

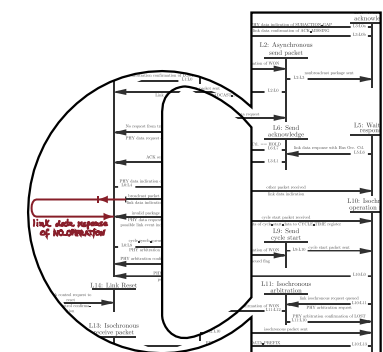


# Ethics for Nerds

An Advanced Course in Computer Science  
Summer Semester 2020

Current Topics C7.1  
Moral Autonomous Systems

What is an autonomous system?



Prof. Holger Hermanns,  
Kevin Baum, Sarah Sterz



UNIVERSITÄT  
DES  
SAARLANDES  
1



# AUTONOMOUS SYSTEMS



<https://www.flickr.com/photos/smoothgroover22/15104006386>



<http://theconversation.com/machines-with-guns-debating-the-future-of-autonomous-weapons-systems-39795>



<https://www.flickr.com/photos/arselectronica/5514133373>

[https://en.wikipedia.org/wiki/File:%D0%A0%D0%BE%D0%B1%D0%BE%D1%82\\_%D0%BF%D1%88%D0%BB%D0%B5%D1%81%D0%BE%D1%81\\_Roomba\\_780.jpg](https://en.wikipedia.org/wiki/File:%D0%A0%D0%BE%D0%B1%D0%BE%D1%82_%D0%BF%D1%88%D0%BB%D0%B5%D1%81%D0%BE%D1%81_Roomba_780.jpg)



[https://en.wikipedia.org/wiki/File:Robomow\\_10\\_City\\_2012-06-05.jpg](https://en.wikipedia.org/wiki/File:Robomow_10_City_2012-06-05.jpg)





## What is an Autonomous System?

⇒ **Characterization:** An autonomous system is a system that performs complex tasks with a high degree of autonomy.

*But that's not autonomy  
as we have seen it already!*

## Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

### Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

### Action Government – Reasons-Responsiveness View

An agent governs her own actions only if her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does.

### Action Government – Incompatibilist View

An agent governs her actions only if her actions cannot be fully explained as the effects of causal powers that are independent of her, even if her beliefs and attitudes are among these effects.

## Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

### Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

### Action Government – Reasons-Responsiveness View

An agent governs her own actions only if her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does.

### Action Government – Incompatibilist View

An agent governs her actions only if her actions cannot be fully explained as the effects of causal powers that are independent of her, even if her beliefs and attitudes are among these effects.

## Autonomy (working definition)

A system is autonomous if and only if it governs its own **behaviour**.

## Behaviour Government – Coherentist View

A system governs its own **behaviour** if and only if it is **caused** to behave as it does because this **cause** coheres with some **internal representation** that represents its **model** of the **behaviour**.

## Behaviour Government – Reasons-Responsiveness View

A system governs its own **behaviour** only if its **computations** that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as it does.

## Behaviour Government – Incompatibilist View

A system governs its behaviour only if its **behaviour** cannot be fully explained as the effects of causal powers that are independent of it, even if its **model of the world** is among these effects.

I as a philosopher say: it's obvious that these artificial agents are not autonomous in the outlined sense, so they are not autonomous at all, and “autonomous system” is a misnomer.



## What, then, is an Autonomous System?

An autonomous system is a system that performs complex tasks with a high degree of autonomy.



