$ *is the* intention revision *of* Δ*, where the set of intention rules* I^* *and the priority function* ρ^* *are defined as follows:*

$$I^* = I^- \cup \{I^+ \setminus I\} \qquad \text{and} \qquad \rho^* = \rho^-.$$

As discussed in Section 3.2, BOID* has a two-fold flow of information. In addition to the original flow of information from observations to goal sets, BOID* integrates information flow starting from anticipations. Figure 2 illustrates the new flow of information as a result of integrating the deliberation component into BOID*.

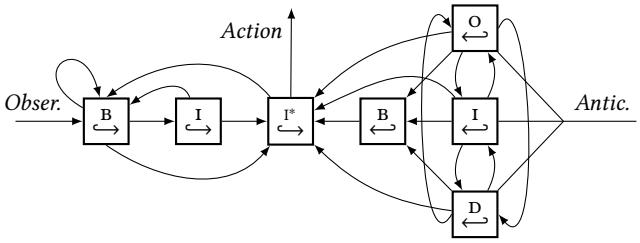

Figure 2: Information flow in BOID* with integrated goal generation-and-deliberation components

A given output of intention reconsideration might seemingly break with the agent type, if we only focus on the default rule direction of a theory. This could, for example, result from a prioritization happening closer to the top rule in a chain of abductive rules.

We can now formalize the dinner plan example from Section 2 to show the workings behind intention formation.

Example 3.1 (Dinner Plan). *We define a BOID* theory* $\Delta_0 = \langle W_0, B_0, O_0, I_0, D_0, \rho_0\rangle$*, where* I_0 *contains the intention rule*

$$\textit{supermarket} \overset{\text{I}}{\hookleftarrow} \textit{buy pasta},$$

and D_0 *contains the desire rules*

$$\textit{buy pasta} \overset{\text{D}}{\hookleftarrow} \textit{dinner} \text{ and } \textit{dinner} \overset{\text{D}}{\hookleftarrow} \top.$$

At first, the goal generation extension of Δ_0 does not specify any prior choices made by the agent. The agent then adopts having dinner as a goal. This corresponds to the only chain of abductive rules

$$(\textit{supermarket} \overset{\text{I}}{\hookleftarrow} \textit{buy pasta}), (\textit{buy pasta} \overset{\text{D}}{\hookleftarrow} \textit{dinner}), (\textit{dinner} \overset{\text{D}}{\hookleftarrow} \top),$$

with $\textit{dinner} \overset{\text{D}}{\hookleftarrow} \top$ as its top rule.

Then, by the condition (a) of Definition 3.7, the agent forms the intention to go to the supermarket and adds the simple intention rule $\top \overset{\text{I}}{\hookrightarrow} \textit{supermarket}$. By Definition 3.7 (c), the agent adds two conditional intention rules, namely $\textit{supermarket} \overset{\text{I}}{\hookrightarrow} \textit{buy pasta}$ and $\textit{buy pasta} \overset{\text{I}}{\hookrightarrow} \textit{dinner}$. Starting from the revised intention rules set I_0^*, we can calculate a new goal generation extension defined as $E^Y = Th_L(\{\textit{supermarket}, \textit{buy pasta}, \textit{dinner}\})$, as expected according to the informal discussion of the example from Section 2. The example shows why default intention rules are crucial in changing the direction of fit from world-to-agent to agent-to-world.

3.4 Horty's and Thomason's Examples in BOID*

In Section 2, we introduced two examples of reasoning with abductive rules that we now want to formalize with the new BOID* extension calculation schema. In Horty's example, we dealt with an agent who encounters a drowning child, while being bound to keep a promise to meet a friend for lunch. This example shows the importance of abductive reasoning for goal selection because the decision on whether to help a child or meet a friend will depend on whether the agent prioritizes $\textit{save life} \overset{\text{O}}{\hookleftarrow} \top$ over $\neg\textit{break promise} \overset{\text{O}}{\hookleftarrow} \top$.

