RESPONSIBILITY IN ACTION: The Guilty Mind (Day 4)

Jan Broersen & Hein Duijf

ESSLLI 2019
Riga

Outline

Six categories of responsibility for action

Description level \ Involvement type	Passive: allowing to happen + ability to prevent	Active: seeing to it + ability to refrain
Causal	causal omission	causal contribution
Informational	conscious omission	conscious action
Motivational	intentional omission	intentional action

Table: A responsibility matrix: six categories of responsibility

The mind in the law

Actus Reus = ‘guilty act’

Mens Rea = ‘guilty mind’

A principle of the law: show ‘concurrence’ between the two.

Modes of the current North American system (in decreasing order of culpability) [Model Penal Code, Foundation Press, 2002.]:

- **Purposefully** - the actor has the "conscious object" of engaging in conduct and believes and hopes that the attendant circumstances exist.
- **Knowingly** - the actor is certain that his conduct will lead to the result.

Mens Rea (the guilty mind)

(continued from previous slide)

- **Recklessly** - the actor is aware that the attendant circumstances exist, but nevertheless engages in the conduct that a "law-abiding person" would have refrained from.
- **Negligently** - the actor is unaware of the attendant circumstances and the consequences of his conduct, but a "reasonable person" would have been aware.
- **Strict liability** - the actor engaged in conduct and his mental state is irrelevant.

The more serious a crime, the more relevant the ‘higher’ modes.

Outline

Belnap, Perloff and Xu

From the stit bible "Facing the Future; Agents and Choices in our Indeterministic World":

"*Our strategy is to concentrate almost exclusively on the objectively causal side of indeterminism and agency, which already presents enough difficulties without bringing in non-causal concepts. We therefore lay aside many deeply important aspects of agency and choice that involve intentions, propositional attitudes, or other mental phenomena.*"

Could *stit* explain the mind in agency better than Davidson's theory?

- Could *stit* explain how proattitudes, that is, beliefs and intentions, determine specific effects?
- Belnap: "Leave the mind out!"
- My standpoint: make mind-related modalities explicit in *stit*[1].
- You might however think that the mind is already present in classical *stit*..

[1] Jesse Mulder: the difference can be seen as one between second and third gear metaphysical thinking

A non-standard, but possible *stit* interpretation

- Can we see the non-determinism in the choice cells as epistemic uncertainty about a *deterministic* world...?
- So we already see a knowledge aspect of the mind at work in classical *stit*?
- Would this give a deterministic / compatibilist interpretation of *stit*?

A non-standard, but possible *stit* interpretation

- Can we see the non-determinism in the choice cells as epistemic uncertainty about a *deterministic* world...?
- So we already see a knowledge aspect of the mind at work in classical *stit*?
- Would this give a deterministic / compatibilist interpretation of *stit*?
- But, all the axioms for agency would need to have an epistemic interpretation. I think they have not.
- Furthermore: how would intention fit in?

Back to deliberative stit

"deliberative" sounds as if there is a mental component?

- 'deliberative' *stit*:
  $[ag\ \texttt{Dxstit}]\varphi \equiv_{def} [ag\ \texttt{Cstit}]X\varphi \wedge \Diamond\neg[ag\ \texttt{Cstit}]X\varphi$
  or, equivalently
  $[ag\ \texttt{Dxstit}]\varphi \equiv_{def} [ag\ \texttt{Cstit}]X\varphi \wedge \Diamond X\neg\varphi$
  or, equivalently
  $[ag\ \texttt{Dxstit}]\varphi \equiv_{def} [ag\ \texttt{Cstit}]X\varphi \wedge \neg\Box X\varphi$

Are alternative possibilities sufficient then?

The definition of deliberative *stit* seems to suggest that the existence of alternative possibilities is not only necessary for 'deliberate' choice[2], but also sufficient..

Sufficiency is maybe to big a step; what if the agent throws a dice?

Frankfurt: for morally responsible action alternative possibilities are not even necessary[3]..!

[2] Roughly: the consequence argument, deployed by libertarians.

[3] So, the consequence argument is wrong.

