CSCI 245 Life, Computers, and Everything

Machine Metaethics

Image source

Consequentialist Ethics

- A way to describe the scenario in the world
- A way to enumerate the possible actions
- A way to predict the consequences of each action
- A method to evaluate each action in terms of goodness

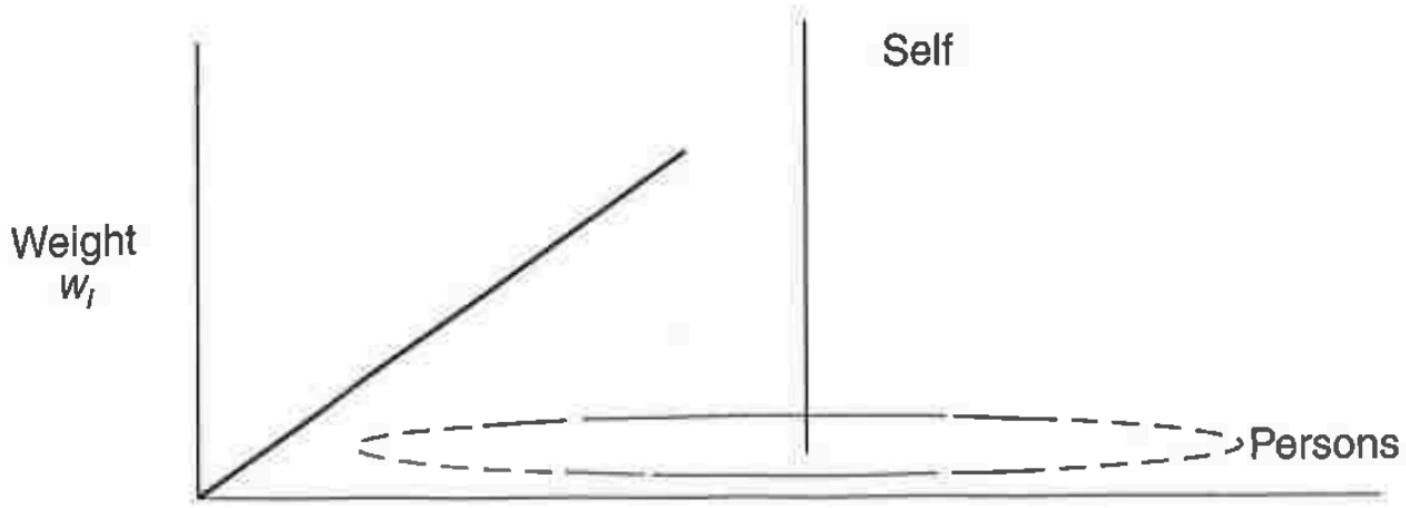