### **System?**

≅ is a set of interacting or interdependent (hardware or software) component parts forming a complex whole



### **performing complex tasks?**

≅ achieving an at least partially pre-defined objective in a reasonable amount of time that is not achievable in a straight-forward way



### **Autonomy?**

≅ capable of adapting their behavior in a way in which they achieve their objectives with no or little human intervention



# AUTONOMOUS SYSTEMS

We can operationalize as follows:

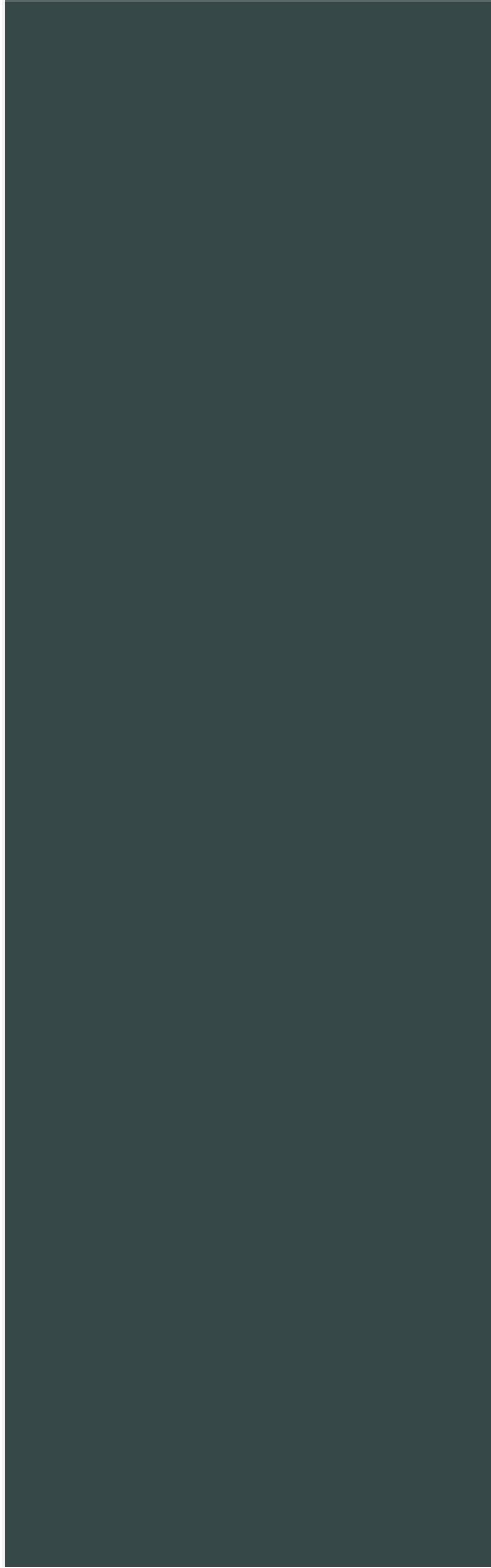
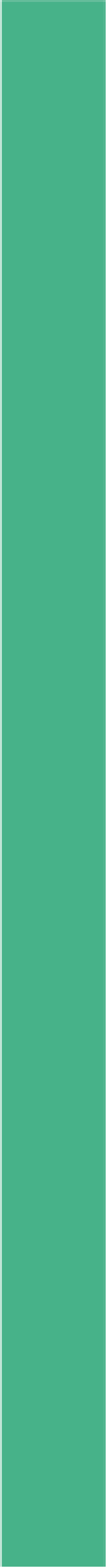
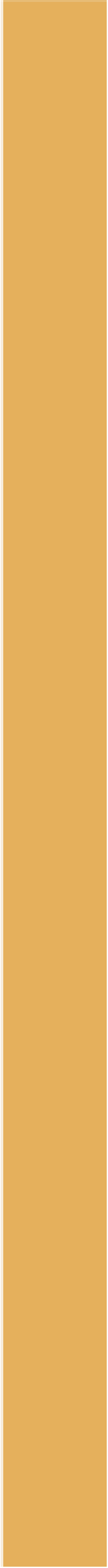
various degrees of **autonomy**

≅ various degrees of independence of human control and intervention for the completion of a complex task:



And with this operationalization our characterization fits, too:

- ⇒ An **autonomous system** is a system that performs complex tasks with a high degree of autonomy.
- ⇒ An **autonomous system** is a system that performs complex tasks with a high degree of independence of human control and intervention.