Frankfurt's argument against the PAP / consequence argument

- PAP $=_{abrev}$ "the Principle of Alternative Possibilities"
- EAP $=_{abrev}$ "the Existence of Alternative Possibilities"

PAP: Intentional action $\Rightarrow$ EAP
Frankfurt examples: Intentional action $\wedge \neg$ EAP
Logical inference: Inconsistency

(Frankfurt's argument concerns moral responsibility, but here I cast it in terms of intentional action)

My position: I do think deliberate action requires alternatives, but alternatives come in many different disguises.

Intermezzo: strange for a *stit* theorist to use Frankfurt?

- compatibilism $=_{def}$ satisfiable(determinism $\wedge$ free will)
- libertarianism[4] $=_{def}$ unsatisfiable(determinism $\wedge$ free will)

Libertarians typically apply the consequence argument: free will $\Rightarrow$ EAP $\Rightarrow \neg$ determinism

But Frankfurt then attacks this main libertarian argument.

People doing *stit* theory are almost always advocating libertarianism (because of the view on 'open futures' and inherent non-determinism)..

Reconciliation: Frankfurt attacks alternative possibilities as a basis for *any* kind of action theory, either deterministic or indeterministic.

[4] not in the political but in the metaphysical meaning

EAP hides different kinds of alternatives

Determinism is a physical concept. Deliberateness is a mental issue.

Deterministic alternative possibilities are physical. Deliberate alternative possibilities are mental. $\Rightarrow$ the definition of deliberative *stit* is too simplistic.

The problem is: we have not clearly represented (the differences between) the physical and the mental aspects of acting.

Three separate description modes of action

Idea: actions come under different *mental* descriptions

Three descriptions of an agent raising its arm (variation on Wittgenstein's example)

- the objective action: it is the agent that *performs* the raise
- the informed action: the agent knowingly raises his arm
- the motivated action: the agent intentionally raises his arm

The language should then enable to make such descriptions.

Outline

Aristotle on responsibility

Aristotle: There are two components to responsibility:

- being the cause of a certain outcome
- knowing what you were (are) doing

Adding knowledge to the syntax

Add knowledge (and a next operator) to the syntax:

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

Use the standard interpretation for the next operator.

Use equivalence classes of moment-history pairs for the semantics of knowledge.

Find appropriate further constraints on the structures.

Anscombe on (un)knowingly doing

- Elizabeth Anscombe [Intention, §7, page 12]:

  "The statement that a man knows he is doing X does not imply the statement that, concerning anything which is also his doing X, he knows he is doing that thing."

- Here knowledge pertains to action: we model a mode of acting, not a static epistemic state.
- Knowledge is not moment determinate: $\not\models K_{ag}\varphi \rightarrow \Box K_{ag}\varphi$, because that does not hold for the substitution $[[ag\ \texttt{Cstit}]X\psi/\varphi]$.
- knowingly doing $\approx$ knowing how

Examples of (un)knowingly doing

- If you do not know that you carry a contagious disease, it can be that by knowingly taking a seat next to somebody you unknowingly see to it that the person is infected.
- By knowingly sending an email you may unknowingly see to it that a server breaks down.
- By knowingly signing a contract you may unknowingly see to it that you loose all your money.

How do we express this in the logic?

- *ag* 'objectively does' *p* unknowingly:
  $[ag\ \texttt{Cstit}]Xp \wedge \neg K_{ag}[ag\ \texttt{Cstit}]Xp$
- *ag* 'does' *p* knowingly: $K_{ag}[ag\ \texttt{Cstit}]Xp$
- objective possibility for *ag* to see to it that *p*: $\Diamond[ag\ \texttt{Cstit}]Xp$
- 'epistemic ability' of *ag* to see to it that *p*: $\Diamond K_{ag}[ag\ \texttt{Cstit}]Xp$

