RESPONSIBILITY IN ACTION: Agency and Ability (Day 2)

Jan Broersen & Hein Duijf

ESSLLI 2019
Riga

The main message from yesterday

Understanding and modelling agency is crucial for ultimately programming responsible intelligent systems.

Between the three central notions involved in responsibility (Agents, Actions, Norms), agency is a theory linking the first two.

Today is about formalizing agency.

Plan for today

1. Formal agency
2. Chellas stit
3. Deliberative stit
4. Time
5. Achievement stit
6. Interventions in stit
7. More variants
8. Ability

Why a logic approach?

My claim was that formalization is required. But why am I advocating to focus on logic?

[1] Kantian generalization? Artificial emotion? The golden rule of ethics? etc.

Why a logic approach?

My claim was that formalization is required. But why am I advocating to focus on logic?

Answer: maybe I am just old-fashioned in believing that reasoning is the key component in modelling responsibility.

- Deontic logic is about reasoning in the context of normative systems. "What should I do?"
- Action logics are about reasoning in the context of action ⇒ practical reasoning. "What should I do?"

But: if we implement deontic action reasoning in a moral artificial agent, we miss a crucial part: what is this agent's 'moral source[1]'?

[1] Kantian generalization? Artificial emotion? The golden rule of ethics? etc.

Ontological commitments of action logics from AI

Logic \ Action	Types	Execution	Reach	Agency	Eff. Des.	Reali.
Modal Action Logic	Yes	conditional	one step	none	open	next
Situation Calculus	Yes	conditional	one step	none	open	next
Dynamic Logic	Yes	conditional	extensive	none	open	end
FO Dynamic Logic	Eff.	conditional	extensive	none	closed	end
Coalition Logic	Yes/No	conditional	one step	group	open	n.a.
sub-game operator	Eff.	conditional	one step	group	closed	n.a.
ATL	Yes/No	conditional	extensive	group	open	strat.
BIAT (Segerberg)	Eff.	conditional	extensive	single	closed	end
PAL / DEL	Eff.	cond. + det.	one step	none	closed	end
Belnap *stit*	No	actual	instant.	multi	open	imm.
Group *stit*	No	actual	instant.	group	open	imm.
XSTIT	No	actual	one step	group	open	next
Strategic STIT	No	actual	extensive	group	open	strat.

What is agency?

- [Davidson in "agency" 1971]: "*What events in the life of a person reveal agency; what are his deeds and his doings in contrast to mere happenings in his history; what is the mark that distinguishes his actions?*"
- Computer science: agency $=$ anything associated to the modelling of agents..
- Closely related: mind-body problem; how does the mind resort effects in the physical world? (and the other way around)
- The de facto standard theory of agency in philosophy is Davidson's *causal* theory.

Davidsonian action theory

- Events are the basic elements of action theory, action can be explained in terms of them
- Events come under different descriptions (a coarse grained view): "the stabbing of Ceasar" = "the killing of Ceasar"
- Agency = reasons for acting *causing* effects in the world (a monistic view). But, why do we not have concrete explanatory causal theories of this phenomenon?
- Anomalous monism: causation only at the event token level (event particulars), no descriptions at the type level

Causality

David Hume: causal truths are empirical truths (that is, they are not analytical)

Democritus (430-380 BC): "*I would rather discover one causal relation than be King of Persia*"

Bertrand Russell: "*The reason why physics has ceased to look for causes is that, in fact, there are no such things. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.*"

Deviant causal chain examples contra Davidson

But, all the criteria for a physical mental state causing an action can be met without this adding up to a case of agency..

"*A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally.*" (Davidson, 1980: 79).

Very similar: The Inexperienced Criminal [Frankfurt, 1978: 157], A Philosopher's Worries [Mele, 1992: 182], The Marriage Proposal [Davis, 1994: 113], etc.

The *stit* (= seeing to it that) picture

- No theory of events
- Actions are identified with effect descriptions (a fine grained view), relative to agents (central operator: $[ag\ \texttt{stit}]\varphi$): "the stabbing of Ceasar" ≠ "the killing of Ceasar"
- *Stit* focusses on the logical properties

Stit theory approaches the problem of agency "top-down", not committing to a Davidsonian "bottom-up" picture.

Stit abstracts away from underlying causal structure. Davidsonian problems of 'mental causation' can simply be left unconsidered.

Since *stit* theory is less committing, it seems more 'cautious'.

The *stit* picture (continued)

- *stit* theory assumes our world is non-deterministic
- *stit* logics define actions as relations between agents and effects: $[ag\ \texttt{stit}]\varphi$ means "agent *ag* ensures the world is among those satisfying φ"
- *Temporal* version of *stit* take 'histories' as worlds: acting = necessitating sets of histories

  So, *stit* may be said to add the witnesses of branching (i.e., choices by agents) to the branching time logics CTL / CTL*

A third view: dynamic logic as a theory of action

- Dynamic logic [Pratt 1976] originates in computer science and is (was) meant for reasoning about programmes.
- Central operator: $[\alpha]\varphi$ with α a complex programmatic schema (a program) built from atomic actions and program constructors.
- One of the program constructors is non deterministic choice $\cup$. This enables 'vague' action types.
- A program 'α' is interpreted as a transition from one program state to a set of other program states.

