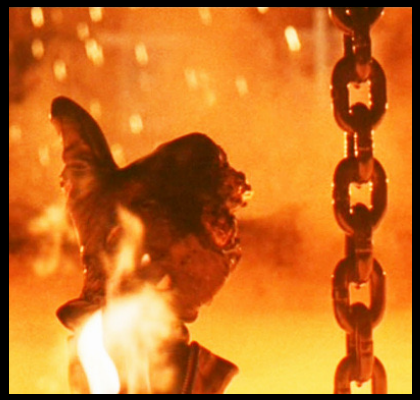
UNIVERSITY OF CAMBRIDGE

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

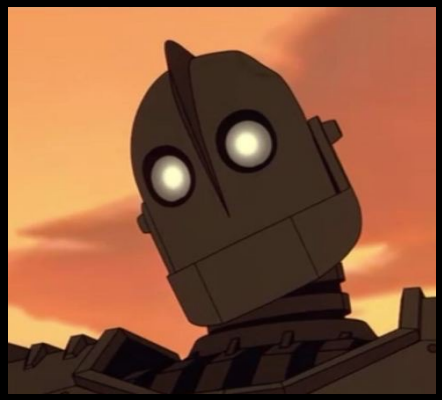
Pathways to artificial moral patiency

Henry Shevlin









Novel beings workshop, 19th September 2019

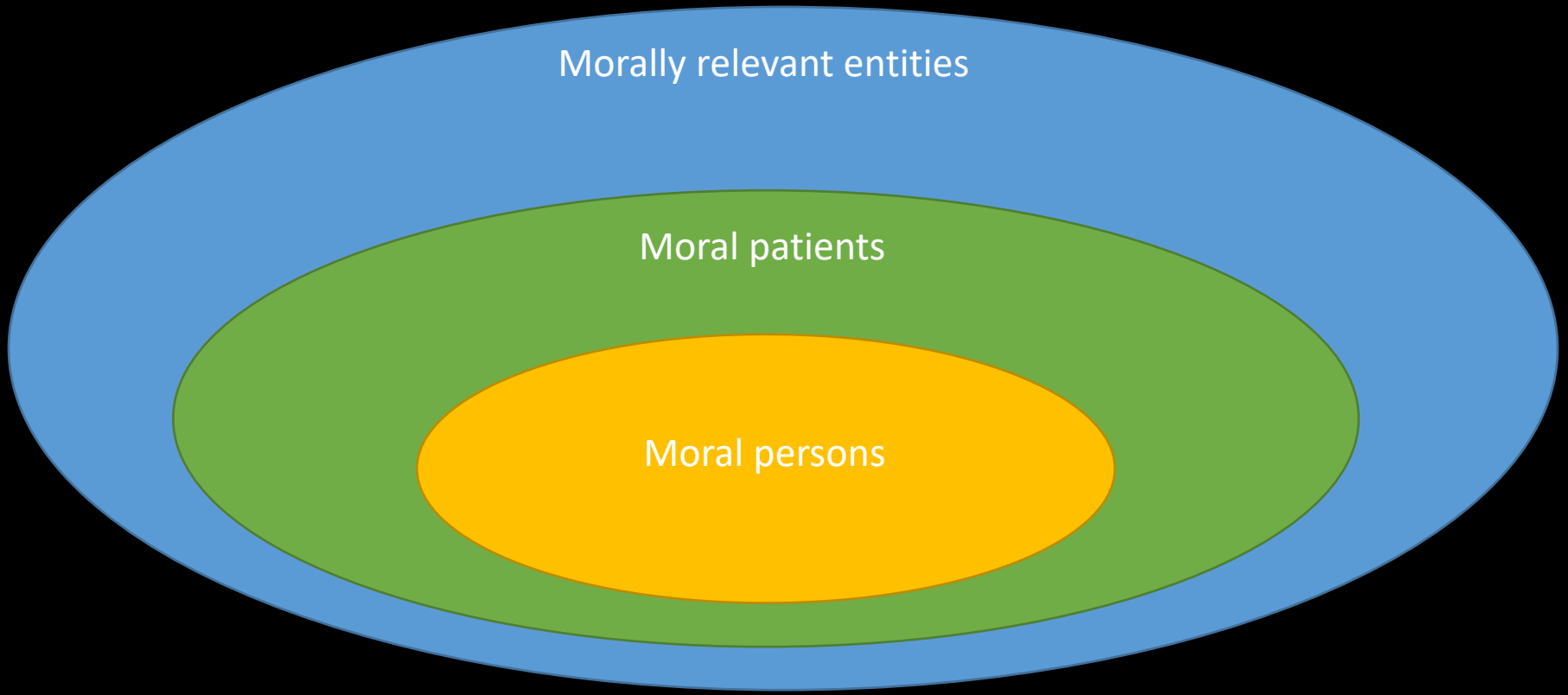
Newcastle University

What is a moral patient?