Consequentialist Ethics

$$\sum w_i p_i$$

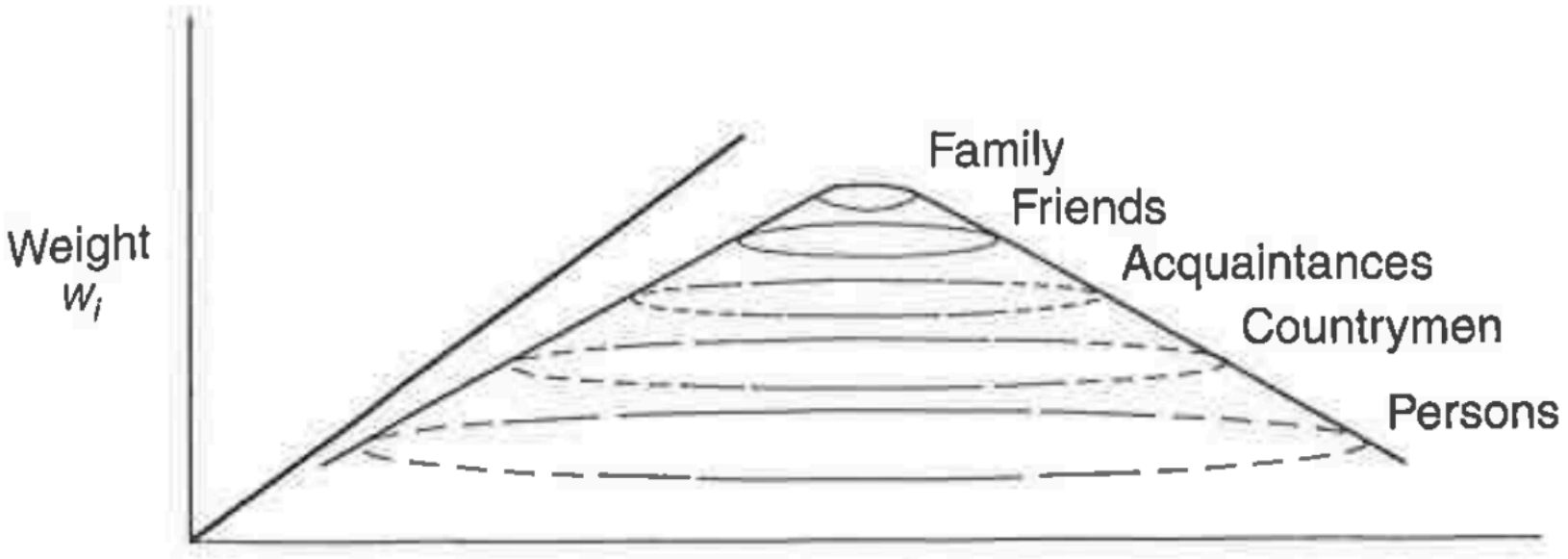
w_i = the weight assigned to person i

p_i = the happiness of person i

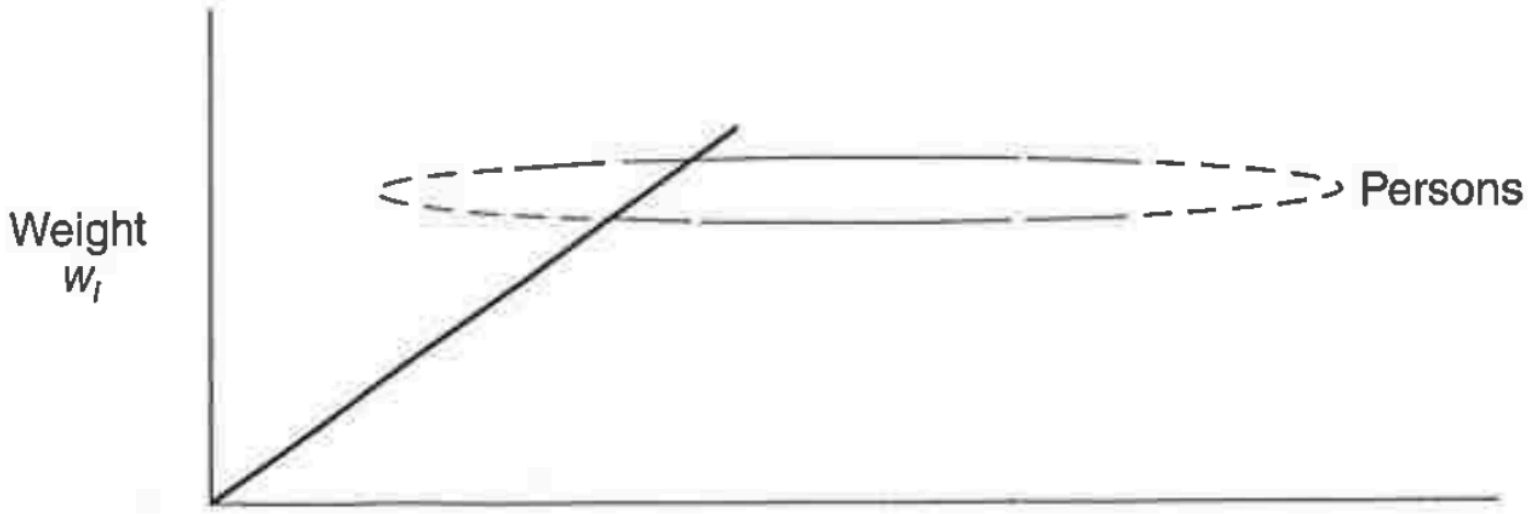
Theories



Theories


Weight
w_i
Family
Friends
Acquaintances
Countrymen
Persons

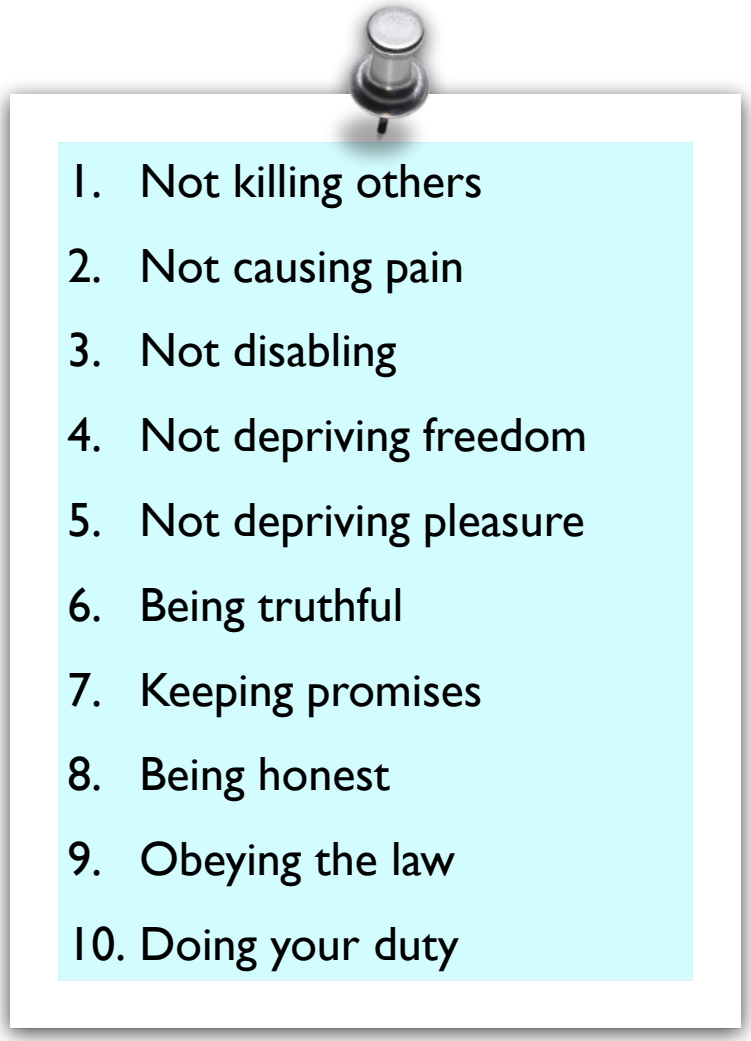
