Philosophy, Artificial Intelligence, and Robotics: The ethical challenges

John P. Sullins
Ph.D. 2002, Ph.D in Philosophy, Computers and Cognitive Science, Binghamton University
Sonoma State University Philosophy Department. 2003-
WE SHOULD WANT TO HAVE AI BEHAVE AT OUR BEST NOT COPY OUR WORST
AI AND MORALS: SHOULD WE (CAN WE?) TEACH HUMAN MORALITY TO MACHINE MINDS?
WHAT IS MORAL AGENCY?
MORAL AGENTS

• How ever we chose to define them, moral agents have:
  • Rights
  • Responsibilities
When Is a Robot a Moral Agent?
AI AND ROBOTS ARE IN OUR LIVES
WARFARE
ON THE STREETS
TAKING CARE OF US
WHAT ARE THEY ACTUALLY USED FOR?

If the agents involved in these actions were humans rather than an AI or robot, then some of the situations would easily be seen as moral situations.
When these situations involve robots is the morality:

- **Illusory**—situation is fallaciously seen as moral situation
- **Pseudo Moral**—partially moral situations but lacking something that will make them fully moral situations
- **Real**—these situations may be novel but they are real moral situations that moral theories must address

I will argue for a version of the latter case
MORALITY AND TECHNOLOGIES
• We typically blame the user not the tool when harm is caused by a person using some tool or technological system.

• If a robot is simply a tool then does not the morality of the situation reside fully in the user and/or designer of the robot?
  • Thus the robot is not a moral agent but only an instrument that advances the moral interests of others.
Here the moral praise or blame is harder to determine.

- Both trainer and dog seem to share it. We praise the skill and sacrifice of the trainers and laud the actions of the dog as well.
- An important emotional attachment is formed between all the agents in this situation but the attachment is strongest towards the dog.
  - Is a good guide dog morally praiseworthy?
• Here the moral praise or blame is harder to determine.
  • Both trainer and dog seem to share it. We praise the skill and sacrifice of the trainers and laud the actions of the dog as well.
  • An important emotional attachment is formed between all the agents in this situation but the attachment is strongest towards the dog.
  • Is a good guide dog morally praiseworthy?
CATEGORIES OF ROBOTIC TECHNOLOGIES

• Telerobot
• Autonomous robot
• Remotely controlled with only minimal autonomous decisions made by the machine.
  • Mars rover, rescue-bot, predator drone, SWORD, etc.
PROGRAMMER(S) -- >  
AI/ROBOT/ENVIRONMENT  ➔  USER

• This is a much more interesting problem
• Programmers are not the primary locus of praise and blame
• Could the robot itself serve as the primary locus?
1. Robots are not persons
2. Unlikely to be so anytime soon
3. So we need a sense of moral agency that does not require personhood
PHILOSOPHICAL VIEWS ON THE MORALITY OF ARTIFICIAL AGENTS

• Daniel Dennett
• Selmer Bringsjord
• Joseph Emile Nadeau
• Floridi and Sanders
DENNETT: WHEN HAL KILLS, WHO IS TO BLAME?

- First robot homicide
  - Malfunctioning robot kills autoworker in Japan, 1981
- Mens rea is required
- Robots can fulfill these requirements and thus can be guilty or fully to blame for their actions
- But “Higher order intentionality is a necessary precondition for moral responsibility.”
- Robots with this will be moral agents
• No such thing as an autonomous robot
• Robots have No free will
• PERI is incapable of doing other than what it is programmed to do unless random factors added to program.
• PERI is not a moral agent
• Nevertheless, robots can be programmed to act ethically
• Only androids operating with a theorem prover are ethical agents.
  • Responsibility and culpability determine ethical status and they require free will
  • An action is caused by free will iff it is caused by reasons

• An android operating from a theorem prover or some combination of technology with this ability could have actions caused by reasons
• Androids are then ethical agents
  • Note, humans do not operate on reasons so are not ethical agents!
FLORIDI AND SANDERS

• Provide an argument for believing that one need not be a robust cognitive agent to be a moral agent

• Ascribing the status of moral agent or patient to an entity requires looking at the entity from the proper level of abstraction.