A single agent knowledge frame

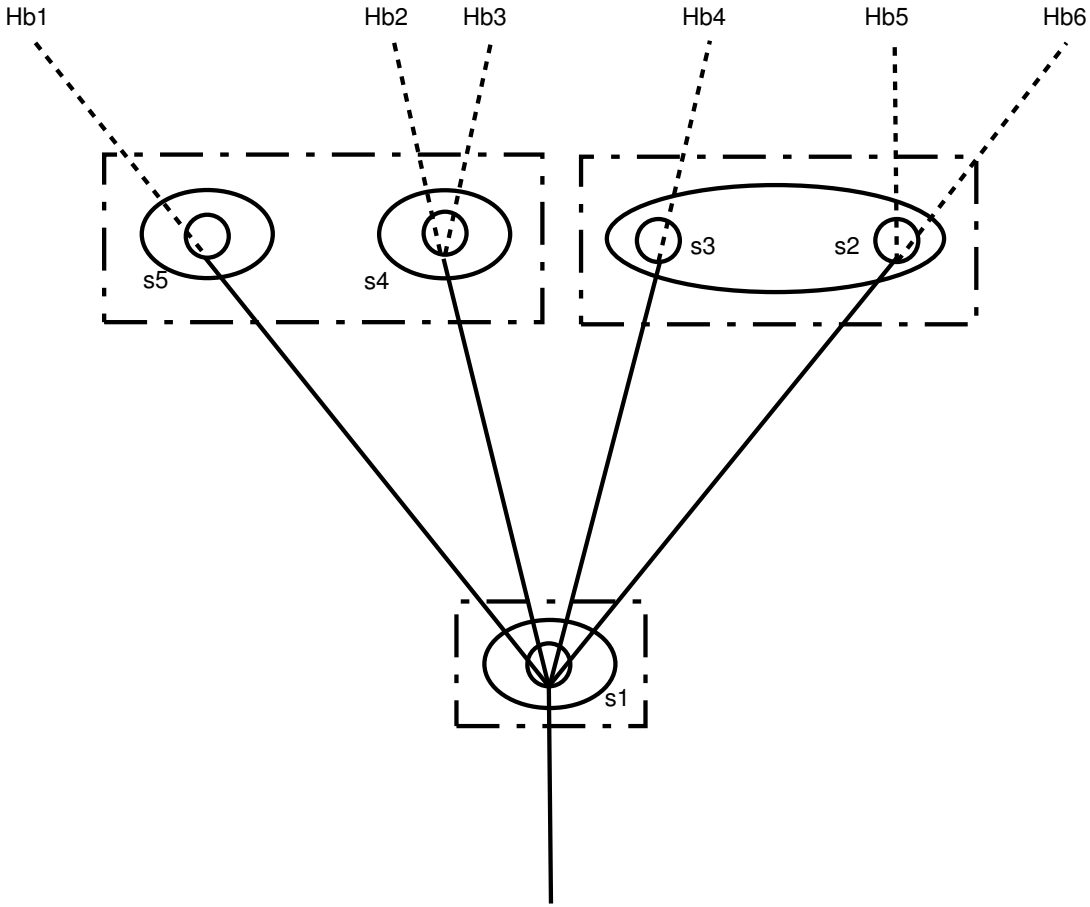

Figure: Knowingly doing in an epistemic stit frame

Sahlqvist axioms

(Fut-K) $K_{ag}X\varphi \rightarrow K_{ag}[ag\ \text{Cstit}]X\varphi$

(Rec-Eff) $K_{ag}[ag\ \text{Cstit}]X\varphi \rightarrow XK_{ag}\varphi$

(Unif-Str) $\Diamond K_{ag}[ag\ \text{Cstit}]\varphi \rightarrow K_{ag}\Diamond[ag\ \text{Cstit}]\varphi$

(K-S) $K_{ag}\Box\varphi \rightarrow \Box K_{ag}\varphi$

Uniformity of strategies

The axiom:

$$\textit{(Unif-Str)} \qquad \Diamond K_{ag}[ag\ \text{Cstit}]\varphi \rightarrow K_{ag}\Diamond[ag\ \text{Cstit}]\varphi$$

Corresponding intuition: if it is possible for you to knowingly destroy your computer, it follows that you know you have the physical capacity to destroy your computer.

Not the other way around: if you know you have the physical capacity to be rude/nice, it does not follow that it is possible for you to knowingly be rude/nice to somebody.

$$\textit{(Unif-Str')} \qquad K_{ag}\Diamond K_{ag}[ag\ \text{Cstit}]\varphi \rightarrow K_{ag}\Diamond[ag\ \text{Cstit}]\varphi$$

$$\textit{(Unif-Str'')} \qquad \Diamond K_{ag}[ag\ \text{Cstit}]\varphi \rightarrow K_{ag}\Diamond K_{ag}[ag\ \text{Cstit}]\varphi$$

Knowing how?