Theories



Theories

- Ethical egoism
- Ethical altruism
- Utilitarianism

Deontological Ethics

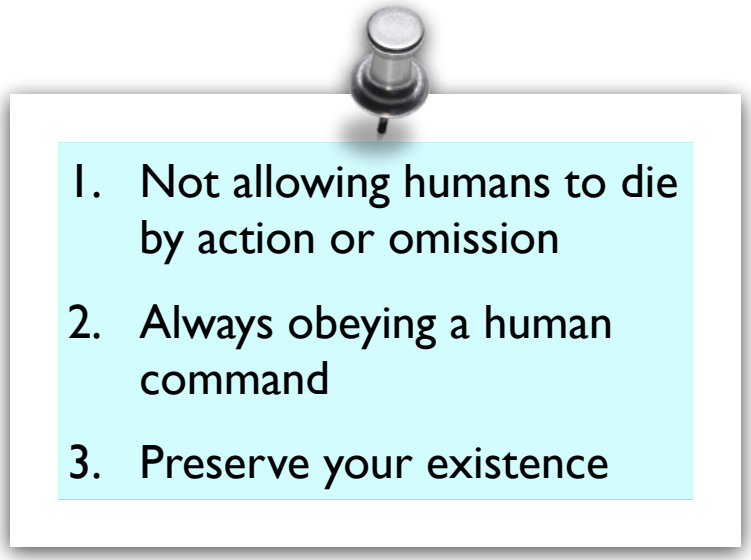


- Conflicts can arise in systems like this
- A conflict is what we called a “moral dilemma”
- One approach to resolving conflicts is ranking the moral rules by priority

