How to make the BOID ethical (and other observations)

Jan Broersen

Department of Philosophy and Religious Studies
Utrecht University, The Netherlands

Deontic logic in 2020-2030
Luxembourg
Oktober 2nd, 2020

Outline

Forward looking responsibility

- We cannot leave it to the deep learning people to make artificial agents *ethical*: they do not have the tools
- But, do we deontic logicians have the tools? Do we even think about it in the right way?
- Our formalisms:
  (1) modal possible world approaches (semantic orientation) +
  (2) rule-based reasoning approaches (syntactic orientation) +
  (3) ...
- Is something missing? I think maybe there is (see end of the talk).

Outline

An architecture for artificial agency

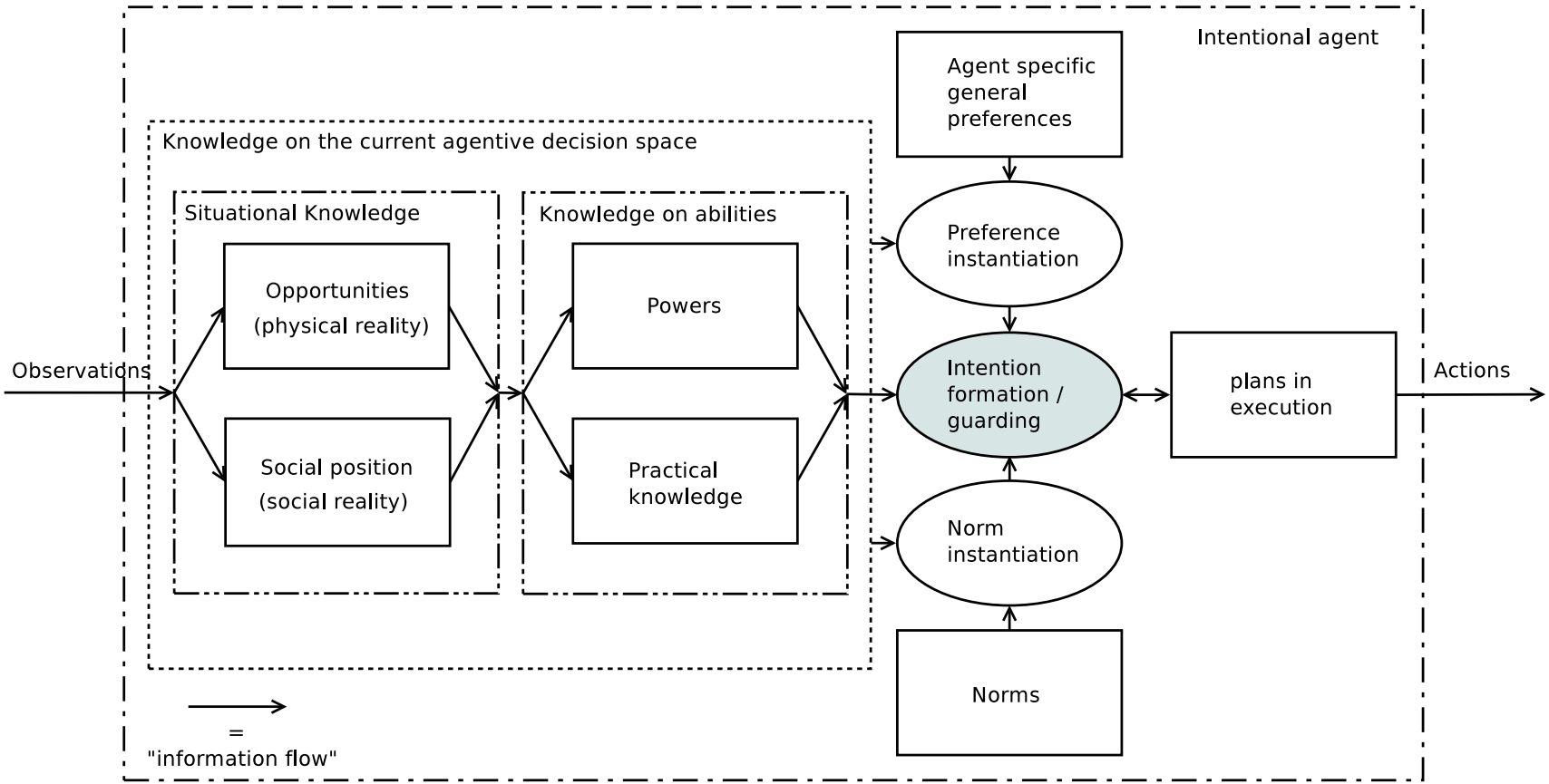


Figure: A pre-formal conceptual model of intentional agency

Outline

What the BOID [Broersen, Dastani, Hulstijn, vd Torre, 2000-2003] is about

The BOID is[1] based on the four mental attitudes Belief, Obligation, Intention and Desire and is:

- a *simple* (!) rule based system for practical reasoning
- an architecture where attitude components and the way they are connected constrain the reasoning
- a prioritised rule based semantics
- a *conflict resolution* mechanism

[1] If you ask the authors you will get 4 different answers

The (a) BOID architecture

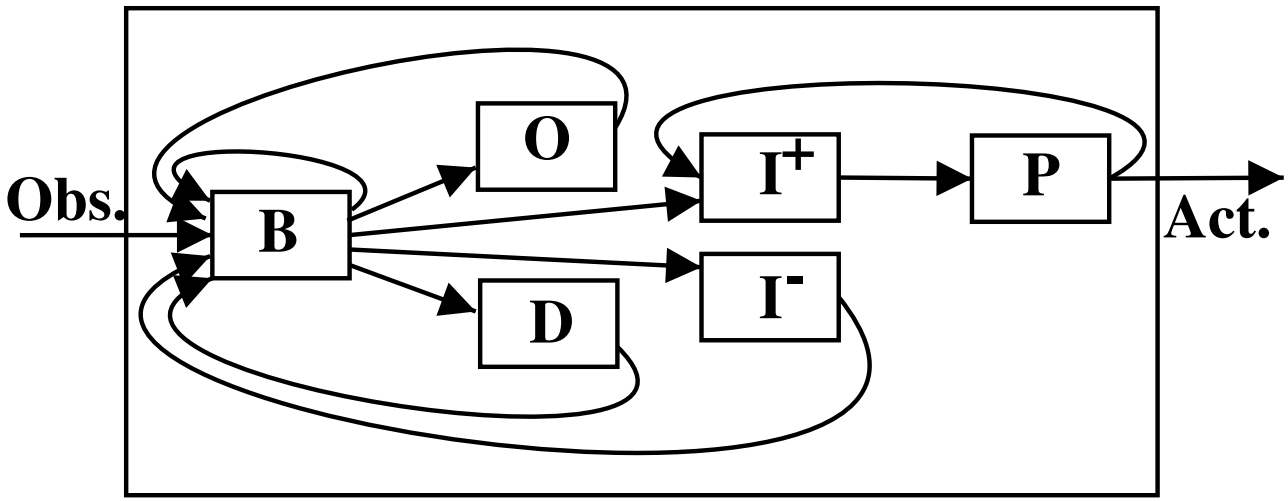


Figure: A BOID architecture and information flow

Assumptions made in the BOID

- Propositional content is for real (this is not an enactivist-type of approach, like Rodney Brooke's)
- Propositional content can somehow (e.g., by sub-symbolic means) be grounded in an agent's environment
- Propositional content is the source (cause?, reason?) for concrete action (through planning)
- Practical reasoning and conflict resolution can suitably be modeled by defeasible rules (Rich Thomason, *Desires and Defaults*; John Horty's book *Reasons as Defaults*)

What the BOID is *not* about

The following possible concepts for artificial agents are not modelled in the BOID:

- self-knowledge,
- consciousness and awareness,
- emotion, feeling,
- understanding, intuition,
- concept learning,
- learning skills, knowing how,
- psychological phenomena (e.g., akrasia, dissonance, bias, etc.),
- communication, language, interpretation.

BOID signatures

Definition (BOID theories, agent types and rule applicability)

Relative to a propositional language L, a BOID theory is a tuple $\Delta = \langle W, B, O, I^{-}, D, \rho \rangle$ with:

- W a subset of L representing observations,
- B, O, I^{-} and D sets of belief, obligation, (prior) intention and desire rules of the form $\alpha \hookrightarrow w$ with α and w elements of L,
- ρ a function from $B \cup O \cup I^{-} \cup D$ to the integers assigning priorities. ρ represents an agent's *type*.

We say that a rule $\alpha \hookrightarrow w$ is *applicable* to a deductively closed subset $E \subseteq L$, iff $\alpha \in E$ and $\neg w \notin E$.

The simplified algorithm

Definition (BOID Extension Calculation Scheme)

Let $\Delta = \langle W, B, O, I^{-}, D, \rho \rangle$ be a BOID theory.
Define
$E_0 = \{W\}$, and for $i \geq 0$
$E_{i+1} = Th_L\big(E_i \cup \{w \mid (\alpha \hookrightarrow w) \in B \cup O \cup I^{-} \cup D$ and
$(\alpha \hookrightarrow w)$ is applicable to E and
$\nexists (\beta \hookrightarrow v) \in B \cup O \cup I^{-} \cup D$ applicable to E
such that $\rho(\beta \hookrightarrow v) < \rho(\alpha \hookrightarrow w)\ \}\ \big)$
Then $E \subseteq L$ is an extension for Δ iff $E = \cup_{i=0}^{\infty} E_i$

This is the simplest version, that assumes ρ is a *total order*.

