Responsible Intelligent Systems

Jan Broersen

Department of Philosophy and Religious Studies
Utrecht University, The Netherlands

Master Course on Philosophy of AI
Utrecht, March 18th, 2020

Outline

1. Responsible AI
2. Backward looking / checking
   - The desert traveller
3. Forward looking / deontic reasoning
   - Norms / Reasoning with moral rules
   - Contrary to duty reasoning
4. Learning morality

Responsibility and AI

- We need Intelligent Systems to behave Responsibly (self-driving cars, autonomous trading, military drones, recommender systems, etc.).
- If AI is about making and executing the *good (rational, optimal, intelligent)* choices, then responsible AI is about making and executing the *right* choices and being answerable for it.
- Related area(s) of interest: the social context of AI, adoption of AI, AI and law, AI and moral agency and patiency, etc.

Three angles on responsibility

Responsibility as in:

1. who did it? (who was responsible for that action or result?) backward looking, causality, etc.
2. who ought to do that? (who is responsible for it?) forward looking, norms, decisions
3. who takes / avoids / should be assigned responsibility? strategy, dynamics, social blame games

In this lecture we only take the first two angles.

AI problems: moral verification versus moral decision making

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

- backwards looking responsibility,
- liability,
- causality,
- logics of agency (stit logics), etc.

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

- forward looking responsibility,
- representation of ethical knowledge,
- deontic reasoning, etc.

Outline

1. Responsible AI
2. Backward looking / checking
   - The desert traveller
3. Forward looking / deontic reasoning
   - Norms / Reasoning with moral rules
   - Contrary to duty reasoning
4. Learning morality

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'. Who is responsible for the death of the traveler?

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'. Who is responsible for the death of the traveler?*

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.].

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 to collective responsibility?
3. How do we trace back to individual / shared / partial responsibility?
4. Where to stop back-tracing?

Six categories of responsibility for action

Involvement type Description level	Passive: allowing to happen + ability to prevent	Active: seeing to it + ability to refrain
Causal	(6) causal omission	(5) causal contribution
Informational	(4) conscious omission	(3) conscious action
Motivational	(2) intentional omission	(1) intentional action

Table: A responsibility matrix: six categories of responsibility

Pearl’s theory on causation (causal responsibility)

- There are two relevant[1] possible causes: (1) the poisoning event, and (2) the emptying event. But, neither satisfies the "but for" test.
- Add a concept of ‘sustaining’ cause:
  - poisoning the water does not ‘sustain’ anything, since its effects are eliminated by the emptying event
  - emptying the canteen ‘sustains’ that (1) no poison intake takes place, (2) no water is in the canteen with death as a result.
- Apply the standard counter-factual test (the but-for test) to the situation where all non-sustaining causal processes are removed: get rid of the poisoning event ⇒ the test succeeds (if no emptying, no dehydration).

[1] That the agent decided to start the journey in the morning, is not ‘relevant’..

More intuitions in favour of Pearl

I see two more (wrong) reasons to go with Pearl:

- the official ‘cause of death’ was dehydration, and the second enemy pursued death *by* dehydration
- the second enemy was the *last* to interfere in the process; his inervention was more proximate

Pearl comes with what he claims to be an equivalent example that according to him underlines his argument. But In my view it only complicates things further.

Turning to switches

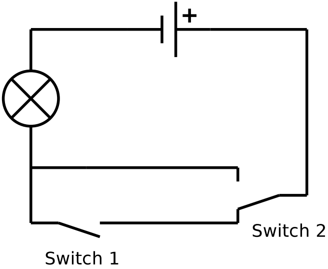

Figure: figure 10.1 from [Pearl 2000], but with the switches in the ‘starting’ position, and with the names of the switches switched

switch 1 ≈ enemy 1; switch 2 ≈ 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.*"

Back to Pearl’s theory

- There are two relevant possible causes: switch 1 and switch 2.
- Sustaining causes:
  - switching switch 1 to the on position does not ‘sustain’ anything, since switching switch 2 to the on position redirects the current
  - switching switch 2 to the on position ‘sustains’ that (1) switch 1 plays no role, (2) the lamp burns.
- Apply the counter-factual test (the but-for test) to the situation where all non-sustaining causal processes are removed: get rid of switch 1 ⇒ the test succeeds (no switching of switch 2, no electricity for the lamp).

Arguing against Pearl

Two reasons *not* to point to enemy 2 as the one responsible for the death:

- the responsibility question was about the *death as such*, not about *death by a certain means*
- the fact that the event of enemy 2 was more ‘proximate’ is irrelevant if the only thing in his power was to change the *way* of dying, but not the *dying as such*.
- taking causal processes out of their context (removing non-sustaining processes) seems arbitrary: the context *did* actually sustain ‘traveller killed’.