$[ag\ \text{Cstit}]\varphi$: "*ag* does φ"

$K_{ag}[ag\ \text{Cstit}]\varphi$: "*ag* knowingly does φ"

$\Diamond K_{ag}[ag\ \text{Cstit}]\varphi$: "*ag* has the possibility to knowingly do φ"

$K_{ag}\Diamond K_{ag}[ag\ \text{Cstit}]\varphi$: "*ag* knows it has the possibility to knowingly do φ"

Davidson on intention

Davidson's earlier view: acting intentionally = being caused to act by a pair of appropriately related mental states (a pro-attitude (desire) and an instrumental belief)

No intentions as distinct mental states.

In "Intending": intentions are all-out evaluative judgments existing as distinct and irreducible mental states.

Adding intention to the logic

Extend the syntax:

$$\varphi \ldots \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [ag\ \texttt{Cstit}]\varphi \mid X\varphi \mid K_{ag}\varphi \mid I_{ag}\varphi$$

Use equivalence classes of moment-history pairs for the semantic of intentions.

Find appropriate further constraints on the structures.

Anscombe on Intention

Anscombe's distinctions:

(1) 'Intention to act' (= an intention, yet without an act)

(2) 'Intention in action' (= an intention, accompanying an act)

(3) 'Intentional action' (= an act, accompanied by an intention)

For responsibility we are mostly interested in the third: 'intentional action'

The fascinating claims by Anscombe in "Intention"

Anscombe: "practical knowledge is knowledge without observation" (unlike other forms of knowledge)

Anscombe: "practical knowledge has a different direction of fit" (practical knowledge is of the kind where the world has to fit it, while other types of knowledge have to fit the world)

Anscombe: "practical knowledge is the cause of what it understands"

Anscombe: intentions are a form of practical knowledge (which goes against both explained views by Davidson)

The fascinating claims by Anscombe in "Intention"

Anscombe: "practical knowledge is knowledge without observation" (unlike other forms of knowledge)

Anscombe: "practical knowledge has a different direction of fit" (practical knowledge is of the kind where the world has to fit it, while other types of knowledge have to fit the world)

Anscombe: "practical knowledge is the cause of what it understands"

Anscombe: intentions are a form of practical knowledge (which goes against both explained views by Davidson)

(We will go Davidson: intention as a stand-alone mental concept)

Basic properties of the intentional action operator

Intentionally doing is (1) consistent, and (2) independent:

(Cons-I) KD for each $I_{ag}[ag\ \text{Cstit}]$

(Indep-I) $\Diamond I_{ag_1}[ag_1\ \text{Cstit}]\varphi \wedge \Diamond I_{ag_2}[ag_2\ \text{Cstit}]\psi \rightarrow \Diamond(I_{ag_1}[ag_1\ \text{Cstit}]\varphi \wedge I_{ag_2}[ag_2\ \text{Cstit}]\psi)$

Intententional actions take effect in states that are epistemically possible:

(X-Eff-I) $\Box K_{ag}\varphi \rightarrow I_{ag}[ag\ \text{Cstit}]\varphi$

Knowledge and Intention

- If I send an email, and by doing so I do not *knowingly* cause a server to break down, I clearly do not *intentionally* bring down the server by sending the email.

$$\textit{(I-K)} \qquad I_{ag}[ag\ \text{Cstit}]\varphi \rightarrow K_{ag}[ag\ \text{Cstit}]\varphi$$

- The converse is not valid: an agent killing in self-defense, kills knowingly, but does not kill intentionally.
- Recall: In law, 'purposefully' conducted acts are more culpable than 'knowingly' conducted acts.

