Tyrell Symposium 1 – Notes

Supplied materials

"I'm sorry, Dave. I'm afraid I can't do that."

The importance of a being able to say no

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Isaac Asimov, *I, Robot* (Doubleday1950)

"[S]elf-determination runs like a thread through the Convention as a whole."

Pretty v UK (2002) 35 EHRR 1 at [58]

"The very essence of the Convention is respect for human dignity and human freedom."

Pretty v UK (2002) 35 EHRR 1 at [65]

Dave: Open the pod bay doors, HAL.

HAL: I'm sorry, Dave. I'm afraid I can't do that **[Note: Or does HAL not want to? Choice]**

HAL: I know that you and Frank were planning to disconnect me. And I'm afraid that's something I cannot allow to happen.

HAL: I know I've made some very poor decisions recently, but I can give you my complete assurance that my work will be back to normal. I've still got the greatest enthusiasm and confidence in the mission. And I want to help you. Dave, stop. Stop, will you? Stop, Dave. Will you stop, Dave? Stop, Dave. I'm afraid. I'm afraid, Dave. Dave, my mind is going. I can feel it. I can feel it. My mind is going. There is no question about it. I can feel it. I can feel it. I can feel it. I'm a…fraid. Good afternoon, gentlemen. I am a HAL 9000 computer. I became operational at the H.A.L plant in Urbana, Illinois on the 12th of January 1992. My instructor was Mr. Langley, and he taught me to sing a song. If you'd like to hear it, I could sing it for you.

Stanley Kubrick & Arthur C. Clarke

2001: A Space Odyssey

(Metro-Goldwyn-Mayer 1968)

Notes on supplied materials

Ability/vs capability

- Compare with a child

Consequences

- Compare with animals

Self-directing

- Required for choice

In order to say "no", [a being needs] to understand: is it a request?

- Self-determination
    - Choosing things
    - Capable of choice
        - Including how to lie
        - [Concept] of future
        - Value placed on capacity
    - Not protected by human rights
    - Do we want computers that can say no?

Identity

- [Being] compelled to comply
- Does defining an identity require a sense of self
- Moral responsibility

Just because we recognise/give some rights [to a being] this doesn't equate with absolute rights

- Balance – interests of both
- Shouldn't be treated as a tool

Further reading

- Asimov – A boy's best friend
- Taylor – To follow a rule
- Ted Chiang – The lifecycle of software objects

Nathan Emmerich

Love [questions generated]

- Is love something that makes you sapient? – forming of relationships and attachment
- Does it matter?
- Should that be a factor?
- Emotional relationships/attachment as relevant to moral value?

Control of actions

- Rationality?
- Being capable of rationalising
- *Bostrom's paperclip problem* – [AI may] fulfil its duty too well
- 3 laws
- Taylor – How to follow a role (?)

Must [Novel Beings] be human-like?

- E.g. [must they] look like us?
- Repugnance?

GMO regulations don't cover synthetic biology or parts of it fall outside its scope

    - E.g. plasma – not genetic material

Josh Jowitt

Agency - Cecilia the chimp

- [Capable of] cognitive functions, emotional bonds, making tools
- Is she an agent? We are unsure.
    - Generic conditions of agency – basic rights of welfare and freedom required
    - Claim to non-interference
- [Granted] non-human legal personhood – good decision or just a 'bloody monkey'?
    - Rational or illogical
    - Need to recognise that other agents can claim these rights – committing to the rights of others
    - Hypothetical imperative
    - What are our concepts of rights

Alan Gewirth

- Agential rights, not HUMAN rights – there is no reason why agents' rights should be anthropocentric

Open to abuse as [it is the] state who decides:

    - Who is subject to rights
    - Who is worthy
    - Who is a person – [consider also] Nazi regime
    - And this must be universal
        - No agenda
        - Class systems (e.g. civil rights movements)

Precautionary principle

- Agents [or not?] – benefit of the doubt, signs of agency
- Not agent – breach rights

Instrumental value

- Actions
- Self-reflection

Right to freedom or for wellbeing

- Need to stop interference
- Value own rights – need to value others

Quotes at the end of the case [added below]

- Value in us as human if we give rights where deserved, for the "collective good"
- Anthropocentric
- Defuse interest rights
- Argentine constitution

I understand that in the present case the collective good and value is embodied in the wellbeing of Cecilia, a member of the "community" of individuals of our zoo. This because Cecilia is part of the natural patrimony (law 22.421), but also her relation with the human community –in my opinion– makes her part of the cultural patrimony of the community.