Example 3.2 ([11] (Adapted)). *We define a BOID* theory* $\Delta_1 = \langle W_1, B_1, O_1, I_1, D_1, \rho_1\rangle$*, where* $W_1 = \{\textit{promise}, \textit{drowning child}\}$*,* B_1 *consists of the belief rules*

$$\textit{promise} \overset{\text{B}}{\hookleftarrow} \textit{meet} \text{ and } \textit{drowning child} \overset{\text{B}}{\hookleftarrow} \textit{help},$$

O_1 *consists of the following obligation rules*

$$\textit{meet} \overset{\text{O}}{\hookleftarrow} \neg\textit{break promises}, \textit{help} \overset{\text{O}}{\hookleftarrow} \textit{save life},$$

$$\neg\textit{break promises} \overset{\text{O}}{\hookleftarrow} \top \text{ and } \textit{save life} \overset{\text{O}}{\hookleftarrow} \top,$$

and I_1 *contains the intention rule*

$$\textit{promise} \overset{\text{I}}{\hookrightarrow} \textit{meet}.$$

The agent type ρ_1 *is* $B \succ \{O, I\} \succ D$*, with an additional priority specification:* $\rho_1(\textit{save life} \overset{\text{O}}{\hookleftarrow} \top) > \rho_1(\neg\textit{break promises} \overset{\text{O}}{\hookleftarrow} \top)$.

The BOID* formalization of the scenario makes explicit some of the reasoning steps that are missing in the original deontic logic formalization. First, the agent had previously selected a goal to meet a friend and formed the (prior) intention rule $\textit{promise} \overset{\text{I}}{\hookrightarrow} \textit{meet}$, which is made clear by the assumption that the agent was on its way to meet a friend. However, the agent then encounters a drowning child, as the set of observations W_1 shows. At this stage, the goal generation extension includes the formula *meet*.

It is only then that the reasoning about the moral conflict starts with the process of deliberation. There are two possible chains of abductive rules, namely

$$(\textit{promise} \overset{\text{B}}{\hookleftarrow} \textit{meet}), (\textit{meet} \overset{\text{O}}{\hookleftarrow} \neg\textit{break promises}), (\neg\textit{break promises} \overset{\text{O}}{\hookleftarrow} \top)$$

$$\text{and } (\textit{drowning child} \overset{\text{B}}{\hookleftarrow} \textit{help}), (\textit{help} \overset{\text{O}}{\hookleftarrow} \textit{save life}), (\textit{save life} \overset{\text{O}}{\hookleftarrow} \top).$$

Both chains are grounded and maximal, but only the second one is prioritized, according to ρ_1 that specifies preference for saving life over not breaking a promise. Thus, there is only one goal deliberation extension, namely $E^\delta = Th_L(\textit{drowning child}, \textit{help}, \textit{save life})$.

As a result of Definition 3.7, the intention revision set I_1^* will include a new conditional intention *drowning child* $\overset{\text{I}}{\hookrightarrow}$ *help*, according to (b). Furthermore, since *meet* and *help* are materially inconsistent, E^δ contains the formula $\neg$*meet*. According to Definition 3.8 (b), this means that the prior intention *promise* $\overset{\text{I}}{\hookrightarrow}$ *meet* has been overridden and, therefore, ρ_1^* assigns value 0 to it.

There are several ways in which the BOID* formalization provides insight into how intelligent agents resolve moral conflicts such as the one we modeled in Δ_1. Reasoning about moral conflicts happens in the goal deliberation component of the BOID*, while the goal generation component concerns output intentions. This conforms to the intuition that the questions of what one *ought to* do are deliberative in character. Perhaps more importantly, assigning priorities to the abductive rules with projectives, that is, to the rules *save life* $\overset{\text{O}}{\hookleftarrow}$ $\top$ and $\neg$*break promises* $\overset{\text{O}}{\hookleftarrow}$ $\top$, provides a more appropriate explanation for the agent's output 'all things considered' obligation to help a child. Assigning a higher priority value to *save life* $\overset{\text{O}}{\hookleftarrow}$ $\top$ than to $\neg$*break promises* $\overset{\text{O}}{\hookleftarrow}$ $\top$ corresponds to what we expect from autonomous agents, namely, to choose those actions that are aligned with the values that they promote.