Dynamic logic as a logic of action

- In the 90ies, many envisioned that dynamic logic could also be a logic of action: [Moore 1980], [Meyer 1988], [Cohen and Levesque 1990], etc.
- However, this only makes sense if we see the basic actions as 'action types'.
- Example: $a =$ 'opening a door', which is not the same as the event of the door opening at a particular time by a particular agent.
- I know of no successful approach (in my view) that manages to add agency to dynamic logic (Segerberg's action theory comes closest).

How do the three theories compare?

- STIT and Davidson agree on particulars, but STIT builds on non-determinism and Davidson on causation
- Davidson and DL agree on a bottom-up-approach, but Davidson builds on particular causal events and dynamic logic on action types
- STIT and DL agree on.. propositional logic and modality

Outline

Chellas stit

The CSTIT syntax:

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

Reading of $\Box\varphi$: "independent of what is happening or currently done by some agent, φ is true".

Reading of $[ag\ \texttt{Cstit}]\varphi$: "agent *ag* sees to it that φ is true".

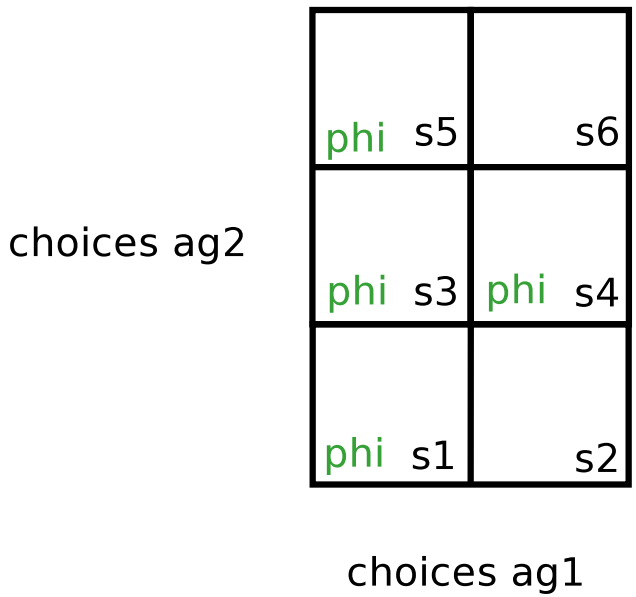
That is, agent *ag* is responsible for the occurrence of φ; independent of all things that might occur, *ag*'s involvement is such that φ is guaranteed.

The idea of the Chellas stit semantics

We use standard S5 normal multi-modal logic; equivalence sets are depicted as squares and rectangles.

For basic Chellas *stit* we do not need temporal structure.

- $s_4 \models \neg[ag_1\ \text{xstit}]\varphi$
- $s_4 \models [ag_2\ \text{xstit}]\varphi$
- $s_4 \models \Diamond[ag_1\ \text{xstit}]\varphi$



Axioms Chellas stit

The Xu axiomatisation:

	The S5 axioms for $\Box$
	For each *ag* the S5 axioms for $[ag\ \texttt{Cstit}]$
(SettC)	$\Box\varphi \rightarrow [ag\ \texttt{Cstit}]\varphi$
(Indep)	$\Diamond[ag_1\ \texttt{Cstit}]\varphi \wedge \ldots \wedge \Diamond[ag_n\ \texttt{Cstit}]\psi \rightarrow$
	$\Diamond([ag_1\ \texttt{Cstit}]\varphi \wedge \ldots \wedge [ag_n\ \texttt{Cstit}]\psi)$
	for $Ags = \{ag_1, \ldots, ag_n\}$

Independence is the main axiom. Why is independence so central / interesting?

Axioms Chellas stit

The Xu axiomatisation:

	The S5 axioms for $\Box$
	For each *ag* the S5 axioms for $[ag\ \texttt{Cstit}]$
(SettC)	$\Box\varphi \rightarrow [ag\ \texttt{Cstit}]\varphi$
(Indep)	$\Diamond[ag_1\ \texttt{Cstit}]\varphi \wedge \ldots \wedge \Diamond[ag_n\ \texttt{Cstit}]\psi \rightarrow$
	$\Diamond([ag_1\ \texttt{Cstit}]\varphi \wedge \ldots \wedge [ag_n\ \texttt{Cstit}]\psi)$
	for $Ags = \{ag_1, \ldots, ag_n\}$

Independence is the main axiom. Why is independence so central / interesting?

It shows that influence on other agents (independence says 'none'), which is central from a responsibility perspective (El Paso) is also the major source for logical complication.