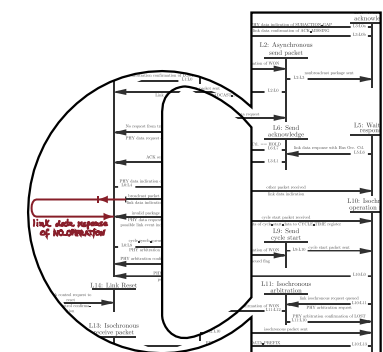


# Ethics for Nerds

An Advanced Course in Computer Science  
Summer Semester 2020

Current Topics C7.2  
Moral Autonomous Systems

Do ethics apply?



Prof. Holger Hermanns,  
Kevin Baum, Sarah Sterz



UNIVERSITÄT  
DES  
SAARLANDES  
11

FIRST THOUGHTS ON BEHAVIOR

Suppose a child was run over:

	human driver	autonomous car
it was impossible to see child in time	tragic, but nothing morally wrong done	tragic, but nothing morally wrong done
possible to see child in general, but (predictably and obviously) visually ill-equipped to do so	blame him, pro tanto it was wrong to drive in the first place	probably the driver did something wrong, or the programmer, or the manufacturer, or the vendor, ...? but most certainly not the car itself

Why do those two diverge?



## THE SCOPE OF MORALITY

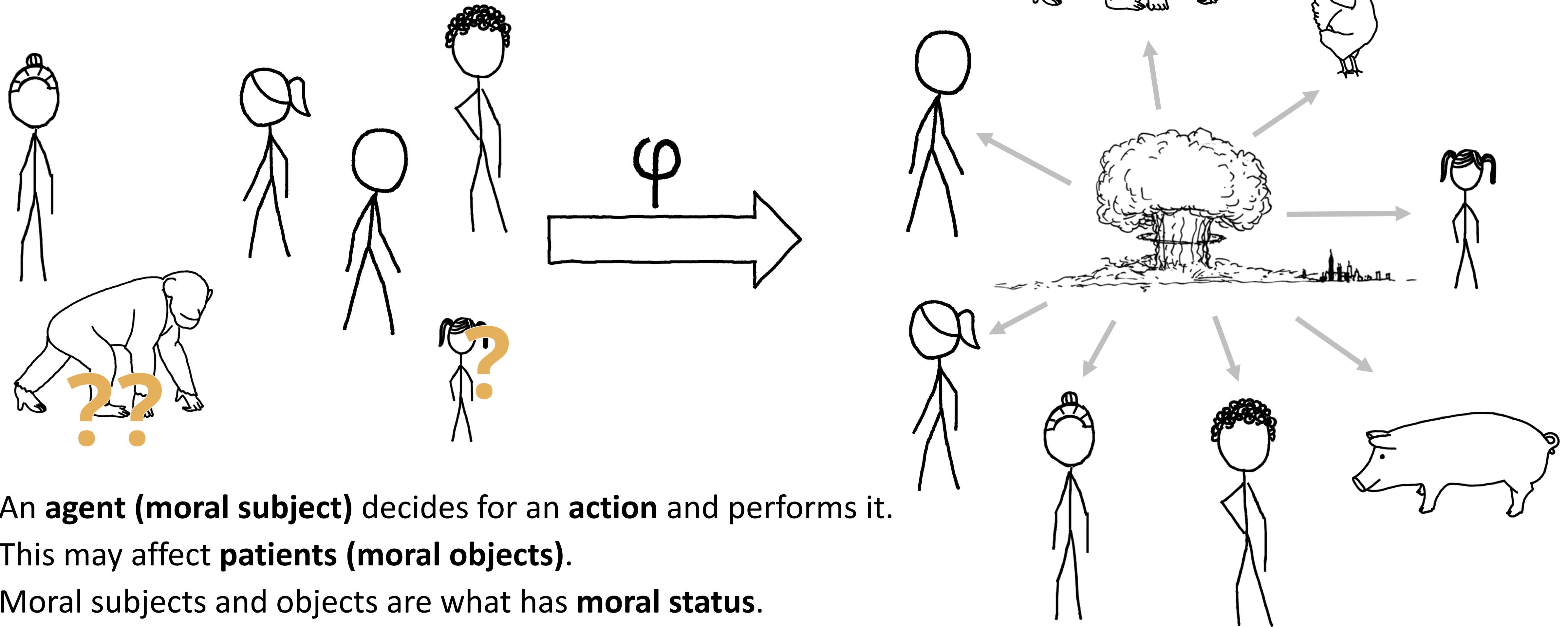
- The **hammer** fell off the board and onto Niko's foot.  
*What a bad hammer! The hammer shouldn't have done this.*
- The **fridge** kept the beverages cold.  
*This is good, as the fridge had the obligation to do so.*
- My **goldfish/hamster/cat/dog** bit me.  
*This was morally impermissible!*
- This **toddler** deliberately broke his plate when he saw that we are having spinach today.  
*He ought not to act like this!*
- These **4<sup>th</sup> graders** beat up Niko.  
*That was clearly wrong.*
- **John** cheated on his wife.  
*How immoral of him!*