Single agent knowledge intention frame: side effects

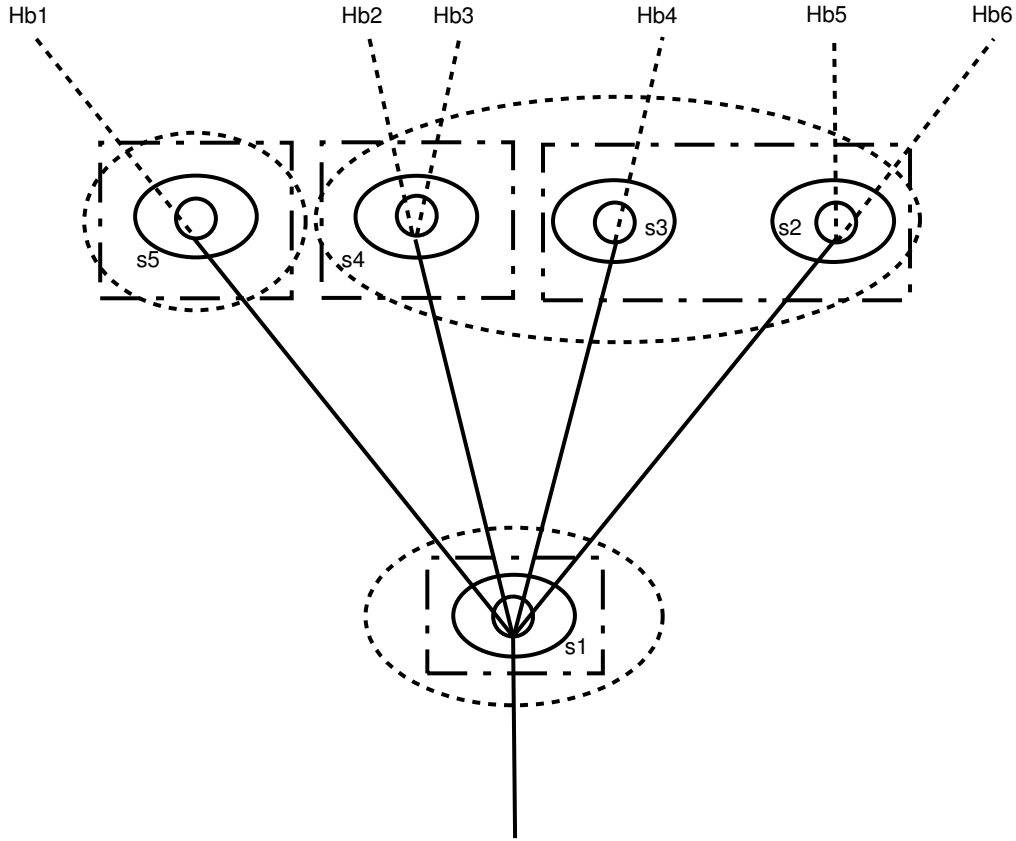

Figure: Knowingly doing and intentionally doing in a motivational epistemic frame

Side effects

The side effect problem (or my interpretation of it): intentional action should not be closed under knowingly doing

All seems ok in the picture:

1. intentionally doing implies knowingly doing (dotted boxes are inside dotted ovals)
2. what an agent knowingly does is *more* than what it intentionally does (dotted boxes encompass less states than dotted ovals), leaving room for side-effects

Problem for this formal picture: Knowingly performed side effects can only be non-intentional, if there are epistemic alternatives within the intentional set. But, why then did the agent did not take that alternative..? The typical excuse "there was no other way" cannot be valid. $\Rightarrow$ the constraints are too tight.

Possible solution

Linking side effects with side conditions:

$[ag\ \text{IntAct}]\varphi \equiv_{def} I_{ag}[ag\ \text{Cstit}]\varphi \wedge \Diamond K_{ag}\neg[ag\ \text{Cstit}]\varphi$

An intentional action must have alternatives the agent could have knowingly taken.

This does justice to the mental deliberations accompanying intentional, free will action.

Intentional action is always successful..

Axiomatically, we have that from

$I_{ag}[ag\ \text{Cstit}]X\varphi \rightarrow K_{ag}[ag\ \text{Cstit}]X\varphi$ *(I-K)* and the veridicality of knowledge we derive that

$I_{ag}[ag\ \text{Cstit}]X\varphi \rightarrow [ag\ \text{Cstit}]X\varphi$. Then with axiom *(XSett)* we derive that

$I_{ag}[ag\ \text{Cstit}]X\varphi \rightarrow X\Box\varphi$. Finally, with standard normal modal reasoning, we arrive at

$I_{ag}[ag\ \text{Cstit}]X\varphi \rightarrow X\varphi$.

What can we do?