(the lower in the ordering, the more preferred)

Comments on the BOID semantics

The BOID applies the greedy approach to prioritized rule-based reasoning: among the applicable rules, always choose the one with highest priority.

The BOID semantics, if lifted to the level of arguments (which are coherent sets of rules), corresponds to 'the weakest link principle' from abstract argumentation (Dung, Prakken, Modgil, etc).

The BOID semantics resembles that of 'prioritized default logic'.

Other semantics/algorithms could be applied (last link principle, contrapositive influence, etc.)

A first example

Example (Chisholm's paradox)

Let $S = \{\top \overset{O}{\hookrightarrow} h, h \overset{O}{\hookrightarrow} t, \neg h \overset{O}{\hookrightarrow} \neg t, \top \overset{I}{\hookrightarrow} \neg h\}$.

agent type ρ: $B < I, O < D$ (desires are the less preferred)

- you need to help
- if you help, you need to tell you will
- if you do not help, you should not tell you will
- you intend not to help

two BOID extensions: $\{\neg h, t\}$ and $\{\neg h, \neg t\}$

The BOID is wrong here: only the second extension is considered correct (but the BOID is correct if ρ: $B < I < O < D$).

Solving this by imposing an order within O-rules is not a way out.

A second example

Example (Drink and drive, adapted from J. Hansen)

Let $S = \{\top \overset{\mathsf{I}}{\hookrightarrow} p, p \overset{\mathsf{D}}{\hookrightarrow} d, d \overset{\mathsf{I}}{\hookrightarrow} \neg dr, p \overset{\mathsf{O}}{\hookrightarrow} dr\}$.

agent type ρ: $B < I < O < D$

- you intend to go to the party
- if you go to the party, you want to drive
- if you drive, you intend not to drink
- if you go to the party, you are obliged to drink with friends

Should you drink? Should you drive?

The BOID goes wrong (does it?), as its only extension is: $\{p, dr, d\}$

because there is no *contra-positive reasoning* for $d \overset{\mathsf{I}}{\hookrightarrow} \neg dr$?

Outline

The simplified picture

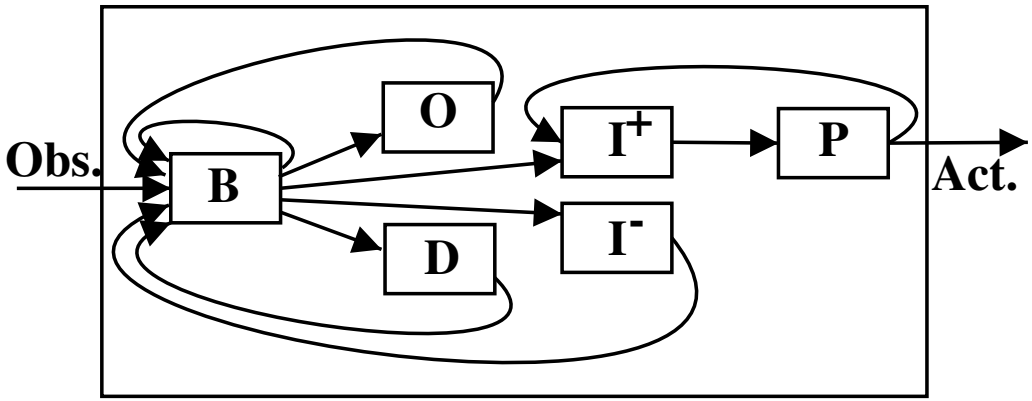


Figure: A BOID architecture and information flow

Why is the planning component P placed *after* the BOID conflict resolution mechanism?

If agents cannot make a proposition true (no plan), why reason about it in the conflict resolution mechanism?

Looking for solutions

- Every proposition might be given a ‘control level’ (link with powers, opportunities and knowing-how)
- Deal with ‘ought implies can’..
- Requires new semantics..
- BOID thinking and action thinking (dynamic logic, stit logic, etc.) have never gone together very well..

Outline

Directions of fit of propositions

proposition-to-world:

- beliefs

world-to-proposition:

- desires
- obligations
- intentions

Suspicion

- The direction of fit of propositions has an influence on the semantics of the rules.
- Why do I think that?
  - Chisholm's example
  - The issue of floating conclusions (Gabbay, Schlechta, Horty, Prakken)

Floating conclusions 1

Example (Competing economic theories, adapted from J. Horty)

Let $S = \{\top \overset{\mathbf{B}}{\hookrightarrow} i, \top \overset{\mathbf{B}}{\hookrightarrow} d, \neg(i \wedge d), i \overset{\mathbf{B}}{\hookrightarrow} edt, d \overset{\mathbf{B}}{\hookrightarrow} edt\}$.