Outline

Deliberate action as the availability of alternatives

- Given the correctness of Newton's laws, it is true of our world that [*Jan* `Cstit`]"F = m $\times$ a".
- And at any moment where the sun disappears at the horizon, it holds that [*Jan* `Cstit`]"sun disappears".
- But, "responsibility for φ" in [*Jan* `Cstit`]φ should mean that *without* the agent's involvement, φ is not guaranteed..

Deliberate action as the availability of alternatives

- Given the correctness of Newton's laws, it is true of our world that [*Jan* `Cstit`]"F = m × a".
- And at any moment where the sun disappears at the horizon, it holds that [*Jan* `Cstit`]"sun disappears".
- But, "responsibility for φ" in [*Jan* `Cstit`]φ should mean that *without* the agent's involvement, φ is not guaranteed..
- 'deliberative' *stit*:
  $[ag\ \texttt{Dstit}]\varphi \equiv_{def} [ag\ \texttt{Cstit}]\varphi \wedge \Diamond\neg[ag\ \texttt{Cstit}]\varphi$
  or, equivalently
  $[ag\ \texttt{Dstit}]\varphi \equiv_{def} [ag\ \texttt{Cstit}]\varphi \wedge \Diamond\neg\varphi$
  or, equivalently
  $[ag\ \texttt{Dstit}]\varphi \equiv_{def} [ag\ \texttt{Cstit}]\varphi \wedge \neg\Box\varphi$

Outline

Adding temporal operators (Belnap, Horty)

The Chellas stit syntax is extended with temporal operators.

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [ag\ \texttt{Cstit}]\varphi \mid F\varphi \mid P\varphi$$

Reading of $F\varphi$: "at some point in the future φ is true".

Reading of $P\varphi$: "at some point in the past φ is true".

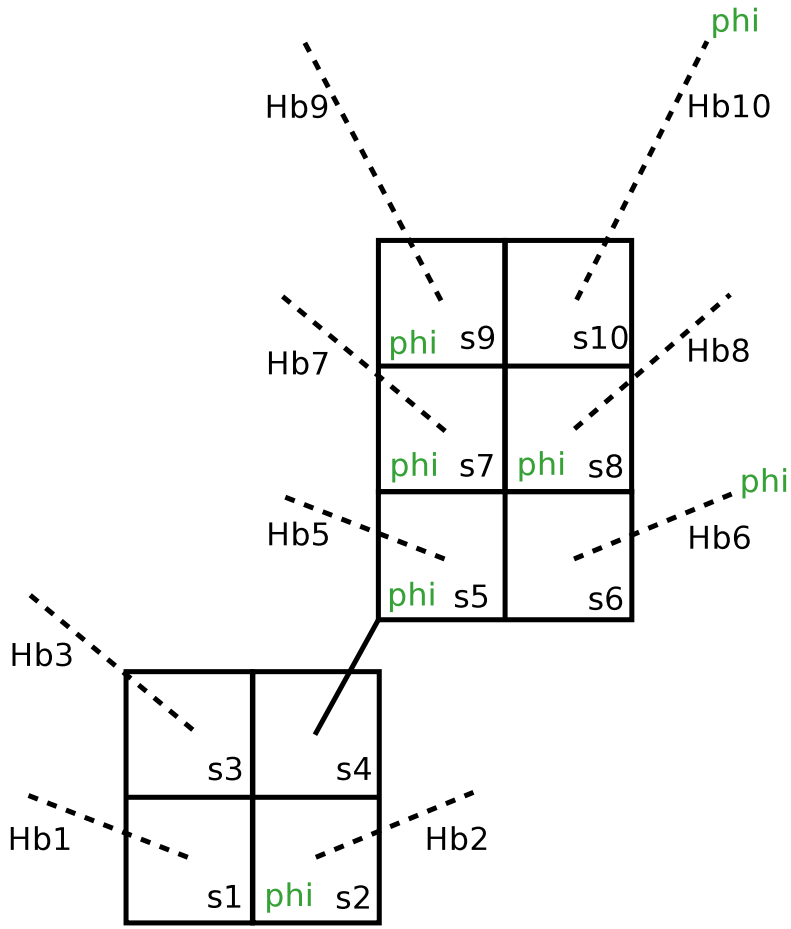
Evaluation of truth is with respect to "state-history" pairs $\langle s, h \rangle$

The idea of the Belnap/Horty *stit* semantics

The evaluation space is now two-dimensional (Arthur Prior: the Ockhamist solution to the problem of future contingencies).

convention: $h_1 \in Hb_1, h_2 \in Hb_2$, etc.

- $\langle s_2, h_2 \rangle \models [ag_1\ \texttt{Cstit}]F\varphi$
- $\langle s_4, h_5 \rangle \models F[ag_1\ \texttt{Cstit}]\varphi$



More on future contingency

Can Fp have a definite truth value if the future is contingent?

One approach: supervaluations

The dispute about the 'thin red line': one history should be singled out as 'special'

The CTL/ATL solution: let path formulas only appear in the context of path quantifiers. Here: $\Box F\varphi$ and $\Diamond F\varphi$ and $[ag\ \texttt{Cstit}]F\varphi$ and $\langle ag\ \texttt{Cstit}\rangle F\varphi$.

