RESPONSIBILITY IN ACTION: The many faces of responsibility (Day 1)

Jan Broersen & Hein Duijf

ESSLLI 2019
Riga

Practicalities

- Lecture: 9.00-10.30 (followed by Coffee Break, next lecture starts at 11.00)
- Location: Room 18
- Small additional break: 9.45 - 9.55

The Lecturers

Jan Broersen

- Professor of Philosophy and Logical Methods in Artificial Intelligence @ Utrecht University
- Recently finished project: "Responsible Intelligent Systems" (2014 - 2019)

Hein Duijf

- PhD in Philosophy @ Utrecht University (2014 - 2018)
- Postdoc in Philosophy @ VU Amsterdam (current)
- Projects: "Responsible Intelligent Systems" and
- "Social Epistemology of Argumentation"
- Themes: Responsibility, Argumentation, Cooperation, and Technology

What to expect?

- Day 1 - The Many Faces of Responsibility
- Day 2 - Agency and Ability
- Day 3 - Responsibility and Rational Action
- Day 4 - The Guilty Mind
- Day 5 - Responsibility in Collective Contexts

What is responsibility?

responsible $\approx$ answerable
if something goes wrong, people can come to you
and demand an answer

An agent is responsible for an action (outcome)
if and only if
we can justifiably/correctly blame or praise her
for that action (outcome)

Without norms, there is no need for responsibility.
Without responsibility, norms lose their grip.

Example sentences

- He was responsible for the breaking of the glass
- You are responsible for not having helped her
- You are responsible for taking care of the children
- It is the buyers responsibility to look for defects
- It is the sellers responsibility to report defects
- He took responsibility for the death of the plants
- He is a responsible person
- Gun manufacturers are not responsible for mass shootings
- Sigaret producers are responsible for lung cancer

Plan for today

1. Action and blame
2. Forward and backward
3. Moral and legal
4. Agents
5. Agency
6. Modes of acting
7. Responsible AI

Responsibility, action and blame

Claim: responsibility issues arise only in contexts with:

- agents (bearers of responsibility and authors of acts)
- acts (by agents, affecting outcomes, often through or in cooperation with other agents)
- norms (to evaluate acts as either blame- or praiseworthy)

Blame (and praise) games

Rational agents will want to avoid responsibility for bad outcomes and assert responsibility for good outcomes.

This results in blame and praise ‘games’.

These games revolve around:

- norms and their violation conditions
- how actions of different agents influence each other
- reasons for performing actions
- finding excuses

The recent ‘El Paso’ blame game

Responsibility for the El Paso mass shooting rests with:

- Patrick Crusius
- Donald Trump
- Guns
- Republicans
- Video games
- The AltRight
- 8chan
- Cloudflare
- Mexican immigrants in the US
- Drag queens and fatherlessness
- Barack Obama

Outline

Forward versus backward looking responsibility

(1) backwards looking responsibility:

- the position a judge typically takes,
- liability,
- causality

(2) forward looking responsibility:

- what should I do?
- moral decision making,
- ethical knowledge and reasoning,
- learning from moral mistakes

Issues in tracing back responsibilities from outcomes

1. World conditions do not follow with certainty from (intentional) actions. Problems:
   - Luck / accidentality
   - Levels of responsibility (relative to uncertainty about effects?)
   - Attempt (what *is* exactly an attempt?)
2. How do we trace back collective responsibility?
3. How do we trace back shared / partial responsibility?
4. Where to stop back-tracing?

Issues in responsible decision making

1. How is moral decision making different from other forms? What are rational moral decisions?
2. Moral dilemma's and excuses.
3. How do we choose between different ethical theories?
4. Is there symmetry between forward and backward looking responsibility? Why not? (moral luck?)

Outline

Quote

"Good people do not need laws to tell them to act responsibly, while bad people will find a way around the laws" - Plato.

(Too pessimistic in my view)

Difference between moral an legal responsibility

Obviously strongly related, but..

- Legal: normative system is explicit (text based)
  Moral: normative system is implicit
- Moral: Intent and Knowledge are crucial,
  Legal: not necessarily (e.g. strict liability)
- Legal: coercion is always an excuse (not only regimentation),
  Moral: heavily debated (PAP and Frankfurt cases!)

Difference between moral an legal responsibility

- Legal: groups can be responsible (corporations),
  Moral: heavily debated
- Legal: normative system reflects (political) agreement on explicit rules and enforcement,
  Moral: not necessarily (Kant: objectively 'true' moral rules)
- Moral: not clear if one can be morally lucky,
  Legal: clearly one can be legally lucky

Upshot: Legal responsibility is a reflection of moral responsibility, but has extra functions (retribution, reform, deterence.)