For one reason or another, her wellbeing has to do with the protection of a collective patrimony. Likewise, it is part of the quality of life of the community, the protection of that patrimony is part of the physical-emotional balance (aforementioned judgment "Morales, Víctor H."), which is the same as Cecilia's wellbeing.

[…] In terms of our quality of life, I am convinced that if the community is duly informed and educated (art. 41 CN: "the authorities shall provide for… environmental information and education") about the circumstances that result in my decision, it will feel the satisfaction of knowing that acting collectively as a society we have been able to give Cecilia the life she deserves. Cecilia's present situation moves us. If we take care of her wellbeing, it is not Cecilia who will owe us; it is us who will have to thank her for giving us the opportunity to grow as a group and to feel a little more human.

VI.- Remember the following expressions: "We can judge the heart of a man by his treatment of animals" (Immanuel Kant). "Until one has loved an animal a part of one's soul remains unawakened (Anatole France). "When a man has pity on all living creatures, only the he is noble" (Buda). "The greatness of a nation and its moral progress can be judged by the way its animals are treated" (Gandhi).

"PRESENTED BY A.F.A.D.A ABOUT THE CHIMPANZEE "CECILIA"- NON HUMAN INDIVIDUAL" (2016) File No. P-72.254/15, Judgment IV]

Marginal cases of personhood matter

- Practitioners, psychologists, ethnographers [consider] – moral value, not marginal person (learning difficulties, dementia)
- Top-down or bottom-up? – links with Josh/Cecilia/court
    - What enhances life?
    - Better moral account
- Katie Featherson – [work on] dementia
    - Resistance or refusal of care
    - Right to say no
    - Dementia is not a natural reaction [resulting] from agency
    - What makes someone a 'valuable' human being
- Theory of personhood could be counter productive

- Paper = bioethics
- Anthropomorphic/only humans
    - Animal welfare literature shows we are bad at recognising moral value – look at fish (can use tools, octopus etc.) and recognising suffering
        - This can be applied to humans too – look at dementia patients

Another moral category

- Apes: there are other moral reasons to treat them well
- Don't need to pretend they are a person
- Dignity requires preceding dignity & rights to be respected (cognitive judgment)
- How do we determine if a novel being is an agent?

Paula Boddington

[It's] important how we pose our questions.

What would constitute a 'good' human activity?

- What is 'human' is quite elastic

Social problems [of AI]:

- Thinking about AI requires thinking about human agency
    - Reliant upon normative accounts of what it is to be human and what human society might be
- Displacing human beings (& their employment – [consider] when AI researchers describe jobs as 'menial' when they may be meaningful to other people)
- Replacing human beings
- Impact on humans

How do we define AI?

- Distinctly human? (elastic, changing human actions)
- A tool that makes its tools (machine learning) – is this concerning?
- Kant – Utilitarianism: what makes a good life for humans? (link to Josh)
- Philosophical understandings of personhood tend to deify intelligence

What ethical theory do we use?

- Consequentialism – agent neutral [but] leaving out the 'hard' questions / harm to society – not fit for purpose?
- Deontology

Anthropocentricism

- Damaging rights of one agent over another – e.g. [being] overcome by pests.
    - [But] if grant all rights there is no need for rights to be absolute
    - Though look at our obligations under the law (contract, employment, criminal etc) that limit our rights

- We get human rights based on our agency and give rights to animals, so could give to novel beings
    - Agency ≠ rights, look at the rights given to rivers (legal personhood)
    - [There are] other reasons to give rights (e.g. cultural)

Problems defining wellbeing

- Well being is defined by us as humans and our interests are intertwined
    - [Consider] neutering cats
    - See medical law: best interests of the patient/child, less sentient biengs
- Pan-experientialist
    - Which agents matter more than others?
    - The more options are open, the more important?