The Ockhamist stit 'solution' is similar, but more 'elegant' (solution in the semantics instead of in the syntax)

More on Belnap/Horty *stit* semantics

- No axiomatization proven (although for the non-group case this seems doable)(update: my visiting PhD student Jianfeng He claims to have a proof).
- The semantics of [*ag* `Cstit`]*Fp* is about *ag* making an instantaneous choice that select histories ensuring *p* at some unspecified moment in their futures. Is that realistic?
  (my opinion is that it calls for strategic versions of stit)

Outline

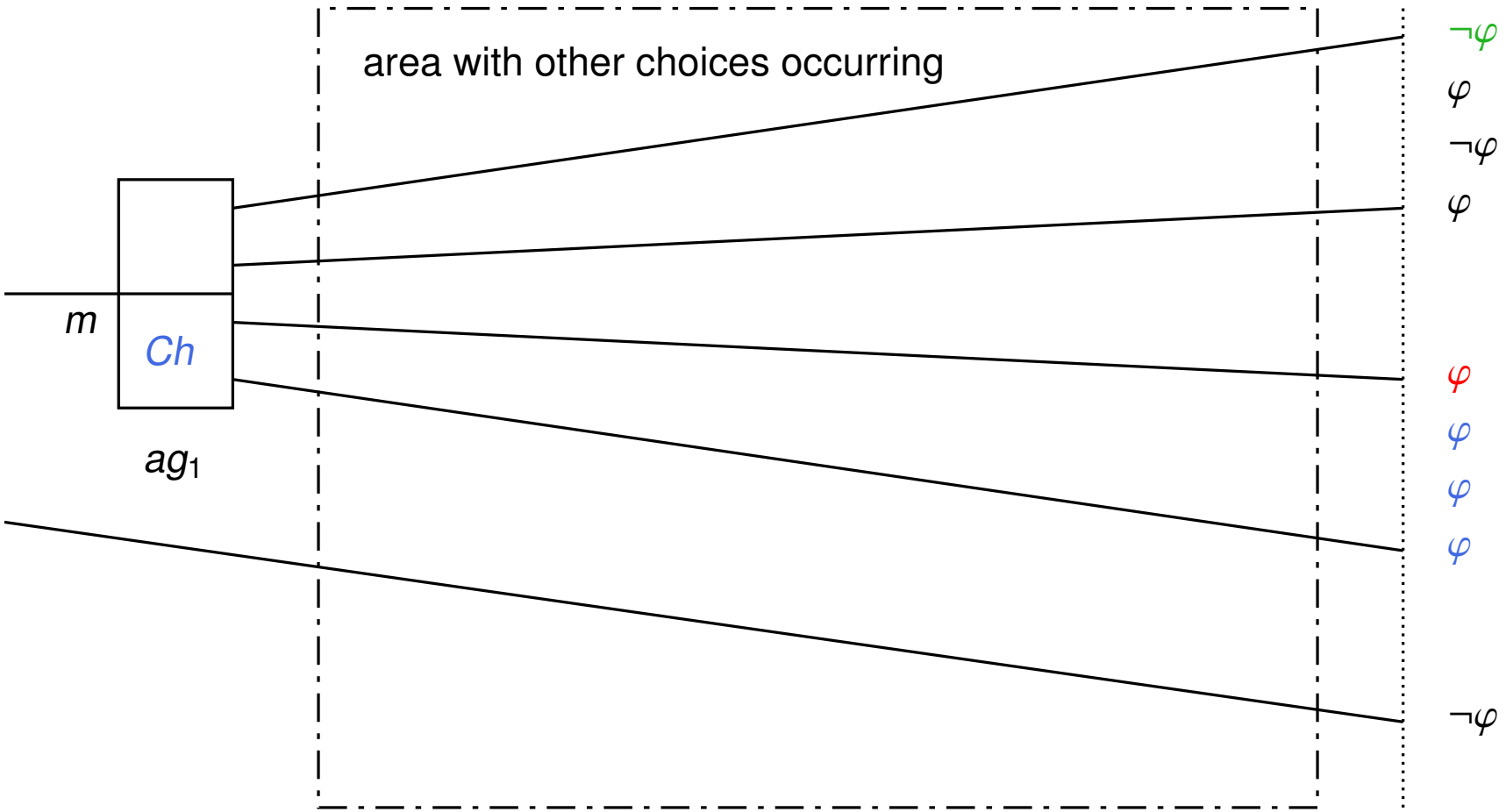
Definition of the achievement stit in "Facing the Future"

Using this auxiliary concept, the rule for evaluating an achievement stit, $[\alpha$ *astit:* $A]$, at a point m/h of a $BT+I+AC$ model $\mathfrak{M}$ can now be set out as follows (§8G.3):

2-4 DEFINITION. *(Truth for the achievement stit)* $\mathfrak{M}, m/h \vDash [\alpha$ *astit:* $A]$ iff there is a moment $w < m$ such that (i) for all moments m_1 $\underline{Choice}^{\alpha}_{w}$-equivalent to m, we have $\mathfrak{M}, m_1/h_1 \vDash A$ for all $h_1 \in H_{(m_1)}$; and (ii) there is some moment $m_2 \in i_{(m)}$ such that $w < m_2$ and $\mathfrak{M}, m_2/h_2 \nvDash A$ for some $h_2 \in H_{(m_2)}$.