- Broadly: a being with morally relevant interests – a being that can be benefitted or harmed in ways that are potentially relevant for ethical decision-making.
- One kind of moral patient: a being capable of suffering.
- But perhaps there are other ways of being a moral patient – we recognize harms that don't involve suffering (e.g., via having your wishes violated, autonomy compromised).
- Note also that there are morally relevant things that are not moral patient patients – e.g., mountains, sacred objects, great works of art.

What is a moral patient?

- I would also suggest that while all persons are moral patients, the reverse isn't true; we might recognize a fish as a moral patient without calling it a person.



Artificial moral patients

- An artificial moral patient (AMP) is thus an artificial being that has interests of moral significance.
- AMPs are just a possibility at this stage, so why bother even thinking about them?
- Already we fail to act to secure the interests of many beings we recognize as moral patients, both people and animals.
- However, many important reasons why we should start thinking about this now!

Artificial moral patients

- First, note that we seem to have no difficulty empathizing with artificial beings.

- The question is whether that empathy is appropriate. We want to make sure we don't empathise inappropriately and waste resources.

Inappropriate empathy

Artificial moral patients

- Note that we seem to have no difficulty empathizing with artificial beings.
- The question is whether that empathy is appropriate. We want to make sure we don't empathise inappropriately and waste resources.
- We also want to make sure we're not blind to moral patiency in non-obvious forms (cf. fish welfare).
- Human history is full of cases where we failed to extend empathy appropriately – if we believe in an 'expanding circle' of moral concern then AI an obvious next target.

Some special considerations for AMPs

- Proliferation
- Inscrutability
- Ease of intervention
- Danger of bad practices becoming entrenched
- Extremes of valence (s-risks)
- But most importantly… it helps us build frameworks for animal moral patiency

Pathway 1: self-disclosure

- But how could we identify an AMP when it came into being?
- We might naively expect from science fiction that AMPs would simply tell us they had interests and feelings.
- However, this could be misleading in light of anthropomorphizing tendencies: remember ELIZA.
- We also shouldn't assume that the first AMPs will be capable of human language or introspection; perhaps more like simple animals than artificial people.

Pathway 2: Suffering

- Alternatively, we might use scientific investigation to establish the presence of suffering. This would *ipso facto* make an AI an AMP by most people's lights.
- Problem 1: how do we identify negatively valenced states? Already tricky for animals; how much harder for AIs.
- Problem 2: how could we tell whether these states were conscious?
- Maybe if cognitive science has a *really* good couple of decades, this could work, but I'm not holding my breath.

Pathway 3: Preferences

- Maybe we could attribute moral patiency to an AI on the basis of its having robust preferences (c.f. Dawkins).
- Problem: what counts as robust preferences? Even Roombas engage in (superficially?) goal-directed behavior.
- No clear dividing line between more appetitive behavior and preferences in 'thick' sense.
- (Also, do unconscious preferences really count?)
- Again, a target for cognitive science, but unclear we'll even know when we've located relevant psychological kind.

Pathway 4: Biological analogy

- Proposal: any AI that is relevantly cognitively similar to an animal which we already consider a moral patient (legally, socially) should be afforded protection.
- Worry 1: existing animal welfare law and attitudes are a fucking disaster.
- Yes, but we have independent reason to fix this.
- Worry 2: how do we identify relevant cognitive parameters for comparison?
- Not as hard as it sounds – experiment and reflection.

Why we don’t have artificial moral patients (yet)

- Familiar list of reasons why AI differs dramatically from ‘BI’.
- Most of these concern general intelligence, i.e., robust and flexible production of behavior.
- AIs are brittle; animals are robust and resilient.
- AIs are hidebound (transfer learning, catastrophic forgetting); animals are flexible.
- Suggests underlying differences in cognitive architecture between animals and current AI that mean biological analogy doesn’t come close to applying (yet).