• Artificial agents are moral agents when they are interactive, adaptable and have sufficiently autonomous state changes
THREE CONSIDERATIONS

• To assess the moral status of autonomous robotic technologies

1. Is the robot significantly autonomous?
2. Is the robot’s behavior intentional?
3. Is the behavior based on a ‘belief’ held by the machine?
AUTONOMY

- AI/Robot is not a telerobot
- AI/Robot has functional independent agency
- AI/Robot level of Autonomy contributes to level of moral agency
INTENTIONALITY

- Can we describe the behavior only by ascribing some ‘intention’ to harm or help?
- Are the actions seemingly deliberate and calculated?
- Free will nor robust internationality is not necessary.
‘BELIEFS’ OF RESPONSIBILITY

• Is the behavior describable using standard folk psychological notions of belief?
• The belief does not have to be ‘real’ in any philosophically rigorous way.
TWO CONCLUSIONS

• Robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and beliefs.

• If the robot can be seen as autonomous from many points of view then it is a robust moral agent, possibly approaching or exceeding the moral status of human beings.
MOVING FORWARD ON AI ETHICS
ARTIFICIAL PHRONESIS

• A process used by artificial systems to achieve
  • At least a functional capacity to
    • Identify ethical situations
    • Properly attend to them
    • Practically and efficiently react
    • Justify actions and choices
Phronesis (Ancient Greek: φρόνησις, phronēsis)

- Practical wisdom or intelligence
- Skill used in social interactions
  - Navigate them with proficiency and excellence
  - Most people have this skill
  - Requires practice to master
For Aristotle, this skill is not a matter of simple logical deduction

- When in situation x, then always do action y

- Instead, it is a skillful practice that requires discernment of nuances encountered in social situations
- Real life social situations are just too complex and unique to properly deduce behavior simply form past experience or to have predetermined reaction for
- Life requires artful improvisation
If phronesis is a real ability humans have and it is not something that can be programmed or learned simply form past experience, then the social robot’s lack of that capacity will be an insurmountable barrier to the machine’s ability to enter into meaningful relations with human agents. That would be a fundamental problem that would prevent successful application of social robotics to all but the most trivial and scripted interactions with human agents.
Autonomy vs. Ethical sensitivity diagram with regions labeled:
- High Autonomy, High Ethical sensitivity: Danger!
- Full moral agency
- Artificial Moral Agents
- Functional morality
- Operational morality
- Today's (ro)bots

The diagram indicates a transition from today's (ro)bots to artificial moral agents as ethical sensitivity increases.
For Dewey, phronesis plays a role in all human thought

- Dewey gives us a fine counterargument that will blunt the apriori claims that AMAs are impossible
- But it also adds a complication
  - Phronesis plays a strong roll in scientific and technological reasoning as well
  - Thus, solving the problem of artificial phronesis will be a requirement for developing AI with the ability to generalize its thoughts over multiple topics
- If Phronesis is computational, then this is not bad news. If it is not, then this is a major limitation to AI
• Even our scientific and technological actions are directed at some perceived good
• It follows that the Aristotelian categories of episteme and techné cannot be clearly distinguished from phronesis
  • Since both require that the agent successfully identify a good towards which these actions are directed
• To be successful at this requires
  • sensitivity to the complex nature of the situations the agent finds itself in
  • accurate account of the contingencies and random outcomes that may affect the attainment of a goal
• Those agents that constantly make accurate evaluations of the value of the goals they set, will be more successful in developing their innate capacities for living than those that do not
MOVING FORWARD

• Only a small number of people working on this problem.
• We need to inspire research and experiments alla Shaner’s proposal
• Is Phronesis a type of consciousness or a capacity of conscious agents only?
• Pace Bryson—we must not conflate artificial phronesis with human phronesis
System Sketch

A system that attempts to build on the Artificial ethical agent to bring us closer to an acceptable AMA
THANK YOU OLLI !!!

Philosophy, Artificial Intelligence, and Robotics: The ethical challenges

John P. Sullins
Ph.D. 2002, Ph.D in Philosophy, Computers and Cognitive Science, Binghamton University
Sonoma State University Philosophy Department. 2003-