Our conception of intentional action

An action is intentional if and only if:

- there is a causal connection (in, e.g., the but-for sense) between the agent's choice and the outcome
- the agent performed the choice with the intention to bring about the outcome
- the outcome actually obtained (success)

Our conception of intentional action

An action is intentional if and only if:

- there is a causal connection (in, e.g., the but-for sense) between the agent's choice and the outcome
- the agent performed the choice with the intention to bring about the outcome
- the outcome actually obtained (success)

But, accidentality rises its ugly face again here..

The hornets example (a deviant causal chain scenario)

"Suppose that Betty kills Jughead. The bullet she fires misses Jughead by a mile, but it dislodges a tree branch above his head and releases a swarm of hornets that attack him and sting him until he dies. [Davidson, 1980]"

- Question: was Jughead's killing of Betty an intentional kill?
- Yes: Jughead intentionally chose to kill Betty, Betty was indeed killed, and there was a causal (but for) connection between the choice and the result
- No: Jughead intentionally attempted to kill Betty, but his action failed. Betty was killed accidentally.

So, how to see this?

This is parallel to Davidson's example of the nervous climber:

Intentional state $\approx$ Intentional action

Nervousness $\approx$ Hornets

So, did we not book progress at all by going to *stit*?

A ‘weaker’ epistemic attitude towards action performance

- Belief instead of knowledge ⇒ intended action can be unsuccessful.

$$\varphi \ldots \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid X\varphi \mid [ag\ \texttt{Cstit}]\varphi \mid B_{ag}\varphi \mid I_{ag}\varphi$$

Actions can be unsuccessful

- Now, the actual history-state pair may not be epistemically accessible.
- Axiomatically, we do not have that from

  $I_{ag}[ag\ \text{Cstit}]\varphi \rightarrow B_{ag}[ag\ \text{Cstit}]\varphi$ *(I-B)* we derive that

  $I_{ag}[ag\ \text{Cstit}]\varphi \rightarrow [ag\ \text{Cstit}]\varphi$, because belief is not like knowledge veridical.

A single agent belief intention frame

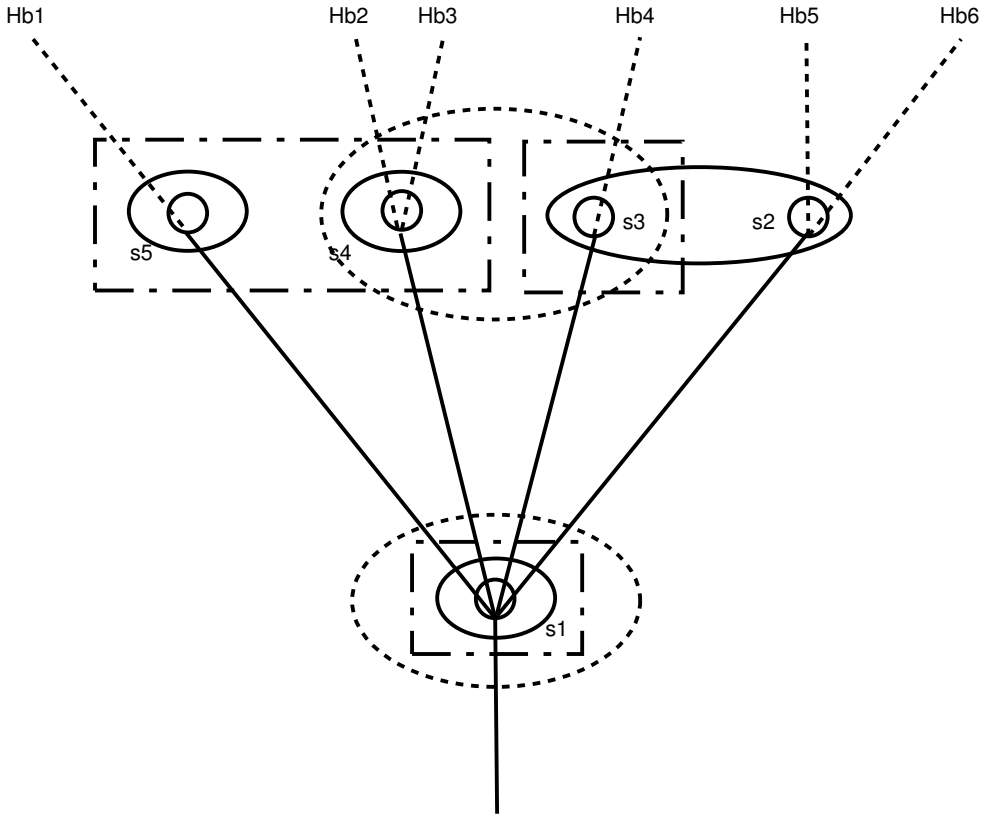

Figure: Unsuccessful action

Is $s1/Hb4$ an attempt?