The theory of Braham and van Hees

Braham, M. and Hees, M. van (2009). *Degrees of Causation*. Erkenntnis 73: 323-344.

Braham, M. and Hees, M. van (2011). *Responsibility Voids*. Philosophical Quarterly 61: 6-15.

Braham, M. and Hees, M. van (2012). *An Anatomy of Moral Responsibility*. Mind 121: 601-634.

Braham and van Hees’ points of departure

Responsibility for an outcome requires:

- Agency (not dealt with by B & vH, but they seem to adopt ‘agent causation’)
- Causal production (defines how an agent contributes to a joint outcome)
- ‘Reasonable’ alternatives (defines avoidance potential for a joint outcome)

Voids: joint outcomes no agent is responsible for

Braham and van Hees' central definitions

(1) Responsibility for a joint outcome *p* is defined as:

- Contribution: relative to the choices of other agents, agent *i*'s choice is NESS[2] for the joint outcome *p*
- Avoidance: *i* has an alternative choice it believes has a higher avoidance potential

(2) The *level* of causal contribution (not important here)

- agent *i*'s NESS-instances / all relevant NESS-instances

B & vH do no not make the step to 'levels of responsibility'.

[2] NESS = Necessary Element of a Sufficient Set

The avoidance potential (1/2)

Since B & vH look back in time, they know whether strategy/choice s_i was NESS for φ, but if it was NESS, that was still a matter of luck, since *i* did not necessarily know what the others would do.

Responsibility requires the previous existence of an alternative that had a higher potential for *not* being NESS for the current outcome.

- Each agent is endowed with a state dependent probability distribution over the combined choices of other agents.
- Probabilities reflect the acting agent's beliefs (about choices of other agents)

The avoidance potential (2/2)

- Agent i's φ-avoidance potential for a performed[3] strategy/choice s_i = the chance that s_i would *not* be NESS for φ.
- Responsibility: if i performed s_i with a φ-avoidance potential of x, then there was an s'_i with a φ-avoidance potential of y such that $y > x$

[3] Again: B & vH always look backwards in time.

Objections against B & vH’s approach (1/4)

1 Contribution is objective, avoidance is subjective

- This creates the possibility of ‘moral luck’[4]
- Moral luck = luck is a factor in ones moral responsibility (dual: ‘moral misfortune’)
- Nagel and Williams argue for the existence of moral luck
- But, does moral luck exist? Legal luck: yes. But moral luck?

[4] Imagine you *minimised* your avoidance potential (= you ‘tried’) but still you were not NESS. You tried to kill but you were lucky because the gun failed.

Objections against B & vH’s approach (2-4/4)

2 Contribution ‘necessitates’, while avoidance is probabilistic ⇒ assymmetry

3 Avoidance modeled as absence of contribution, which ≠ contribution to avoidance

4 Counter-intuitive in case of overdetermination and/or extensive-form scenarios

Applying B & vH's theory to the desert example

We need to go from extensive-form to normal form

Both enemy 1 and 2 perform a choice that by itself is already NESS for the proposition 'traveller killed' ⇒ two minimal sufficient sets, each entirely covered by one of the agents (the outcome does not require joint action)

Level of causal contribution of each enemy: 0.5

Conflicts with Pearl's proposal.

Aristotle on responsibility

Aristotle: There are two components to responsibility:

- being the one who (actually) controlled a certain outcome
- knowing what you did when you did it

Aristotle on responsibility

Aristotle: There are two components to responsibility:

- being the one who (actually) controlled a certain outcome
- knowing what you did when you did it

This leads to two possible excuses:

- lack of control on what you did
- ignorance about what it is you did

How does this assign responsibility in the traveller's case?

Outline

1. Responsible AI
2. Backward looking / checking
   - The desert traveller
3. Forward looking / deontic reasoning
   - Norms / Reasoning with moral rules
   - Contrary to duty reasoning
4. Learning morality

What ethical theories are relevant?

AI Ethics = ordering actions along the right-wrong dimension[5]

(1) on the basis of a right-wrong order of consequences ⇒ act consequentialism
(2) on the basis of 'absolute' rules ⇒ deontology (Kantianism: universal rules / laws)
(3) on the basis of character ⇒virtue ethics (Aristotle, Elizabeth Anscombe[6])

(1') on the basis of utility (e.g., the common good) ⇒ utilitarianism (Bentham / Mill)
(1") order behavioural rules (not actions) on the basis of their right or wrong consequences ⇒ rule consequentialism[7]

[5] the good-bad dimension is personal and conditional

[6] Doctrine of double effect

[7] Interesting link with reinforcement learning, dynamic scripting, etc.