This formidable definition (which is equivalent to that of §8G.3) can be grasped more easily by reference to Figure 2.4, depicting a situation in which $[\alpha$ *astit:* $A]$ is true at m/h, as a result of an action by α at the prior moment w, known as a *witness*, which action is determined as effective by A being possibly false at the moment m_2, known as the *counter*.[6] The evaluation rule embodies two re-

The achievement stit [Belnap and Perloff 1992], explained in a model



Explainging the semantics of achievement stit

The achievement stit has 4 essential elements in its truth condition. Informally, these are:

$\langle m, h\rangle \models [ag\ \texttt{astit}]\varphi \Leftrightarrow$
there is a choice *Ch* such that:
(1) *Ch* is a choice of agent *ag*
(2) *Ch* is in the past of m and led to the current situation $\langle m, h\rangle$
(3) *Ch* necessitated that currently φ
(4) *Ch* had an alternative that would not have necessitated that currently φ

Can there be more than one witness?

Can there be more than one witness?

No.

An earlier witness obeying (3) and a later witness obeying (4) do not go together.

So, if there is a witness, it is unique.

Properties of the achievement stit

- There are only two tastes of (causal) responsibility: effects are either (1) necessitated (the active mode) or (2) allowed (the passive mode)
- If an agent is responsible for an effect φ, no other agents involved after the moment of acting carry any responsibility (the same is not true for agents before).

The logic of the achievement stit

The operator $[ag\ \texttt{astit}]\varphi$ has been axiomatised for groups G (in stead of individuals ag) and relative to specific classes of structures (e.g. not allowing so-called 'busy choosers').

Axiomatisations are quite involved (which is why I do not show them) and meta-logical results (e.g. decidability) very difficult to prove.

Hope: a simple axiomatisation of achievement stit relative to our simplified structures (not shown today).

Applying the achievement stit to the desert example

Looking back from the proposition ‘death’,

- Assume enemy 2 had no local causal alternative (he did not carry fresh water with him)
- Then enemy 1 ‘necessitated’ death (assume..)
- Enemy 1 had an alternative
  (assume there are histories where 1 did not poison the water and 2 did not empty the canteen)
- So, Enemy 1 is responsible, Enemy 2 is not

But:

Applying the achievement stit to the desert example

Looking back from the proposition ‘death’,

- Assume enemy 2 had no local causal alternative (he did not carry fresh water with him)
- Then enemy 1 ‘necessitated’ death (assume..)
- Enemy 1 had an alternative
  (assume there are histories where 1 did not poison the water and 2 did not empty the canteen)
- So, Enemy 1 is responsible, Enemy 2 is not

But:

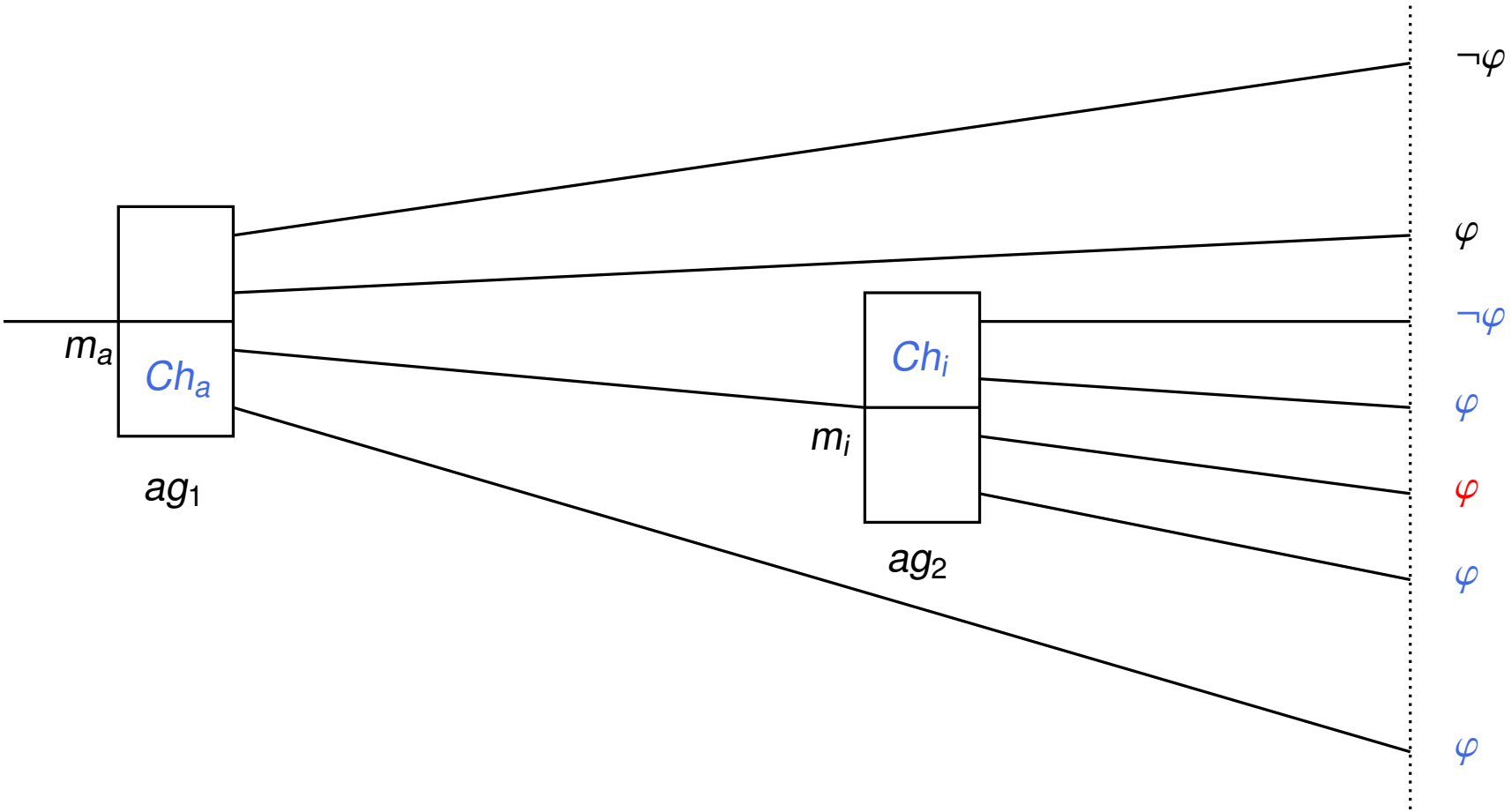
- If enemy 2 had the alternative to refill, *even though he did not take it*, causal responsibility would jump to enemy 2 alone!?

Outline

What are interventions (in our framework)?

- Performed by agents, not by non-human nature (different from Anscombe)
- Also an intervention by an agent is a *genuine* choice: something can only be an intervention if an agent has the possibility *not* to intervene
- An intervention cancels *directedness* of a course of events at a certain property. If an available intervention is not performed, we say the non-intervention is itself *directed* at that property.

The interventionist stit by way of an example



In words

We look for a past situation $\langle m_a, h \rangle$ whose associated choice Ch_a is a witness for φ, making φ's true 'in general'.

'in general' means that exceptions $\neg\varphi$ are allowed only if 'within the temporal cone' determined by $\langle m_a, h \rangle$, we can find a witness for the exception $\neg\varphi$ that is such that the intervening agent could also *not* have intervened.