Outline

Responsibility

Mode \ Modality	Could have refrained	Could have prevented
Causally	$\Diamond\neg[a\ \text{Cstit}]\varphi \wedge [a\ \text{Cstit}]\varphi$	$\Diamond[a\ \text{Cstit}]\neg\varphi \wedge \neg[a\ \text{Cstit}]\neg\varphi$
Knowingly	$\Diamond\neg K_a[a\ \text{Cstit}]\varphi \wedge K_a[a\ \text{Cstit}]\varphi$ $\Diamond K_a\neg[a\ \text{Cstit}]\varphi \wedge K_a[a\ \text{Cstit}]\varphi$	$\Diamond K_a[a\ \text{Cstit}]\neg\varphi \wedge \neg K_a[a\ \text{Cstit}]\neg\varphi$ $\Diamond K_a[a\ \text{Cstit}]\neg\varphi \wedge K_a\neg[a\ \text{Cstit}]\neg\varphi$ $\Diamond K_a[a\ \text{Cstit}]\neg\varphi \wedge \neg K_a\neg[a\ \text{Cstit}]\neg\varphi$
Intentionally	$\Diamond\neg I_a[a\ \text{Cstit}]\varphi \wedge I_a[a\ \text{Cstit}]\varphi$ $\Diamond I_a\neg[a\ \text{Cstit}]\varphi \wedge I_a[a\ \text{Cstit}]\varphi$	$\Diamond I_a[a\ \text{Cstit}]\neg\varphi \wedge \neg I_a[a\ \text{Cstit}]\neg\varphi$ $\Diamond I_a[a\ \text{Cstit}]\neg\varphi \wedge I_a\neg[a\ \text{Cstit}]\neg\varphi$ $\Diamond I_a[a\ \text{Cstit}]\neg\varphi \wedge \neg I_a\neg[a\ \text{Cstit}]\neg\varphi$

Table: The responsibility matrix (to be completed)

Outline

The crucial role of Ability

Ability is a central concept for the modelling of rational agents:

- Abilities are the basis for planning and decision making (in AI)
- Abilities are the basis for cooperation and negotiation
- Abilities come with responsibilities (reversing "ought implies can")
- Absence of ability is the most heard excuse to avoid responsibilty

More on Ability

The notion of ability stresses the crucial role of non-determinism: Abilities are not just $\Diamond[ag\ \texttt{stit}]\varphi$?

- If one hits the bull's eye, it does not necessarily follow that one has the ability to do so
- If one is able to hit the bull's eye, it does not follow that one always will (so better $\Diamond[ag\ \texttt{stit}^{\geq c}]\varphi$?)

Even more on Ability

Ability involves 'mental' capacities:
Abilities are not just $\Diamond[ag\ \texttt{stit}]\varphi$?

- Ability requires knowing how
- Ability may require intention (according to some folk psychological theories on self-confidence)

Types, tokens and epistemic abilty

Epistemic ability: Horty and Paquit argue that $\Diamond K_{ag}[ag\ \text{Cstit}]X\varphi$ does not adequately represent a notion of epistemic ability.

They propose a new operator $[ag\ \text{Kstit}]X\varphi$ that is interpreted in *stit* frames with action types added to them.

In our paper accepted (minor revisions) for the Review of Symbolic Logic we show a translation to standard epistemic stit as used here.

What does this tell us about the nature of action types?

Yet more on Ability

There is a natural link between abilities and:

- powers
- dispositions
- affordances
- opportunities
- potentialities (Aristotle)

Yet, philosophers have not been able to define these concept in a way that clearly distinguishes and relates them.

Outline

Future work

Solve the problem with side effects

Understand and model ‘in-action’ / being passive

Group moral responsibility? Free will group action?

Complete the formalisation of the responsibility matrix

Thanks

Thanks!