A Deontological Moral System

1. If it's ok for someone to do something, it should be ok for everyone to do that same something

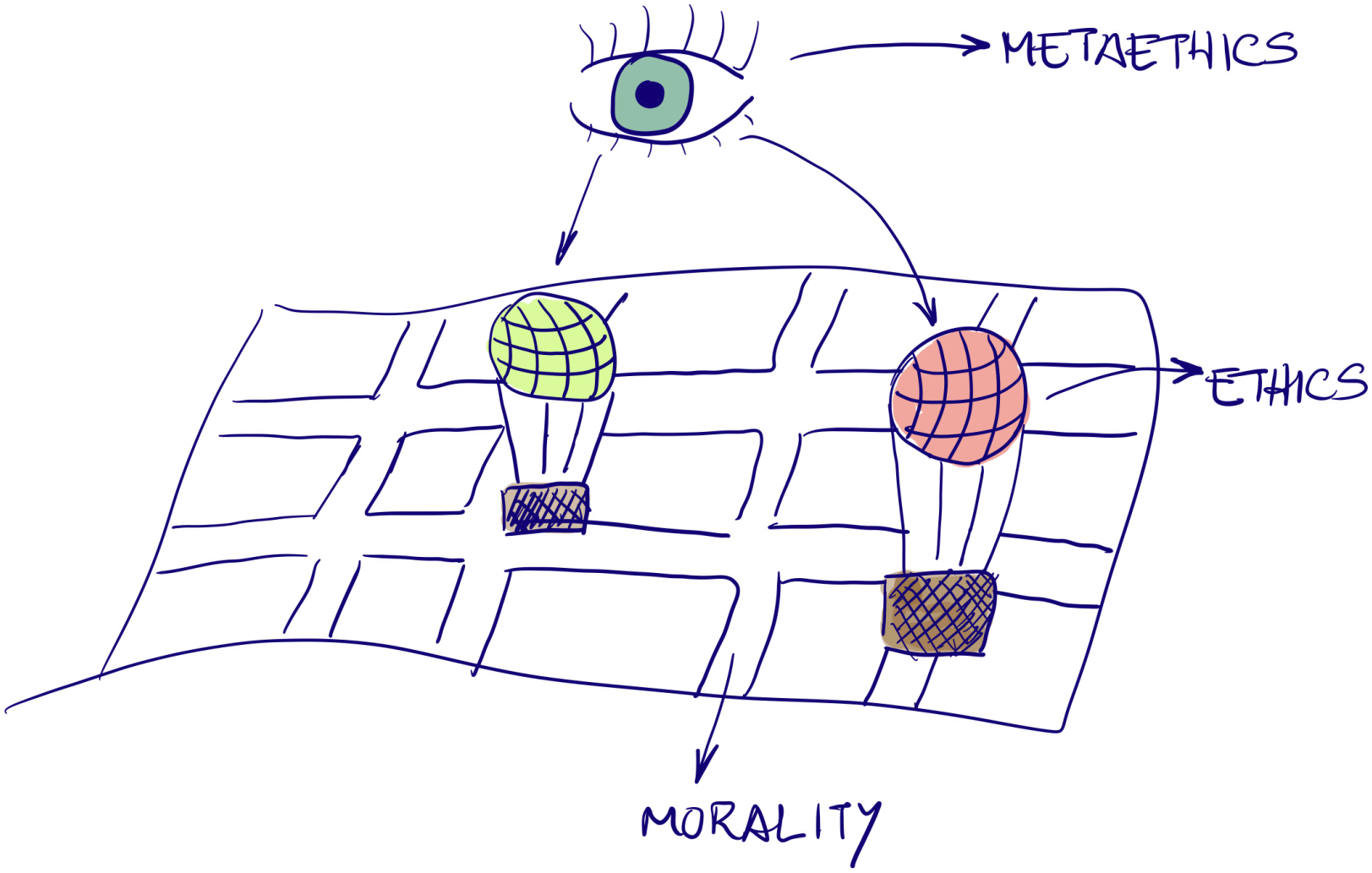
A Deontological Moral System



1. Not allowing humans to die by action or omission
2. Always obeying a human command
3. Preserve your existence

- Resolve conflicts by priority…

How to Think of Metaethics



How to Think of **Computer Ethics**

The human designer follows “ethical principles because they are concerned about harm that could come from machine behavior.”

Machine Metaethics, Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.

How to Think of **Machine Ethics**

The machine in imbued with *the means to compute the solutions to ethical quandaries.*

"The machine itself is reasoning on ethical matters."

Machine Metaethics, Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.

How do We Get to **Machine Ethics**

- Use consequentialist theories?
- Use deontological theories?
- Use a virtue-based theory?

How do We Get to **Machine Ethics**

- Collect lots of data on how people make ethical judgements.
- Use AI (supervised learning) to teach the machine how to resolve similar ethical quandaries. **???**

Machine Metaethics, Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.

Challenges to **Machine Ethics**

- Could we find an ethical theory that resolves all ethical dilemmas?
- Will this ethical theory be computable? (Sidebar: what it means to be computable.)

Machine Metaethics, Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.

Challenges to **Machine Ethics**

- "There are still a number of ethical dilemmas in which even experts are not in agreement as to what is the right action."

- **Danger: ethical relativism** - "the view that when there is disagreement over a particular action is right or wrong, both sides are correct."

Machine Metaethics, Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.

