Robots, AI, and philosophy

Arto Laitinen

arto.laitinen@uta.fi

Professor in Social Philosophy

University of Tampere

- Are fears and hopes concerning robotics adequate? Overestimated in the short run and underestimated in the long run?
- Developments in robotics quickly lead to deep philosophical questions about
    - **Responsibility:** who is responsible for what autonomous robots do?
    - **The nature of persons:** are robots moral persons now, or in the future?
    - **Ethics:** what moral codes should be programmed into robots (e.g. robot cars)?
    - **Social philosophy:** should we rethink the basic categories of a modern society after the AI revolution?

Could robots be (e-)persons?

- Consider the nuanced parliamentary motion concerning Europe-wide legislative framework for robotics and AI:
- for purposes of legislation on liability and responsibility, a new legal status of

  > “electronic persons, with specific rights and obligations, including that of making good any damage they may cause”.

- What would it mean to think that in addition to human manufacturers, programmers, trainers, owners or users and so on, the robot itself would have rights and obligations?

Why would we want robots to be persons?

- Autonomous robots seem to create a ”responsibility gap”: due to self-learning and autonomous functioning, what they do does not seem to be anyone else’s responsibility
    - Manufacturers, programmers, trainers (the length of training period -> degree of responsibility), owners or users and so on
- But
- Unless the robots themselves are fit to be held responsible, there is a responsibility gap: a planned, created responsibility ”vacuum” where no-one is responsible.
    - Analogous to forces of nature in that sense: no-one is responsible for a volcano erupting. Would we want to create such ”volcanoes” to our midst?

Robots, as we currently know them, do not meet conditions of personhood.

- 1) Persons are *rational beings*.
  - AI very good in tasks that require computable intelligence; and they can learn
- 2) Persons are beings to which states of consciousness are attributed, or to which psychological or mental or *intentional predicates* are ascribed.
  - Robots can have functional equivalents of beliefs and desires;
  - good memory ("access consciousness"); but no feel, *phenomenal* conscioucness,
- 3) Whether something counts as a person depends in some way on an attitude taken toward it, a *stance adopted* with respect to it.
  4) The object toward which this personal stance is taken must be capable of *reciprocating* in some way.
  - Robots do not have moral sentiments: respect, love, esteem, moral indignation;

- 5) Persons must be capable of *verbal communication*.
  6) Persons are distinguishable from other entities by being *conscious* in some special way: there is a way in which we are conscious in which no other species is conscious. Sometimes this is identified as *self*-consciousness of one sort or another.
  - Again, robots do not have even less special consciousness

  (The list from Dennett: ”Conditions of Personhood”)

What would it take to build a robot person?

- It would need to have the physical basis of experiential consciousness, it would e.g. be able to feel pain
- And it would need to be a process that has ”stakes” or ”vital interests” in the world, like living beings.
- Arguably, that would require certain type of hardware, and not just software. Consciousness seems to be associated with certain physical systems and not others. The key need not be whether it is carbon-based or silicon-based, but what kinds of connections take place.
- Why is it that, say, the cerebral cortex produces consciousness but the cerebellum does not, or what is it that underlies the difference between wakefulness, sleep, coma, and anaesthesia? It is not fully known. The best attempts are very controversial and even speculative, such as the “integrated information theory”.

What would it take to build a robot person?

- Further, robots as we know them, do not lead lives, that is, engage in “far-from equilibrium dynamic processes”, whose end amounts to the thing ceasing to exist, dying.
- A candle flame is a self-organizing, self-maintaining process, and bacteria (and other living beings) are *recursively* self-maintaining processes. They have a normative stake in the world: they must maintain the process on pain of ceasing to exist.
- Presumably that would have to be so with robots, before it would make sense to think they have rights, or that we owe them anything

What new legal category for robots?

- Supposing then that the category of electronic persons is misleading, how should we categorize them? One counter-proposal is that robots should be slaves (Joanna Bryson), or less polemically, *e-servants* instead of e-persons.
- We build them for our purposes, no great mystery: they are what we build them to be.
- They will have the capacity to learn and function autonomously to the extent that we build them that way.

Do responsibility gaps matter?

- Suppose then that robots will not be persons; moral agents responsible for their own doings.
    - Neither is any other piece of technology we live with.
- Assess acceptable levels of risk & design insurance schemes:
    - Strict liability: duty to compensate harms even when not one's fault
    - Cf. traffic legislation:
        - registration of individual machines,
        - compulsory insurance schemes (compensate when faulty)
        - General insurance scheme, which compensates when the faulty party not identified, or cannot compensate etc.

Self-deriving vehicles and runaway trolleys

- In Immanuel Kant’s (1724-1804) theory, everyone ought to be treated as ends and not mere means, each human being has an infinite worth instead of a measurable value: price. No-one is to be sacrificed in the name of the general good.
- What if the situation is such that someone must die and we must choose between one and five, as in the so called “trolley cases” or “rescue cases”?
- These highly theoretical discussions have now surfaced in the context of robot cars, which perhaps need to be programmed in advance to make a decision in such cases.
- These cases are abstractions to help discuss whether in addition to all other morally relevant features, the numbers of persons involved make a moral difference at all: or is the value of each person literally infinite?

Runaway trolley cases: one lesson

- If a runaway trolley is going to hit and kill five people, is it morally permissible for a bystander to turn the switch, so that the trolley will only hit and kill one person? More lives would be saved.
- If someone is standing on a bridge, is it permissible to push a button that will drop that person to the rails, and stop the trolley before it hits and kills the five?
  - No, that would be to treat that person on the bridge as a ”mere means”. The causal role makes a moral difference.
- If a runaway trolley with five passangers is heading towards destruction, is it permissible for a bystander to turn the switch, so that the trolley will only hit and kill one person instead of the five passangers?
  - The difference between passanger vs not may make a moral difference; here we are not discussing numbers alone. Not a ”pure” trolley case.
- One lesson: saving five acceptable if not treating as a ”mere means”

Rescue cases and acceptability to all: an additional lesson?

- If a rescue boat can save either five (island A) or one (island B), what should it do?
- Is there a principle you would consider acceptable, if you didn’t know whether you will be the rescuer, one of the five on the first island A, or the lone person on the second island B?
  - rescue 5? Unfair to the B-person?
  - rescue 1? Unfair to the A-persons?
  - lottery between A and B? Unfair to the A-persons?
  - weighted lottery between A and B? All have a fair chance!
- With random number generators, would be feasible with robot cars
- Weighted lottery seems most acceptable *for the designer, manufacturer, programmer,* who in a sense avoid making the decision: leaves it to ”fate”.

What purposes should robotics serve?

- What, then, are the purposes that we should design robots to serve, in transportation, service robotics, healthcare, education, low-resource communities, public safety and security, employment and workplace, and in entertainment (to mention the areas that the Stanford University AI100 report of 2016 focuses on)?
- The question is fundamentally the same as what goals should we design our institutions, such as schools and hospitals, to serve, or ultimately, how should we design human societies. To promote well-being, equality, autonomy, justice and so on – the very principles that political philosophy has always debated.