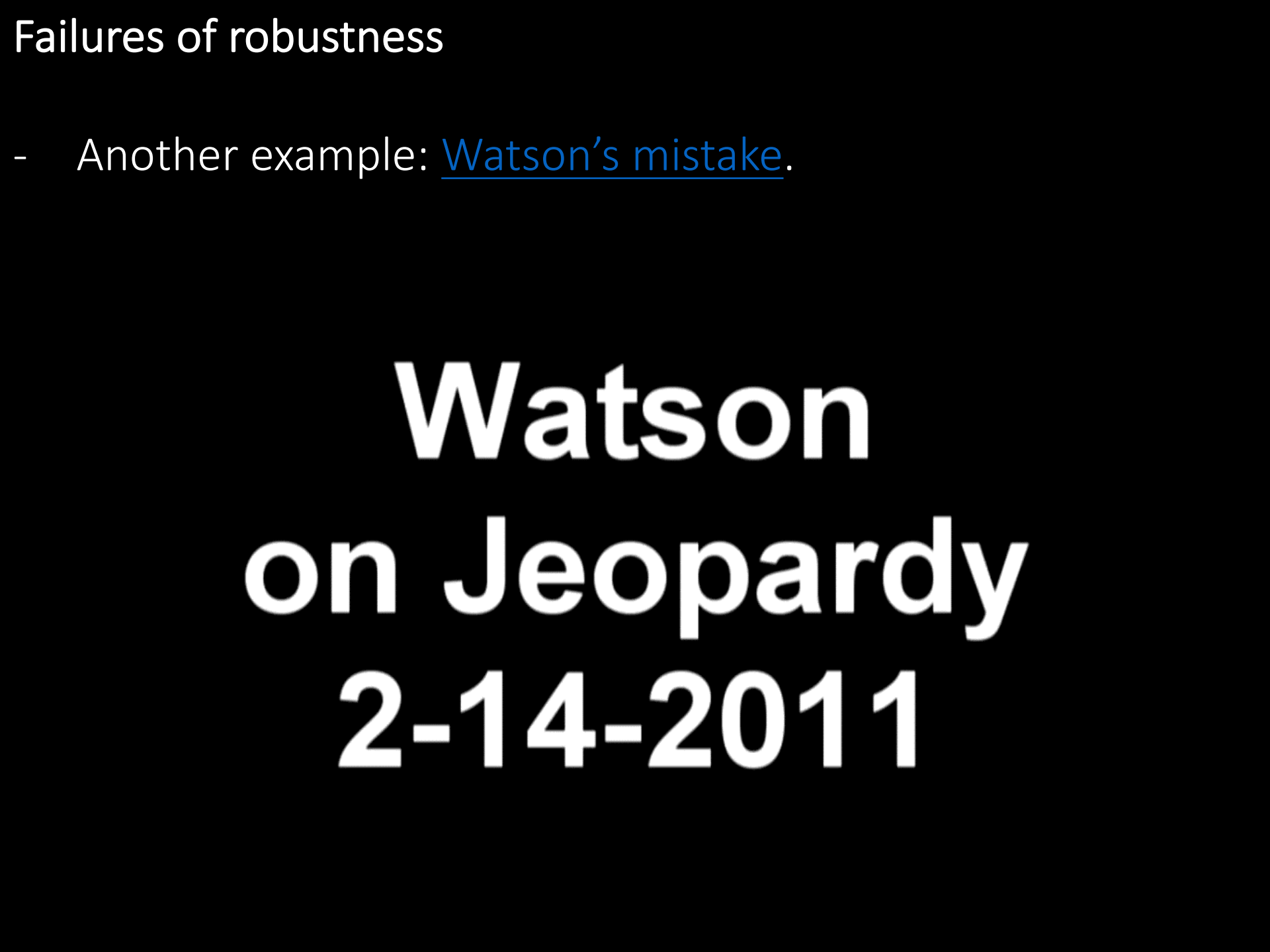
Examples of failures of robustness

- Living things are quite robust: aside from moths and lamps, they don't have many simple 'failure modes'.

- By contrast, robots/AIs are glitchy and vulnerable.

Failures of robustness

- Another example: Watson's mistake.



Failures of flexibility in AIs

- This difference exemplified by Frostbite challenge.

- Although AI outperforms humans, it lacks flexibility.

- Similar, often learns slowly via many examples.

- Try this character challenge...