Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible

Daniel Tigard
Ph.D. Candidate
Department of Philosophy

Tulane University

Symposium on Regulating Consciousness: 27th April 2018
Newcastle University

Some background on Responsibility in A.I.

Assumption: Responsibility is a necessary condition for just war, morally permissible medical practice, etc.

- Matthias (2004): The use of machines (learning automata, operating with unfixed rules) creates a "<u>responsibility gap</u>, which cannot be bridged by traditional concepts of responsibility…"
- Sparrow (2007): possible loci of responsibility [for war crimes]…
  - Programmer?
  - Commanding Officer (or Operator)?
  - Machine itself?

    NONE!
- Thus, morally impermissible to deploy autonomous machines in war, medical practice, etc.

Guiding Questions & Agenda

Can we hold machines responsible (e.g. for harms in warfare or medical practice)?

Yes!

The question, then, is *HOW?*

(1) Artificial Moral Agency

(2) How Agency does and doesn’t matter

(3) Pluralism in Moral Responsibility

(4) Locating Responsibility in Learning Automata

(1) Artificial Moral Agents (AMAs)

- Allen & Wallach (2009): AMAs = artificially intelligent (AI) systems within the circle of moral agents
- **Moral agency** is *very complex*, traditionally entails…
  - Capacities for deliberation, free-will (“control condition”)
  - Capacities for understanding, say, right from wrong (“epistemic condition”)
- Each of the conditions for moral agency presupposes **consciousness** (Himma 2009)
- AI cannot (yet?) have consciousness. Thus, can’t be “moral agent.”
- Still, AI can have *functional morality*: “its architecture & mechanism allow it to do many of the same tasks” (Allen & Wallach)

(2) How Agency Does & Doesn’t Matter

- P.F. Strawson (1962): responsibility is a function of being susceptible to “natural human reactions to the good or ill will or indifference of others towards us”
- Reversal of traditional concepts of responsibility
  - *Holding* is conceptually prior to *Being* responsible
- Agency is secondary. Facts of responsibility are determined by our practices (‘reactive attitudes’, blaming/praising, etc.)
- But agency matters: we don’t hold *anyone/anything* responsible!
- Holding others responsible is *<u>not</u>* a singular/unified enterprise…

(3) Pluralism in Moral Responsibility

- Watson's "Two Faces"
  - Blame: to *attribute* something (a moral fault) to an agent

    "Aretiac" face – concerns one's character ("deep self")
  - Blame: *holding* someone accountable

    "Accountability" face – concerns our practices (rewarding, punishing, etc.)
- Shoemaker's Tripartite Theory
  - Attributability: attributing decision/action (fault) to one's character

    Requires agent's capacity for cares/commitments
  - Accountability: holding one accountable (for poor "regard")

    Requires agent's capacity for empathy
  - Answerability: demanding reasons/justifications for one's judgment

    Requires agent's capacity for deliberative decision-making

(4) Locating Responsibility in Learning Automata

- Hold automata "answerable" – demand reasons/justifications
  - AI can consider multitude of competing reasons (better than us!) and can respond to demands for reasons by citing goal-directed programming &/or learned causal processes
- "Attribute" decisions/actions to automatas' "self" (murky!)
  - Given unique environments & processes learned, something *like* a unique "character" can be developed over time (although not proper cares/commitments)
- Hold automata to "account" – reward/punish to encourage/discourage
  - Consequential justifications can be "understood" and can be effective, despite ineffectiveness of desert-based accounts

(1) Demand reasons → (2) Attribute action → (3) Hold to account

Conclusion: Responsibility “Gap” Revisited

- The responsibility gap created by learning automata “cannot be bridged by traditional concepts of responsibility…”
- Perhaps! But rather than abandoning the project of trying to bridge that gap (& rather than relying on artificial conceptions of agency), we can adapt our existing practices of holding others responsible.
- What’s artificial is not the moral agency; it’s our application of otherwise natural responsibility ascriptions (i.e. *artificial moral responsibility*).
- Plausible theoretical foundation for moral responsibility in AI, can provide basis for development & application of important legal norms.

Thank you!

References

Allen, C and W Wallach (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford UP.

Himma, K (2009). "Artificial agency, consciousness, and the criteria for moral agency." *Ethics and Information Technology* 11: 19–29.

Matthias, A (2004). "The responsibility gap: Ascribing responsibility for actions of learning automata." *Ethics and Information Technology* 6: 175–183.

Shoemaker, D (2015). *Responsibility from the Margins*. Oxford UP.

Sparrow, R (2007). "Killer Robots." *Journal of Applied Philosophy* 24: 62–77.

Strawson, PF (1962). "Freedom and Resentment." *Proceedings of the British Academy* 48: 1–25.

Watson (1996). "Two Faces of Responsibility." *Philosophical Topics* 24: 227–248.

Daniel Tigard
Ph.D. Candidate
Department of Philosophy
dwtigard@gmail.com