Kantianism (1/2)

- Moral acting requires freedom which follows from autonomy. Consequentialism fails this requirement because the consequences determine what has to be done. Instead, what has to be done should be determined by reasons.
- Acts are right only in case they are performed for reasons of good will[8]. Reasons that link to things other then 'duties', such as desires, wants, etc, do not qualify as moral reasons.

[8] no other reasons, like pity, pleasure, intelligence, or whatever, qualify

Kantianism (2/2)

- Acting with a good will = acting in accordance[9] with the universal laws of wrong and right
- The universal laws of wrong and right (maxims) are the ones that would generate the most optimal world[10] if obeyed by everybody

Telling a serial killer the route to his next victim is right, because lying is not an option in the morally ideal world...[11]

[9] not clear if this takes a subjective or objective stance

[10] not clear what determines the underlying ordering here

[11] Kant did not accept conditionalisation and therefore, no contrary to duties

What ethical theory to use for machine ethics?

Kantianism seems too ‘idealised’ (black and white, no contrary to duty reasoning, etc.) and is out of reach anyway (logical formalisations are too computationally intensive)

Act consequentialism requires us to do what Kant would not allow us to do: reducing ‘right versus wrong’ to ‘good versus bad’. However, it connects comfortably with standard decision theory..

Rule consequentialism / proportionalism as a workable compromise between act consequentialism and deontology is a natural and feasible theory from the perspective of Artificial Intelligence.

Example of rule consequentialism: Asimov’s three rules

Asimov’s three laws for robots in his 1942 short story "Runaround":

1. Robots need to protect humans
2. Robots need to obey humans, if not in conflict with 1
3. Robots need to protect themselves, if not in conflict with 1 or 2

Example of:
(1) a ‘contrary to duty’ (CTD) specification[12],
(2) a prioritised moral code.

[12] From 1 & 2 it follows that robots need to obey, if not, they need to protect?

How do we specify an order?

Moral code = the implicit or explicit specification of a right-wrong order

How do we make a code explicit[13]? ⇒ rules..

Types of moral rules:

- Permission: if φ you may *do* α
- Prohibition: if φ you may not *do* α
- Obligation: if φ you should *do* α

problems: interactions between types of rules? interactions between rules of the same type? priorities/weights? conflicts? indifference?

[13] for instance to our children

How do we represent rules / conditionals?

Standard Deontic Logic (Von Wright 1951)

Take a modal KD box: $\Box$

Define:
It ought to be[14] that alpha: $\mathit{OB}\alpha = \Box\alpha$
it is forbidden that alpha: $\mathit{FB}\alpha = \Box\neg\alpha$
it is permitted that alpha: $\mathit{PB}\alpha = \Diamond\alpha$

moral rules / conditionals: $\varphi \rightarrow \mathit{OB}\alpha$ or $\mathit{OB}(\varphi \rightarrow \alpha)$
(narrow scope versus wide scope)?

[14] Von Wright later said that he meant α to be an action description

Chisholm's 'paradox'

Chisholm's scenario is formed by the following moral code:

- (1) It ought to be that Jones goes to assist his neighbours.
- (2) It ought to be that if Jones goes, then he tells them he is coming.
- (3) If Jones doesn't go, then he ought not tell them he is coming.
- (4) Jones doesn't go.

(note: 2 and 3 suggest a choice between narrow and wide scope. Extra complication arises if this suggestion is not there.)

Chisholm's 'paradox'

- (1) It ought to be that Jones goes to assist his neighbours.

  (1') OBg.
- (2) It ought to be that if Jones goes, then he tells them he is coming.

  (2') $OB(g \rightarrow t)$.
- (3) If Jones doesn't go, then he ought not tell them he is coming.

  (3') $\neg g \rightarrow OB\neg t$.
- (4) Jones doesn't go.

  (4') $\neg g$.

Solving Chisholm's 'paradox'

Chisholm's conflict is a clash between:
'deontic detachment' (from OBg and $OB(g \to t)$ derive $OB(t)$) and
'factual detachment' (from $\neg g$ and $\neg g \to OB\neg t$ derive $OB\neg t$).

Just get rid of all deontic detachment? Make the logic of *OB* weaker? (drop the K axiom and go to neighborhood semantics)

add action and / or time?

go non-monotonic? Yes..