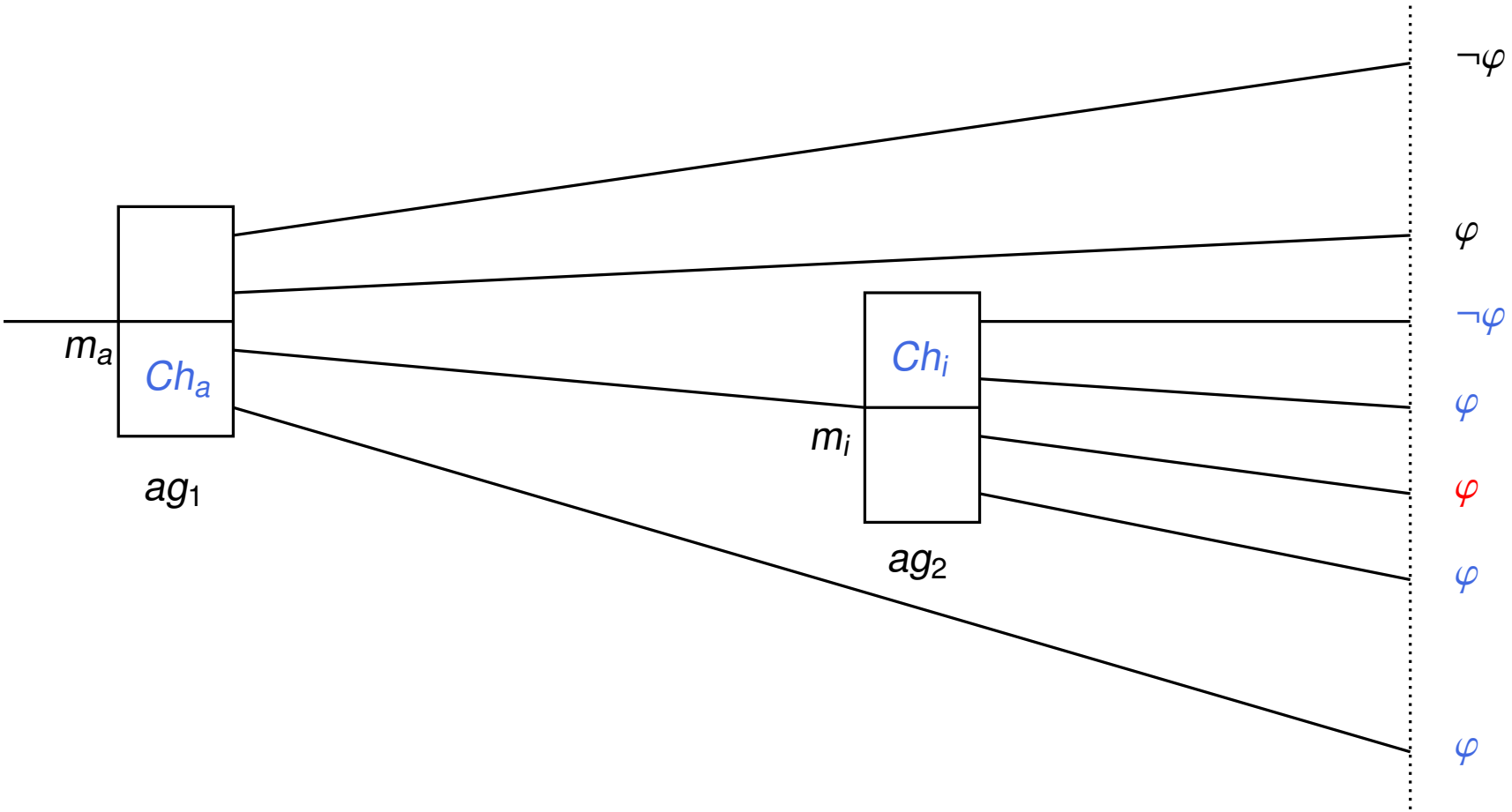
In words

We look for a past situation $\langle m_a, h \rangle$ whose associated choice Ch_a is a witness for φ, making φ's true 'in general'.

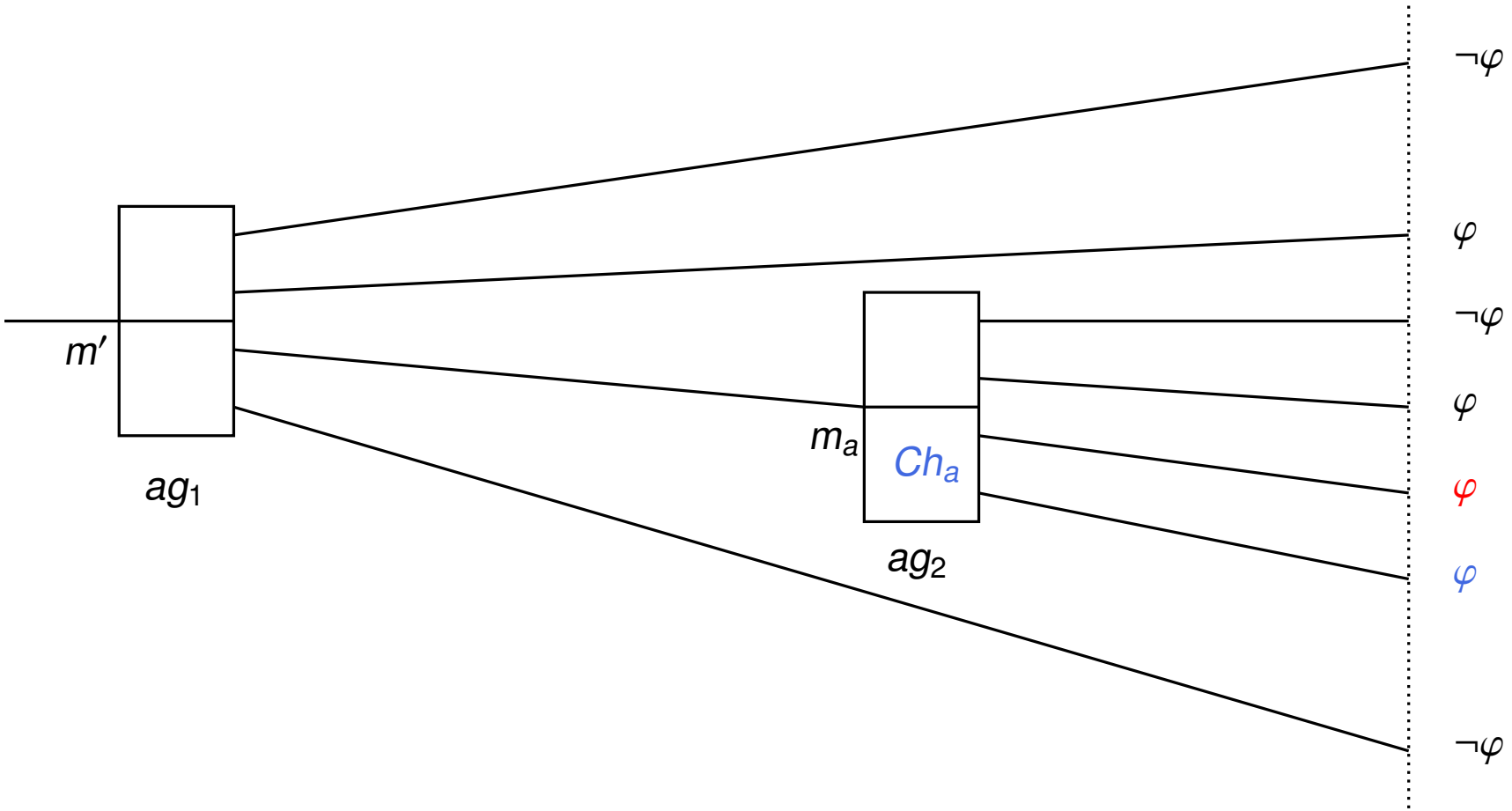
'in general' means that exceptions $\neg\varphi$ are allowed only if 'within the temporal cone' determined by $\langle m_a, h \rangle$, we can find a witness for the exception $\neg\varphi$ that is such that the intervening agent could also *not* have intervened.