Thomason's example illustrated why abductive rules of the type *decaf available* $\overset{\text{B}}{\hookleftarrow}$ *have decaf* are of importance in deliberation. The rule specifies conditions under which the desire to have decaf might be realistically pursued. Abduction is thus also a method to integrate the planning component into the BOID architecture.

Example 3.3 ([20] (Adapted)). *We define a BOID* theory* $\Delta_2 = \langle W_2, B_2, O_2, I_2, D_2, \rho_2 \rangle$, *where* B_2 *contains the belief rule*

$$\textit{decaf available} \overset{\text{B}}{\hookleftarrow} \textit{have decaf},$$

and D_2 *contains the desire rule*

$$\textit{have decaf} \overset{\text{D}}{\hookleftarrow} \top.$$

The problem of 'wishful thinking' discussed in [20, p. 707] can be avoided in the BOID* formalization. Since W_2 is empty and B_2 does not contain any '$\hookrightarrow$' rules, the agent does not have any information on whether decaf is available or not. Thus, it cannot be the case that the agent derives, on the basis of the available rules in B_2 and D_2, that decaf is available. This is the advantage of keeping 'informational' and 'motivational' components apart.

In Δ_2, the goal generation extension is the closure $Th_L(W)$. In the deliberation component, the agent forms an intention rule *decaf available* $\overset{\text{I}}{\hookrightarrow}$ *have decaf*. This conditional intention is formed based on the chain of abductive rules

$$(\textit{decaf available} \overset{\text{B}}{\hookleftarrow} \textit{have decaf}), (\textit{have decaf} \overset{\text{D}}{\hookleftarrow} \top),$$

following Definition 3.7, part (b). The intention revision set I_2^* thus includes the rule *decaf available* $\overset{\text{I}}{\hookrightarrow}$ *have decaf*. But this rule would not be triggered by iterating goal generation with the revised intention set. The reason is that the agent does not have any information on the status of the condition *decaf available*.

If we assume, instead, that the agent has the relevant information about the availability of coffee, e.g., by adding the rules $\top \overset{\text{B}}{\hookrightarrow}$ *coffee available* and *coffe available* $\overset{\text{B}}{\hookrightarrow}$ *decaf available* to B_2 [20, p. 704], the intention to have decaf becomes a part of the goal generation extension computed with I_2^*. On the contrary, were it the case that, e.g., $\top \overset{\text{B}}{\hookrightarrow} \neg$*decaf available* is in B_2, the agent would strictly believe that there is no decaf, which means that Definition 3.7 would not result in the intention formation resulting in adding *decaf available* $\overset{\text{I}}{\hookrightarrow}$ *have decaf*. Moreover, according to (d), the agent would form a negative intention $\neg$*decaf available* $\overset{\text{I}}{\hookrightarrow}$ $\neg$*have decaf* as result of believing that it cannot meet necessary conditions for having decaf.

4 DISCUSSION

In this paper, we proposed a new BOID* system for resolving conflicts between mental attitudes, with intention reconsideration as the core method of BOID*. By making a distinction between the default and abductive direction of rules, BOID* enables for a richer model of informational (beliefs), motivational (obligations and desires) and deliberative (intentions) attitudes. The idea of combining abductive and default reasoning techniques has been previously explored for modeling informational attitudes only [15], but we argue here that the combination of the two methods shows its full potential when informational attitudes are juxtaposed with motivational attitudes.

By representing motivational attitudes with abductive rules, BOID* has a method to make explicit the reasoning behind assigning a higher priority to one motivational attitude over another, as shown in Horty's example. It is also intuitive to have such considerations in a separate, deliberation component. The new deliberation component opens up a possibility to explicitly model intention reconsideration as intention rule expansion and contraction. Additionally, separating BOID* deliberation extensions for motivational attitudes avoids the problems resulting from calculating 'mixed' goal generation extensions with both motivational and informational attitudes, such as 'side effect' and 'wishful thinking' problems. The latter problem has been recognized in the original BOID setting and dealt with in [7].