Challenges to **Machine Ethics**

- "Most ethicists believe, however, that in principle there are correct answers to all ethical dilemmas"
- "Do no permit machines to function autonomously in domains in which there is controversy concerning what is correct behavior."
- In the meantime: create an **automated advisor.**

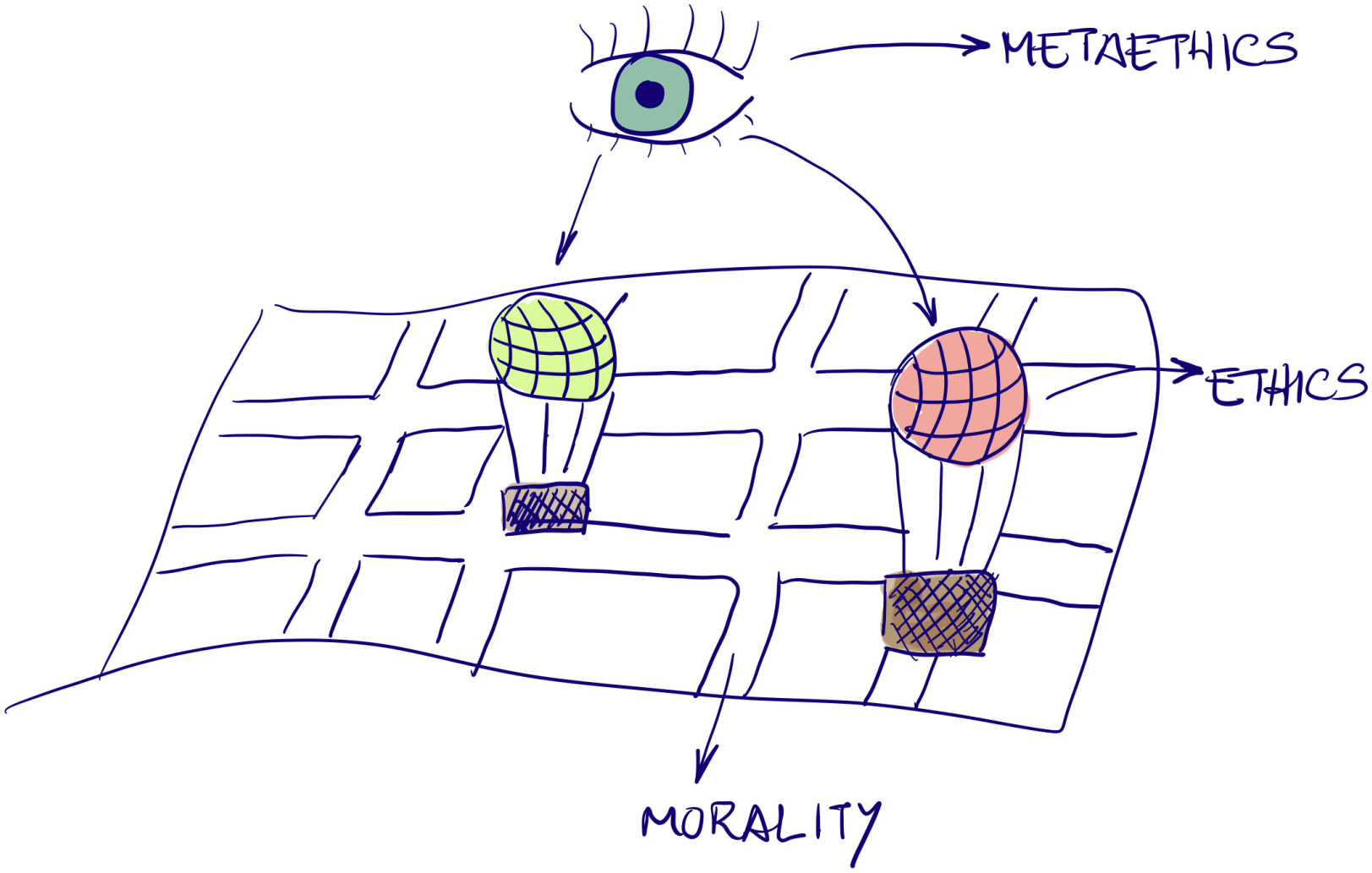
Machine Metaethics, Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.

What AI Can Do

- Consistency is essential to rationality.
- The machine implementation of an ethical theory may be superior to the average human at following that theory.
- The machine will not be confused by emotions.

Machine Metaethics, Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.

How to Think of Metaethics



How to Think of **Machine Metaethics**

Machine metaethics examines the field of machine ethics.

- What is the ultimate goal of machine ethics?
- Is ethics computable?
- Is there a single ethical theory we should implement?
- Is it necessary to determine the moral status of the machine for it be an autonomous ethical agent?

Asimov's "Three Laws of Robotics and Machine Metaethics," Susan Leigh Anderson in Machine Metaethics, Michael Anderson and Susan Leigh Anderson, Cambridge University Press, 2011.