The Moral Responsibility Gap and the Increasing Autonomy of Systems

The Central Thesis

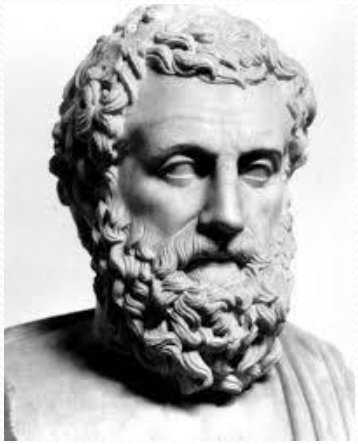
- As systems become more autonomous, engineers and users have less control over the systems' actions.

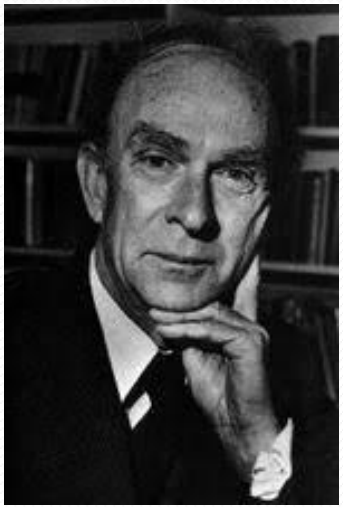
- Andreas Matthias, 2004: *"Nobody has enough control over the machine's actions to assume [moral] responsibility for them."* This is the 'responsibility gap'.

- This paper goes deeper into the problem by elucidating two dimensions of the moral responsibility gap: the causal and the epistemic.

Moral Responsibility

- To be answerable for an action or event in such a way that we might be praised or blamed for it.
- Aristotle: we are only morally responsible for our voluntary actions. If we could not have done otherwise, or if we do not know what we are doing, then we are not morally responsible for our action.
- P.F. Strawson: assigning moral responsibility is an inherently social practice.
- Fischer and Ravizza: a control theory of moral responsibility.
- Prospective and retrospective moral responsibility





The Moral Responsibility Gap

- **Causal Control:**

  What causal influence do engineers have over the actions that the systems take?

- **Epistemic Control:**

  How far can engineers know, understand, explain how the system reaches a decision of the actions that it takes?

Moral responsibility and non-autonomous systems

- **Causal Control:**
  Robust control (but not complete control)

- **Epistemic Control:**
  Robust control

Moral responsibility and autonomous systems

- **Causal control:**
  Control becoming less robust

- **Epistemic control:**
  Control becoming **substantially** less robust

Specific questions

- C. How can we account for emergent behaviour?
- C/E. If we can only assure system-types, how should we navigate the unpredictability of system-tokens?
- E. To what level of confidence should we be able to distinguish a learning error from a design error?
- E. Is it possible to distinguish between a learning error and a novel solution to a problem?
- E. What standards should we use to reconcile trade-offs, e.g. explainability vs. effectiveness?

Thank you!

Engineer

Autonomous system