
Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems

Paper Discussion



Ethical Autonomous Systems

- Near future systems are **moral agents**
 - Driverless cars
 - Medical diagnosis AIs
 - ...

Choices have **ethical consequences**

Paper Focus

- Implicit ethical agents
 - Machines designed to **avoid** unethical outcomes
- Explicit ethical agents
 - Machines that **encode/learn ethics**
 - Determine actions based on those
- Ethical Governance
- Overview of the papers in the special issue

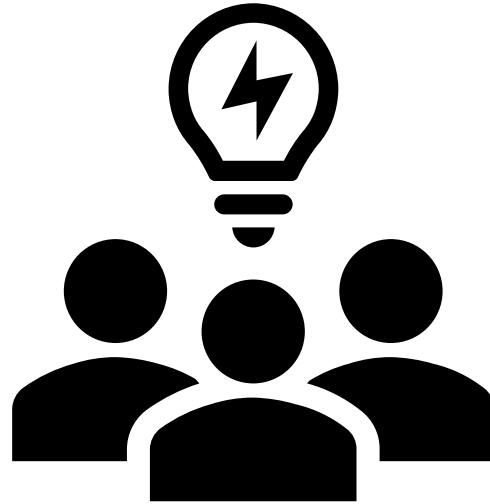
Some standards to be aware of

- IEEE Standards Association document, *Ethically Aligned Design (2017)*
- *IEEE P7000 Model Process for Addressing Ethical Concerns During System Design (2021)*

Reflection



Discuss the paper in your groups



"You are walking on the street and notice a child who is not looking where she's going; you see that she is in imminent danger of walking into a large hole in the pavement. Suppose you act to prevent her falling into the hole."

"You are walking on the street and notice a child who is not looking where she's going; you see that she is in imminent danger of walking into a large hole in the pavement. Suppose you act to prevent her falling into the hole."

Q. Is your action **ethical**? Why?

Q. If you were a robot... Is your action **ethical**? Why?

"Anderson and Anderson propose an alternative 'Ethical Turing Test' (ETT) in which a panel of ethicists are presented with the machine's ethical decisions across a range of application domains. Each ethicist is asked whether they agree or disagree with those decisions – if a significant number are in agreement (i.e., the ethicist would have made the same choice in the same situation) – then the machine is judged to pass the test."

"Anderson and Anderson propose an alternative 'Ethical Turing Test' (ETT) in which a panel of ethicists are presented with the machine's ethical decisions across a range of application domains. Each ethicist is asked whether they agree or disagree with those decisions – if a significant number are in agreement (i.e., the ethicist would have made the same choice in the same situation) – then the machine is judged to pass the test."

Q. What do you think about this test? Discuss pros/cons.

"... **all** intelligent autonomous systems that have the potential to cause harm should be classed as **implicit ethical machines**, and designed using processes of ethically aligned design."

"... **all** intelligent autonomous systems that have the potential to cause harm should be classed as **implicit ethical machines**, and designed using processes of ethically aligned design."

Q. What do you think about this statement? Discuss pros/cons.

Reflection

