Robotethiek en Robo-ethiek

Gert-Jan C. Lokhorst

TU Delft

Utrecht 13 feb 2013

TU Delft

Motivatie

- Toenemend gebruik van robots, o.a. op het slagveld (“drones”)—met duizenden doden als gevolg.
- Maar ook civiele toepassingen: huishoudrobots, carebots. Wat te doen als er iets fout gaat? Hoe dat te voorkomen?

Robotethiek

Ethisch redeneren door robots zèlf

TU Delft

Robo-ethiek

Ethisch redeneren door mensen *over* robots

TU Delft

Basis

1. G.J.C. Lokhorst. Computational meta-ethics: Towards the meta-ethical robot. *Minds and machines*, 21: 261–274, 2011.

2. G.J.C. Lokhorst and M.J. van den Hoven. Responsibility for military robots. In P. Lin, K. Abney, and G. A. Bekey, eds., *Robot Ethics: The Ethical and Social Implications of Robotics*, pp. 145–156. The MIT Press, Cambridge, Massachusetts, 2012.

3. G.J.C. Lokhorst and M.J. van den Hoven: Statues and robots on trial: From robot ethics to roboethics, te verschijnen.

Het Prytaneion

Rechtszaken tegen onbezielde voorwerpen. Plato over Magnesia:[1]
If some soulless thing should take away the soul from a human being, except in cases where it's a lightning bolt or some such missile coming from a god, but in the case of any of the other things that may kill someone through his falling upon it or its falling upon him—the next of kin should appoint the closest neighbor to be judge, and thus remove the impious impurity from him self and the whole family. The convicted party they should cast beyond the borders, just as was stated in the case of the class of living things (Plato, Laws 873e–874a).

[1] "De hele geschiedenis van de Westerse filosofie is een serie voetnoten bij Plato."

. . . the Prytaneion. What is this court? If a stone or a piece of wood or iron or something of that kind strikes a man, and one does not know who threw it, but knows and holds that object which caused the death, a lawsuit is brought against such objects in the Prytaneion (Demosthenes, Against Aristocrates 76).

. . . At the Prytaneion: in this court they are tried for homicide whenever a person is known to have been killed but the person who committed the homicide is missing. One delivers the written accusation to the basileus, and the basileus makes an announcement through the herald and forbids this man who killed so-and-so to set foot in holy places and the land of Attica. Also, if an inanimate object falling on someone hits him and kills him, a trial is held for it in this same court and it is cast beyond the frontier (Patmos scholiast, Commentary on Demosthenes, Against Aristocrates 76).

The court at the Prytaneion gives judgment concerning homicides even if their identity is not known; it also gives judgment concerning inanimate objects which have fallen on someone and killed him. This court was presided over by the phylobasileis, whose duty it was to remove beyond the border the inanimate object which had fallen upon the man (Pollux, Onomasticon 8.120).

The Court in the Prytaneum, as it is called, where they try iron and all similar inanimate things, had its origin, I believe, in the following incident. It was when Erechtheus was king of Athens that the ox-slayer first killed an ox at the altar of Zeus Polieus. Leaving the axe where it lay he went out of the land into exile, and the axe was forthwith tried and acquitted, and the trial has been repeated year by year down to the present (Pausanias 1.28.10–11).

If the actual offender is unknown, the writ runs against "the doer of the deed." The King and the tribe-kings also hear the cases in which the guilt rests on inanimate objects and animals (Aristotle, Constitution of Athens 57.4).

Theagenes

Een voorbeeld:
When he [the Olympic victor Theagenes] departed this life, one of those who were his enemies while he lived came every night to the statue of Theagenes and flogged the bronze as though he were ill-treating Theagenes himself. The statue put an end to the outrage by falling on him, but the sons of the dead man prosecuted the statue for murder. So the Thasians dropped the statue to the bottom of the sea, adopting the principle of Draco, who, when he framed for the Athenians laws to deal with homicide, inflicted banishment even on lifeless things, should one of them fall and kill a man.

But in course of time, when the earth yielded no crop to the Thasians, they sent envoys to Delphi, and the god instructed them to receive back the exiles. At this command they received them back, but their restoration brought no remedy of the famine. So for the second time they went to the Pythian priestess, saying that although they had obeyed her instructions the wrath of the gods still abode with them. Whereupon the Pythian priestess replied to them:—"But you have forgotten your great Theagenes." And when they could not think of a contrivance to recover the statue of Theagenes, fishermen, they say, after putting out to sea for a catch of fish caught the statue in their net and brought it back to land. The Thasians set it up in its original position, and are wont to sacrifice to him as to a god (Pausanias 6.11.6–8).

Er waren soms onduidelijkheden. Bv. het geval van een jongen die per ongeluk iemand met een speer doodde. Tegen wie moest er in dit geval een rechtszaak worden begonnen? De jongen? De speer? Beide? Geen van beide?