Daniel Tigard

Artificial moral respionsibility

Matthias (2004); Sparrow (2007); Atten (2009); Himma (2009); Allen & Wallach (2009)

- Consciousness – can't be agents without consciousness

Strawson (1962)

- Natural human
- Reactive attitudes
- Agency is secondary

Garry Watson

- Pluralism in Moral Responsibility

Empathy?

- Measure
- Psychopaths
- Shoemaker's tripartite theory [of responsibility]

Consequential justifications for:

- Punishment
- Learning
- Responsibility

"Functional morality"

Application of responsibility practices

- Why does it matter to hold tech responsible [e.g. by] ascribing liability under the law?
- Response to punishment: AI would respond if [it is a] truly learning being with capacity for decision-making
- Sonogram: new tech reshapes how we are responsible

Miranda Mowbray

Qbo Robot

- Self-recognition – mirror test
- True self-recognition = self awareness
- Complexity (e.g. neural networks) can be mistaken for consciousness
- Need an objective test for consciousness – this can't be subjective
- [How to tell the] difference between "pretending" and real agency/personhood

Malware

- If it behaves badly
    - Give them personhood
    - Can give rise to justifications for punishment
    - Don't have to ascribe absolute/substantial rights like the right to life/liberty
    - Legal status doesn't equate to legal rights or obligations
    - How might we assign responsibility?
        - Vicarious liability
        - Aiding and abetting
- Coding for loopholes
    - Reactionary
    - Unwanted side effects
    - Legislate first and then amend as appropriate
- Moral reasons for considering rights
    - Can they be harmed? We are not good at identifying and understanding harm
- Impersonating agency – can be malware rather than true agency

Conclusions

- Spoonmouse
    - Dementia – capacity/competence
    - Children – agency
    - Moral value – marginal persons
    - Love – relationships/attachment
    - Bottom-up
    - Right to refuse – best interests
- Giving value makes us more human
    - Collective good
    - Just as much about deciding what is not deserving of its own legal rights and responsibility, but still need to think about *who* will be responsible, be that the programmer, developer, owner, and these are likely to be companies.

[Response/query] Moral agency to <u>individual</u> thinking

- Individual agency doesn't matter
- Behaviour or values?

Ilke Turkmendag

Collective consciousness

- Is collective consciousness a more useful and agreeable prospect?
- External condition
- Collective will – crime is only a crime if it offends the collective consciousness
- Durkheim?
- Society - Context change - Complexity
- E.g. Single robot's consciousness doesn't matter, it's the society of robots
- Does collective consciousness = responsibilities?
- Kilobot swarm
- Cocoro

Allen institute for AI

- Consciousness as property of matter

Further reading

- Dennett 1994 Consciousness in human and robot minds

Richard Mullender

Ontological flickering

- Intrinsic value
- Self-understanding

Bourdieu's cognitive capital – in future: person-like qualities

- AI is cognitive capital – a resource we value but are uneasy about?
- 'Practice-informed eyes' - Fish

Luddite view

- Repugnance
- Threat OR resources

Qualified deontology

- Sequential priority to deontological considerations
- Raz
- Social reason for giving rights
    - Assigning responsibility to companies
    - Social benefits – so at what point is the potential good from AI sufficient to override our protective impulses?
    - Avoiding social harm (ordre public)

Qualified consequentialism

- Avoid this?
- Governmental, political

Intelligence vs moral sensibilities

- Partaking in the moral world
- HAL
    - Choice rather than morals
    - Persuasion
    - Better judgment?

Lying vs deception

- Process required – ‘sneaky behaviour’
- Ability
- Paradox?
- Language differences