Outline

Quotes on who are responsible

"There is an expiry date on blaming your parents for steering you in the wrong direction; the moment you are old enough to take the wheel, responsibility lies with you." - J.K. Rowling

"We must reject the idea that every time a law's broken, society is guilty rather than the lawbreaker. It is time to restore the American precept that each individual is accountable for his actions." - Ronald Reagan

"It is our collective and individual responsibility to preserve and tend to the environment in which we all live." - Dalai Lama

What counts as an agent that can be responsible?

A moral community consists of moral agents and moral patients.

Only agents can be morally responsible. But what counts as an agent? Many questions.

- Are animals agents?
- Are children agents?
- Are groups agents?
- Are ‘future generations’ patients?

More on the category of agents

Many necessary conditions for qualifying as a moral agent / patient (emotions, feelings, reasoning, intentionality, sociality, etc.), but no agreement on conjunctions that give sufficient conditions.

Or: is attribution of agency enough (Dennett) to qualify as an agent?

A straightforward claim: "only agents can exhibit agency". But, what is that?

Outline

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?*"
- Incorrect use in 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

Deviant causal chain examples contra Davidson

Normal form examples: The nervous climber (not the intentional state to let go the rope caused the fall, but the nervousness) [Davidson, 1980: 79]. = The Inexperienced Criminal [Frankfurt, 1978: 157], A Philosopher's Worries [Mele, 1992: 182], The Marriage Proposal [Davis, 1994: 113]

Extensive form examples: The hornets / wild pigs (not the bullet killed the victim, but the alerted hornets / pigs) [Bennett, 1965; Davidson, 1980, Essay 4: 78]. = The Sheriff and the Bank Robber (Brand, 1984: 18), The Murderous Nephew (Chisholm, 1966: 29-30 and Brand, 1984: 17-18), Of B's and Bees (Mele, 1987: 56).

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."

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 "bottom-up" picture that assumes an underlying structure of events and causation.

What *stit* abstracts away from: the difference between an agentive effect and any 'causal' way in which it is obtained. Problems of mental causation are simply 'bypassed'.

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

Does *stit* explain agency better than Davidson's theory?

- Does *stit* explain how (primary) reasons, that is, beliefs and intentions determine specific effects?
- Belnap: "Leave the mind out!"
- Is the mind already accounted for by modelling acting as the necessitation of effects?
  "the stabbing of Ceasar" ≠ "the killing of Ceasar" reveals a difference in intention?
- My standpoint: to get a better understanding of how *stit* theory can model agency we have to make mental modalities explicit in *stit*.

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

Aristotle on responsibility

Aristotle: There are two components to responsibility:

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

Aristotle on responsibility

Aristotle: There are two components to responsibility:

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

This leads to two possible excuses:

- being forced to do what you did (coercion: something else forced/caused your action)
- ignorance of what you did (unknowingly doing)

The traveller in the desert [McLaughlin, 1925]

Two enemies independently intent to kill a person travelling through the desert.

In the night the first enemy poisons the water in the victim's canteen.

Right after that the second enemy, not knowing about the poison, empties the canteen.

The next day the person is found dead and the official cause of death is 'dehydration'.

Questions

- who is to blame?
- who dit it?
- who intended the bad outcome?
- who ‘knew’ there would be a bad outcome?
- Is there symmetry between forward and backward looking responsibility?
- difference between moral and legal?

Questions

- who is to blame?
- who dit it?
- who intended the bad outcome?
- who 'knew' there would be a bad outcome?
- Is there symmetry between forward and backward looking responsibility?
- difference between moral and legal?

Judea Pearl (IJCAI '99): the second enemy is the 'actual cause' and the concept of 'actual cause' is the basis on which a theory of responsibility must be built [Halpern, Pearl, Chockler, etc.].