Uit andere culturen zijn zaken bekend tegen bomen, rotsen, pijlen, zwaarden, (instortende) huizen, boten, afgodsbeelden en (valleien beschadigende) gletschers (Evans 1906, Hyde 1917, Finkelstein 1981).

- Waarom deden de mensen dit? Herhaling voorkomen, de morele orde herstellen, uiten van woede, "achieving closure" ...
- Leeft voort in Amerikaanse rechtspraak, met name rechtszaken waarin schepen worden beschuldigd.

Rechtszaken tegen drones

Drones (unmanned aerial vehicles). Hebben al duizenden tegenstanders of vermeende tegenstanders gedood.

- We zouden *die drones*, of *degenen die ze ontwierpen*, of *maakten*, of *degenen die ze bedienden*, of *daartoe de opdracht gaven*, of *de Amerikaanse overheid*, verantwoordelijk willen stellen voor hun gebruik in de strijd.
- Maar wie precies? Tijd voor een terugkeer van het Prytaneion?

Artificial Ethical Agents

(Artificial) Ethical Agents: 3 soorten (Jim Moor)

1. Implicit ethical agents—gedragen zich in overeenstemming met bepaalde morele regels
2. Explicit ethical agents—houden zich aan bepaalde morele regels die expliciet zijn ingebouwd
3. Full ethical agents—kunnen ethische beslissingen nemen en ethisch rechtvaardigen

"My recommendation is to treat explicit ethical agents as the paradigm example of robot ethics." (Jim Moor)

Explicit ethical agents

Hoe te bouwen? Niemand weet het. De bestaande ethische theorieën schieten te kort.

- Consequentialisme (kijk naar de gevolgen van acties, hun gevolgen, hun gevolgen etc ad inf): geen stop conditie, run-away situatie, computationeel zwart gat
- Utilisme: maximalisatie van geluk. Wat is dat? Hoe meten we dat? Hoe brengen we een robot dat aan zijn verstand?
- Kant: universaliseerbaarheid. Iedereen dient zich volgens regels te gedragen die iedereen acceptabel zou vinden als iedereen ze zou volgen. Gigantisch gedachtenexperiment. Universaliseerbaarheid/generalisatie is niet iets eenduidigs. Vb: joodse onderduiker.

- Prima facie duties: wat als er meerdere conflicterende plichten zijn?
- Aristoteles' deugd-ethiek: wat als er meerdere conflicterende deugden zijn?
- Goddelijke commando theorie: idem ditto; bovendien geven verschillende goden verschillende commando's. Hoe te kiezen?

Kortom

We hebben een “comprehensive model of moral decision making” (Wallach 2010) nodig, maar we hebben er geen. “Scandal of ethics”, “scandal of AI.”

- De wetten schieten ook te kort. Conventies van Genève: non-combatants (degenen die niet meedoen aan de strijd) moeten zoveel mogelijk worden gespaard. Maar wat is een non-combatant? “Gebruik je gezonde verstand.” Hoe moet een robot dat doen?

Extra probleem

Drie soorten robots:

1. Op afstand bestuurde robots
2. Autonome robots
3. Semi-autonome robots

Verantwoordelijkheid voor als er iets fout gaat. In geval 1 en 2 duidelijk. Maar in geval 3?

- Bij rampen met socio-technische systemen is dat ook zelden duidelijk. *Herald of Free Enterprise* (193 doden), ICE Eschede (101 doden). "Many hands problem." Naam van probleem, niet van de oplossing van het probleem. Multicausaliteit.

Implicit Ethical Agents

Vooruitzichten zijn beter. Deugd-ethiek voor robot-deugden. Moed, compassie, vasthoudendheid, etc. Geen probleem met verantwoordelijkheidstoeschrijvingen.

- We moeten wel duidelijk zijn over de eisen die we willen inbouwen.
- Bv. de conventies van Genève: die moeten dan eerst net zo duidelijk worden gemaakt als bv. de regels van het schaken.

Full Ethical Agents on Trial

Mensen voor het gerecht. Mensbeelden:

- "Babbelbox." Implementatie in robots moet kunnen: Franse postmodernistische filosofie generator (Bulhak 1996).
- "Volkspsychologie." Wensen, overtuigingen, intenties, plans. Angst voor straf, hoop op beloning. Toepassen op robots: bizar. "Mechanische deuropener in staking." (Latour 1988)

Conclusies

- Explicit ethical agents zijn een fata morgana.
- Geen autonome robots: de mens (getraind militair personeel dat bekend is met de conventies van Genève e.d.) moet hoe dan ook "in the loop" blijven.
- Geen robot ethiek maar roboethiek, inclusief ethiek voor het ontwerpen van robots.
- De ethiek van de ontwerpers etc. is doorslaggevend: we moeten de robots de waarden meegeven die *ons* aan het hart liggen.