- we will have inflation according to one group of economists
- we will have deflation according to another group of economists
- we cannot have inflation and deflation at the same time
- inflation will likely lead to an economic downturn
- deflation will likely lead to an economic downturn

Undermining: the floating conclusion '*edt*' does not seem to be justified.

The BOID goes wrong (?), as its extensions are: $\{i, edt\}$ and $\{d, edt\}$

Floating conclusions 2

Example (Housing problems)

Let $S = \{\top \overset{\mathbf{D}}{\hookrightarrow} a, \top \overset{\mathbf{D}}{\hookrightarrow} u, \neg(a \wedge u), a \overset{\mathbf{B}}{\hookrightarrow} \neg h, u \overset{\mathbf{B}}{\hookrightarrow} \neg h\}$.

- I want to start a study in amsterdam
- I want to start a study in utrecht
- if I start a study in amsterdam, it will be difficult to get a house
- if I start a study in utrecht, it will be difficult to get a house

No undermining: the floating conclusion $\neg h$ does seem justified.

The conflicting rules do *not undermine* each-other because of the direction of fit of the propositions involved?

Anscombe: we *cannot be mistaken* about our intentions, because of their direction of fit (so, opposition between intentions is different from opposition between beliefs)

Side effects

Example (Going to the dentist (Cohen and Levesque))

Let $S = \{\top \overset{\mathbf{I}}{\hookrightarrow} d, d \overset{\mathbf{B}}{\hookrightarrow} p\}$.

- I intend to go to the dentist
- if I go to the dentist, I believe I will have pain

BOID extension: $\{d, p\}$.

What is the direction of fit of the proposition p? In this case obviously that of a belief, not of an intention.

Is p a belief about the future? Should we represent time?

Side effects

Example (An adaptation of the dentist)

Let $S = \{\top \overset{\mathbf{D}}{\hookrightarrow} d, d \overset{\mathbf{B}}{\hookrightarrow} p\}$.

- I want to go to the dentist (he/she bought a Tesla, I want to see it)
- if I go to the dentist, I believe I will have pain

wrong BOID extension: $\{d, p\}$.

I obviously do not want the pain. But I also do not believe I will have pain (p should not be in an extension at all)? After all, it is only a desire to go there, not a 'decided upon' intention.

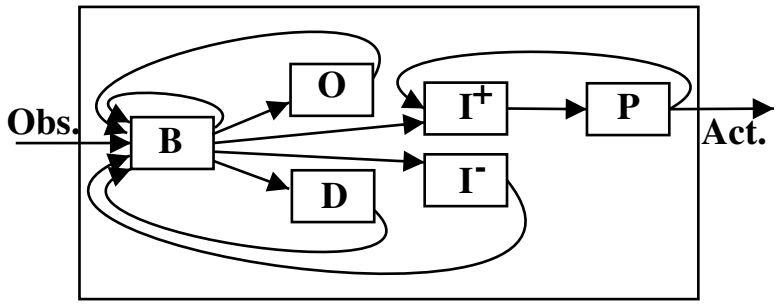
Looking for solutions

Obvious first step to make:

- give a semantics in terms of extensions where propositions are assigned a direction of fit
- define rationality constraints as invariants such extensions need to obey

Outline

Hume's guilotine



- In the BOID, along the way in the reasoning, the direction of fit of propositions *flips*. Is that a problem?
- Hume: it is problematic to make claims about what *ought* to be on the basis of statements about what *is*. But that is what we do?

Will this BOID be ethical?

- So, where is the moral source?
- Answer: in the deontic rules. And the real moral source is the programmer who provided them.
- "you need to help you neighbour" cannot count as a "true" moral source. Imagine we need to put in rules for all such situations. We need to go more general!
- Alternative moral source: learn the rules from examples through inductive logic programming $\Rightarrow$ moral relativism
- Idea: can we not make an extra BOID component containing a formal moral source with guiding general moral principles like "fairness", "proportionality", "tolerance", etc.?

From deontic logic to artificial ethical agents

- If we use deontic logic to add deontic reasoning to an artificial agent, according to what ethical theory (e.g., which of the big three) does it then make ethical decisions?
- Deontic stit logic, dynamic deontic logic: consequentialism. I/O-logic, prioritised default logic: rule consequentialism.
- Is it necessary that an artificial agent is also an artificial patient in order to be moral (it can only project 'feelings' in others if it has them itself)? (emotivism)
- Can we only make agents that can deal with legal contexts, not ethical contexts?

Outline

Conclusions / comments

- I believe it might pay off to consider directions of fit and control elements in coming to a BOID semantics
- It might be worth thinking about general moral principles like 'fairness' and try to formalise them
- We need to keep things *simple* to have an implementable approach!

Thanks

Thanks for you attention!