Turning to switches

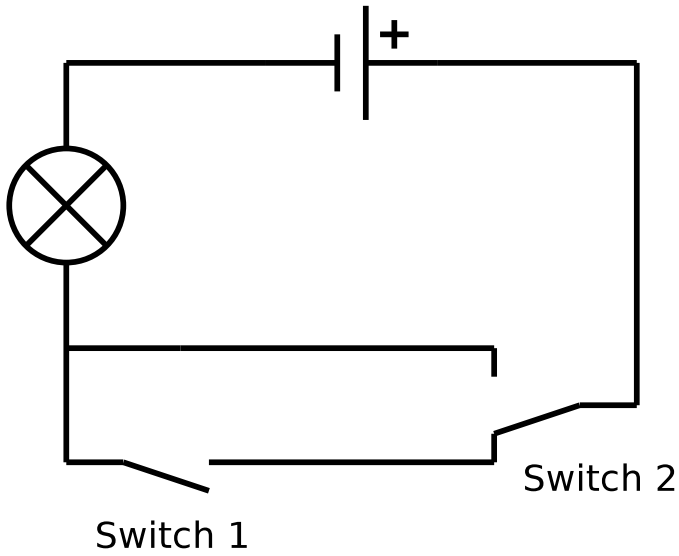

Figure: figure 10.1 from [Pearl 2000], but with the switches in the 'starting' position, and with the names of the switches switched

switch 1 $\approx$ enemy 1; switch 2 $\approx$ enemy 2

Concerning the situation where both switches are no longer in the starting position, Pearl writes: "*Switch 2 (and not switch 1) is perceived to be causing the light, though neither is necessary.*"

Outline

Artificial Intelligence: the three main approaches

Symbolic AI	**Sub-symbolic AI**
Top-down intelligence Modelling AI-concepts	Bottom-up intelligence Optimisation of Algorithms
Logic Agent programming (Inductive) logic programming Planning Emotion modelling	Artificial neural networks Genetic algorithms Decision tree algorithms Search and heuristics

Probabilistic AI
decision and game theory (PO)MDPs Bayesian networks Reinforcement learning

An agent architecture (for symbolic approaches)