Back to the picture: is the choice of ag_2 not a more obvious witness? And if so, ag_2 did it!?

The interventionist stit by way of an example



The interventionist stit; here agent 2 *is* the witness



The central ideas for an interventionist version of stit

An agent *ag* is responsible for having seen to it that φ, if and only if

1. φ actually holds
2. there is a witnessing choice Ch_a of *ag* somewhere in the past that was '*directed at* φ', which means:
   - if no other agent intervenes, φ is guaranteed to occur, and
   - intervention of an agent in a 'φ-directed' course of events, implies it must also have the choice *not* to intervene
3. there is no earlier witness Ch' obeying these properties

New operators

We introduce 3 new core backward (bw)operators:

[*ag* bw.dir] reads "agent *ag* acted in a way directed at φ"
[*ag* bw.init] reads "agent *ag* initialized directedness at φ"
[*ag* bw.prep] reads "agent *ag* has prepared for an initialisation of φ"

Our three forms of responsibility:

(1) Preparing: $[ag\ \texttt{bw.prep}]\varphi$

(2) Initializing: $[ag\ \texttt{bw.init}]\varphi$

(3) Not intervening: $[ag\ \texttt{bw.letgo}]\varphi \equiv_{def} [ag\ \texttt{bw.dir}]\varphi \wedge \neg[ag\ \texttt{bw.init}]\varphi$

The interventionist version of stit:
$[ag\ \texttt{intv.stit}]\varphi \equiv_{def} \varphi \wedge [ag\ \texttt{bw.init}]\varphi$

Theorem

On a given structure:

- *achievement implies directedness*
- *achievement does not imply the interventionist stit*

Theorem

Interventionist stit witnesses are unique

Etc...

Applying the interventionist stit to the desert example

Looking back from the proposition ‘death’,

- Assume enemy 2 had no local causal alternative (he did not carry fresh water with him)
- Then enemy 1 ‘necessitated’ death
- Enemy 1 had an alternative
  (assume there are histories where 1 did not poison the water and 2 did not empty the canteen)
- So, Enemy 1 is responsible, Enemy 2 is not (he could not intervene)

But:

Applying the interventionist stit to the desert example

Looking back from the proposition 'death',

- Assume enemy 2 had no local causal alternative (he did not carry fresh water with him)
- Then enemy 1 'necessitated' death
- Enemy 1 had an alternative
  (assume there are histories where 1 did not poison the water and 2 did not empty the canteen)
- So, Enemy 1 is responsible, Enemy 2 is not (he could not intervene)

But:

- If enemy 2 had the alternative to refill, while he did not take it, causal responsibility would be with *both* enemy 1 and 2. Enemy 1 'initialised' and enemy 2 'let it go' (could have intervened).

Outline

Many more stit variants..

- strategic stit: $[ag\ \texttt{sstit}]\varphi$ with the reading "*ag* strategically sees to it that φ"
- spatial stit: $[ag\ \texttt{estit}]\varphi$ with the reading "*ag* is effective for φ in Newtonian time and space" (which requires the spatial opportunity and the agentive power to do so)
- probabilistic stit: $[ag\ \texttt{stit}^{\geq c}]\varphi$ with the reading "*ag* sees to it that φ obtains with a chance of minimally c"

Outline

Ability in AI

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

- Abilities are the basis for planning
- Abilities are the basis for cooperation and negotiation
- Abilities come with responsibilities (reversing "ought implies can")

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)

$\Rightarrow$ Thursday's lecture on mental attitudes (the guilty mind)

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.

Thanks

Thanks!