In our future work, we plan to replace priorities with meta-rules specifying that, for instance, obligations are more important than desires in some situations. The BOID* architecture implementations could take several useful directions. One of them is to build AI models based on BOID* that can detect conflicts between habit and the will, and hence serve as a method to outsource the will in case habits intrude. Another interesting implementation could integrate BOID* algorithms for anticipatory reasoning with statistical learning models to enhance their explanatory power [17].

Finally, we presented abduction as an appropriate reasoning method to incorporate one of the key features of autonomous goal generation, namely anticipation. Research in biological systems shows that autonomous living systems need to have a sufficient level of anticipatory behavior, not only reactive behavior patterns. By anticipating goals and abductively inferring how to obtain those goals, BOID* agents present a step forward in understanding the principles that are essential for fully autonomous artificial agency.

ACKNOWLEDGMENTS

Our work is supported by the Dutch Research Coucil (NWO) project *Empowering Human Intentions through Artificial Intelligence.*

REFERENCES

[1] Gertrude E. M. Anscombe. 1963. *Intention* (2nd ed.). Cornell University Press.
[2] Grigoris Antoniou. 1997. *Nonmonotonic Reasoning*. Cambridge, MA: MIT Press.
[3] Michael E. Bratman. 1979. Simple intention. *Philosophical Studies* 36, 3 (1979), 245–259.
[4] Michael E Bratman. 1999. Davidson's theory of intention. In *Faces of intention: Selected essays on intention and agency*. Cambridge University Press, 209–224.
[5] Jan Broersen, Mehdi Dastani, Joris Hulstijn, and Leendert van der Torre. 2002. Goal generation in the BOID architecture. *Cognitive Science Quarterly* 2, 3-4 (2002), 428–447.
[6] Jan Broersen, Mehdi Dastani, and Leendert van der Torre. 2001. Resolving conflicts between beliefs, obligations, intentions, and desires. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Salem Benferhat and Philippe Besnard (Eds.). Springer, 568–579.
[7] Jan Broersen, Mehdi Dastani, and Leendert van der Torre. 2002. Realistic desires. *Journal of Applied Non-Classical Logics* 12, 2 (2002), 287–308.
[8] John Broome. 2001. Are intentions reasons? And how should we cope with incommensurable values? In *Practical Rationality and Preference: Essays for David Gauthier*, Christopher W. Morris and Arthur Ripstein (Eds.). Cambridge University Press, 98–120.
[9] John Collier. 2008. Simulating autonomous anticipation: The importance of Dubois' conjecture. *BioSystems* 91, 2 (2008), 346–354.
[10] Gilbert Harman. 2000. Desired Desires. In *Explaining Value: And Other Essays in Moral Philosophy*. Oxford University Press, 117–136.
[11] John F. Horty. 2003. Reasoning with moral conflicts. *Noûs* 37, 4 (2003), 557–605.
[12] John F. Horty. 2012. *Reasons as Defaults*. Oxford University Press.
[13] Kirk Ludwig. 2015. What are conditional intentions? *Methode: Analytic Perspectives* 4 (2015), 30–60. Issue 6.
[14] Charles S. Peirce. 1974. *Collected Papers of Charles Sanders Peirce*. Vol. 1/2. Harvard University Press.
[15] David Poole. 1989. Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence* 5, 2 (1989), 97–110.
[16] Raymond Reiter. 1980. A logic for default reasoning. *Artificial Intelligence* 13, 1-2 (1980), 81–132.
[17] Farhad Shakerin and Gopal Gupta. 2020. Whitebox induction of default rules using high-utility itemset mining. In *Practical Aspects of Declarative Languages*, Ekaterina Komendantskaya and Yanhong Annie Liu (Eds.). Springer International Publishing, 168–176.
[18] Peter Singer. 1972. Famine, affluence, and morality. *Philosophy & Public Affairs* 1, 3 (1972), 229–243.
[19] Yao-Hua Tan. 1997. Is default logic a reinvention of inductive-statistical reasoning? *Synthese* 110, 3 (1997), 357–379.
[20] Richmond H. Thomason. 2000. Desires and defaults: A framework for planning with inferred goals. In *Seventh International Conference on Principles of Knowledge Representation and Reasoning, KR 2000*. 702–713.