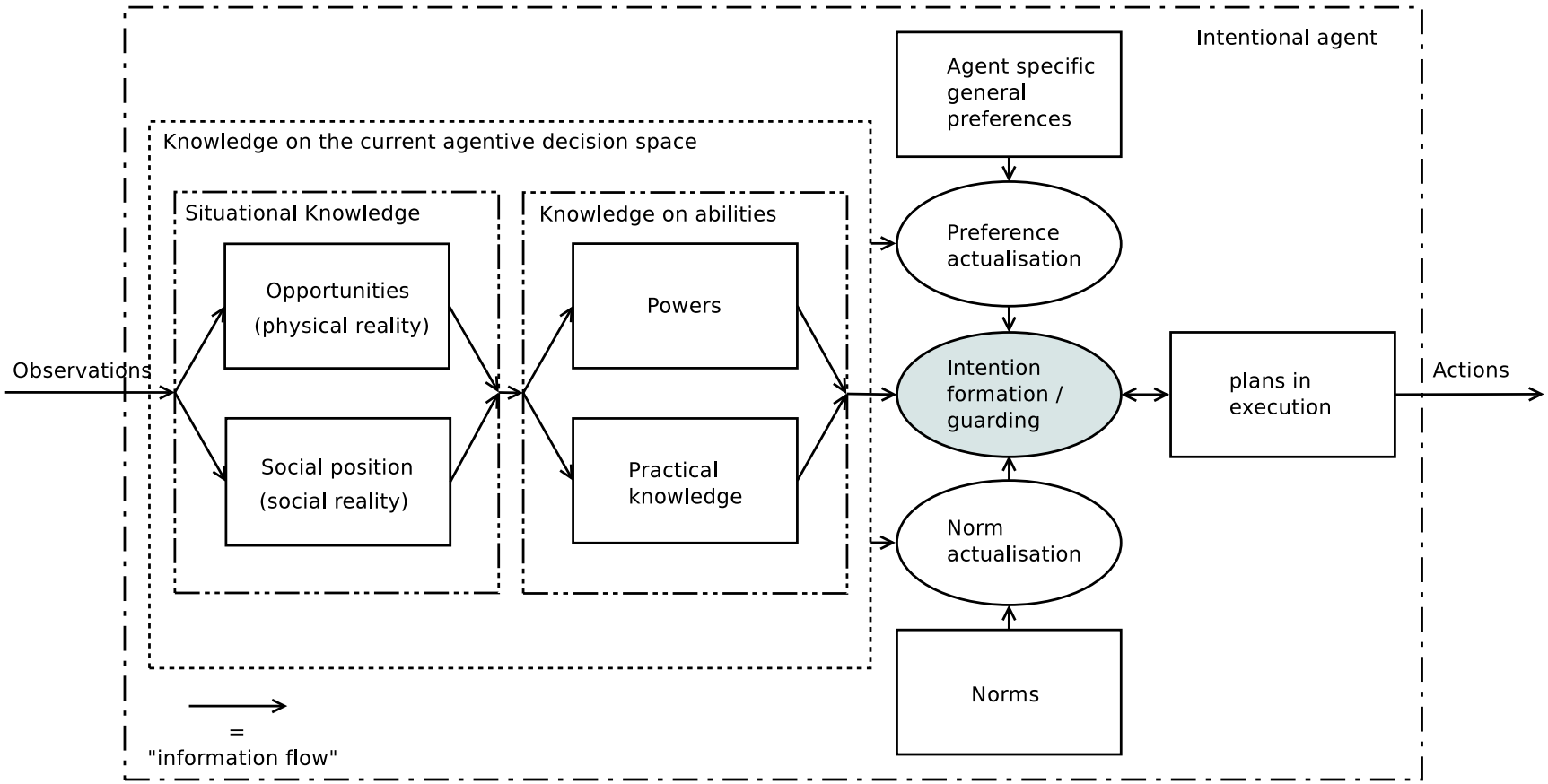

Figure: A pre-formal conceptual model of intentional agency

AI and forward and backward looking responsibility

(1) Responsibility checking = 'outside-the-machine' point of view:

- backwards looking responsibility,

(2) Machine ethics = engineering / 'inside-the-machine' point of view:

- forward looking responsibility,

Personhood for AIs

Not many grant personhood to AIs in exactly the same way as they do to humans. However:

- many AI students do
- some philosophers do

Note: people are extremely reluctant to grant moral patiency to AIs, even if they want to grant them moral agency.

Ripken, 2009: "legal personality can be given to just about any object if it is deemed to serve the ends of justice."

European Parliament, 2016

Motion concerning civil law rules on Robotics (paragraph 59f):

".. creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently."

Robotics Open Letter, 2018

"From a technical perspective, this statement offers many bias based on an overvaluation of the actual capabilities of even the most advanced robots, a superficial understanding of unpredictability and self-learning capacities and a robot perception distorted by Science-Fiction and a few recent sensational press announcements. From an ethical and legal perspective, creating a legal personality for a robot is inappropriate whatever the legal status model."

Agents in AI

The definition of an agent from (symbolic) AI:
(1) autonomous, (2) goal directed, (3) knowledge-based.

AI people say: we are not interested in the question if our agents are ‘proper’ agents, it is all about behaviour. And if attribution of agentive properties helps to get there, attribution is all we need.

Artificial agency?

- Machine functions and cognitive functions will merge.

  Neuro implants, neural interfacing, neuroprostheses, etc.
- If AIs cannot exhibit agency, it does not follow that we cannot formalize agency or reason about agency.

Questions

Recall the El Paso mass shooting example. Do you think responsibility attributions will become more or less difficult with an AI doing the shooting?

The responsibility gap for subsymbolic AIs

"*The responsibility gap: Ascribing responsibility for the actions of learning automata*" by Andreas Matthias, in Ethics and Information Technology 6: 175-183, 2004.

- Warns against the prospect of there being a responsibility gap in sub-symbolic learning machines.
- In my opinion, Matthias wrongly identifies learning with sub-symbolic AI techniques.
- Matthias is one of the philosophers being sceptic about formal and symbolic methods, so he feels there is real danger ahead.

Our (my) view

- We need formalisations of responsibility in order to build / check / control responsible intelligent systems[1].
- The alternative: teach machines to be responsible like we teach our children..
  Is problematic, because:
  - might not work if they are not on an equal footing with us.
  - we will get to that point only gradually, if ever.
  - allowing machines to make mistakes should not be combined with granting them a great deal of causal powers[2].
  - controlling the direction of learning.

[1] And we need pre-formal conceptual models of responsible agency to build a formal framework.

[2] as in algotrading

MIT's moral machine project

http://moralmachine.mit.edu

Paper in Nature appeared in October.

Conclusions: moral preferences differ across regions. Dutch people turn out to be less forgiving towards pedestrians that violate laws!

But, does this have anything to do with morality?

Quotes on how responsibility relates to freedom

"You may believe that you are responsible for what you do, but not for what you think. The truth is that you are responsible for what you think, because it is only at this level that you can exercise choice. What you do comes from what you think." - Marianne Williamson

"Man is condemned to be free; because once thrown into the world, he is responsible for everything he does. It is up to you to give meaning to life." - Jean-Paul Sartre

"Most people do not really want freedom, because freedom involves responsibility, and most people are frightened of responsibility." - Sigmund Freud

Thanks

Thanks!