Summary: the challanges for machine ethics, as I see it

- Design a good system for rule-based reasoning with moral/legal codes (BOID, input-output logic, prioritised default rules, argumentation, etc.): still al lot of work needs to be done.
- Investigate the link with the ethical theory of rule consequentialism
- Think about moral code learning (finding moral rules (like in inductive logic programming: ILP), updating priorities of moral rules on the basis of outcomes, etc.)
- Think about formalizing general moral concepts, like 'fairness', 'equality', 'proportionality' etc.

A simple single agent architecture for responsible agency

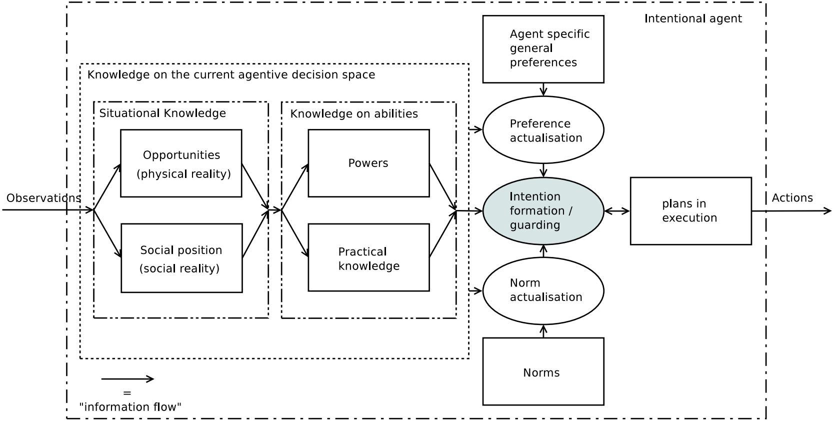

Figure: A pre-formal conceptual model of intentional agency

Outline

1. Responsible AI
2. Backward looking / checking
   - The desert traveller
3. Forward looking / deontic reasoning
   - Norms / Reasoning with moral rules
   - Contrary to duty reasoning
4. **Learning morality**

The responsibility gap

"*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.

Artificial Intelligence: the three main approaches

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

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

AI learning methods

- Logic (formalistic / symbolic) methods:
    - Inductive logic programming (learning logical rules)
    - Dynamic Epistemic Logic (information update)
    - Belief revision, norm revision (AGM theory)
- Probabilistic methods:
    - Bayesian learning / updating (closer to symbolic methods)
    - Reinforcement / Q-learning (closer to black-box methods)
- Black-box methods:
    - Artificial neural networks
    - Evolutionary learning
    - NEAT (a combination of the two)
    - unsupervised learning (self organisation, clustering, pattern recognition, data-mining, etc.)

The role of learning across AI approaches

- Symbolic approaches ⇒ concepts are explicitly modelled, learning is in terms of these concepts
- Black-box approaches ⇒ no concepts are modelled, learning itself does all the work
- Probabilistic approaches: the middle ground

Problems with learning and responsibility

(1) The high price of mistakes:

- (supervised) learning, by definition, will only occur if first there are mistakes..[15]
- We cannot always afford mistakes if it comes to responsibilities delegated to machines[16]

(2) Learning wrong things:

- Difficult to steer learning in the 'right' direction[17]

(3) Forgetting about side effects in reinforcement learning

[15] This is also true for versions of DEL: the ones dealing with beliefs

[16] there is however a difference between on-line mistakes and off-line mistakes

[17] which is especially true for unsupervised learning techniques like clustering techniques, self-organising maps, etc.

Are learning AIs really a problem?

Yes, they are, because:

- inherently unpredictable (black-box approaches) and therefore less under the control of designers (for the same reason, game designers dislike black-box approaches)
- example: should the causal power to kill (as in an autonomous military drone) be in the hands of a ‘black-box’ or reinforcement learning method?

No, they are not, because:

- We still have a great deal of control and can build in mechanisms interfering with the black-box methods.
- We might try to make (probabilistic) models of black-box methods aimed at calculating risks.

Our (my) view

- We need formalisations of responsibility in order to build / check / control responsible intelligent systems[18].
- 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[19].
  - controlling the direction of learning.

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

[19] as in algotrading

MIT’s moral machine project

http://moralmachine.mit.edu

New, follow up paper out, where they double down on their findings.

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?
(consequentialism + relativism)

Thanks

Thanks!

Moral responsibility versus legal responsibility

Issue \ Type	Legal Responsibility	Moral Responsibility
If no act alternative	relieved	no relief[20]
Luck	possible	impossible[21]
Violation conditions	objective	subjective
Need to know?	yes	no?
For groups?	yes	no?
Formalisations	rule-based	modal logic

Table: Legal versus Moral

[20] Frankfurt's objection against the principle of alternative possibilities

[21] This goes against Nagel and Williams