nonsense



standard way of  
talking

## THE SCOPE OF MORALITY

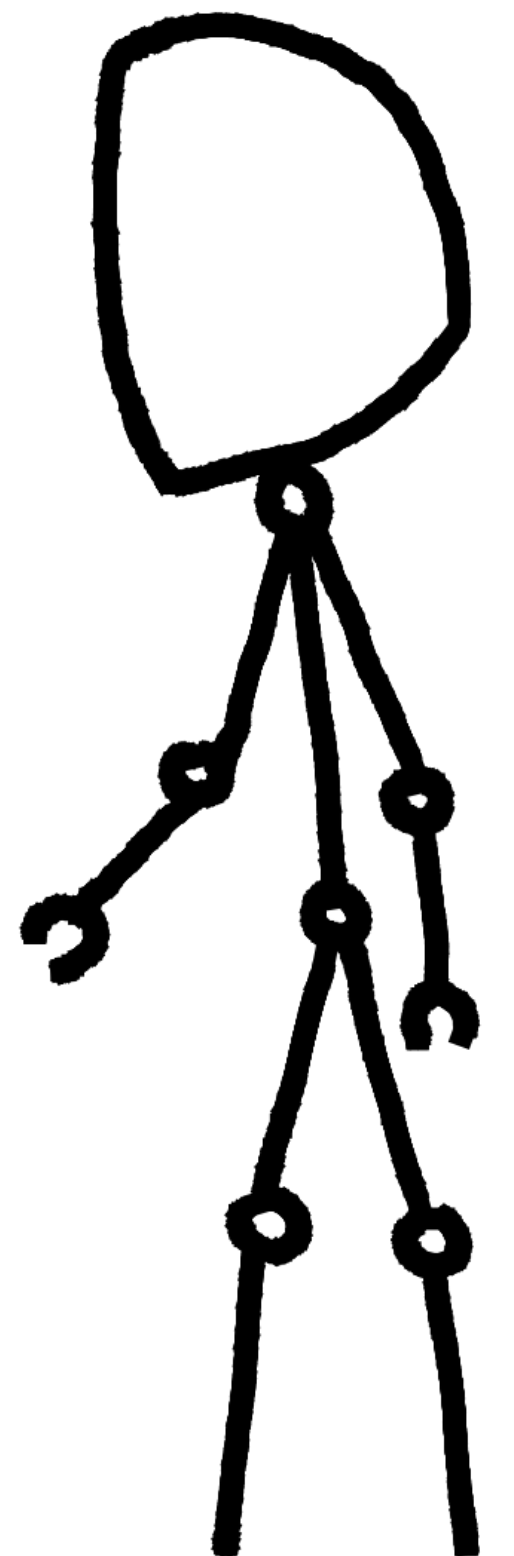


An **agent (moral subject)** decides for an **action** and performs it.  
This may affect **patients (moral objects)**.  
Moral subjects and objects are what has **moral status**.

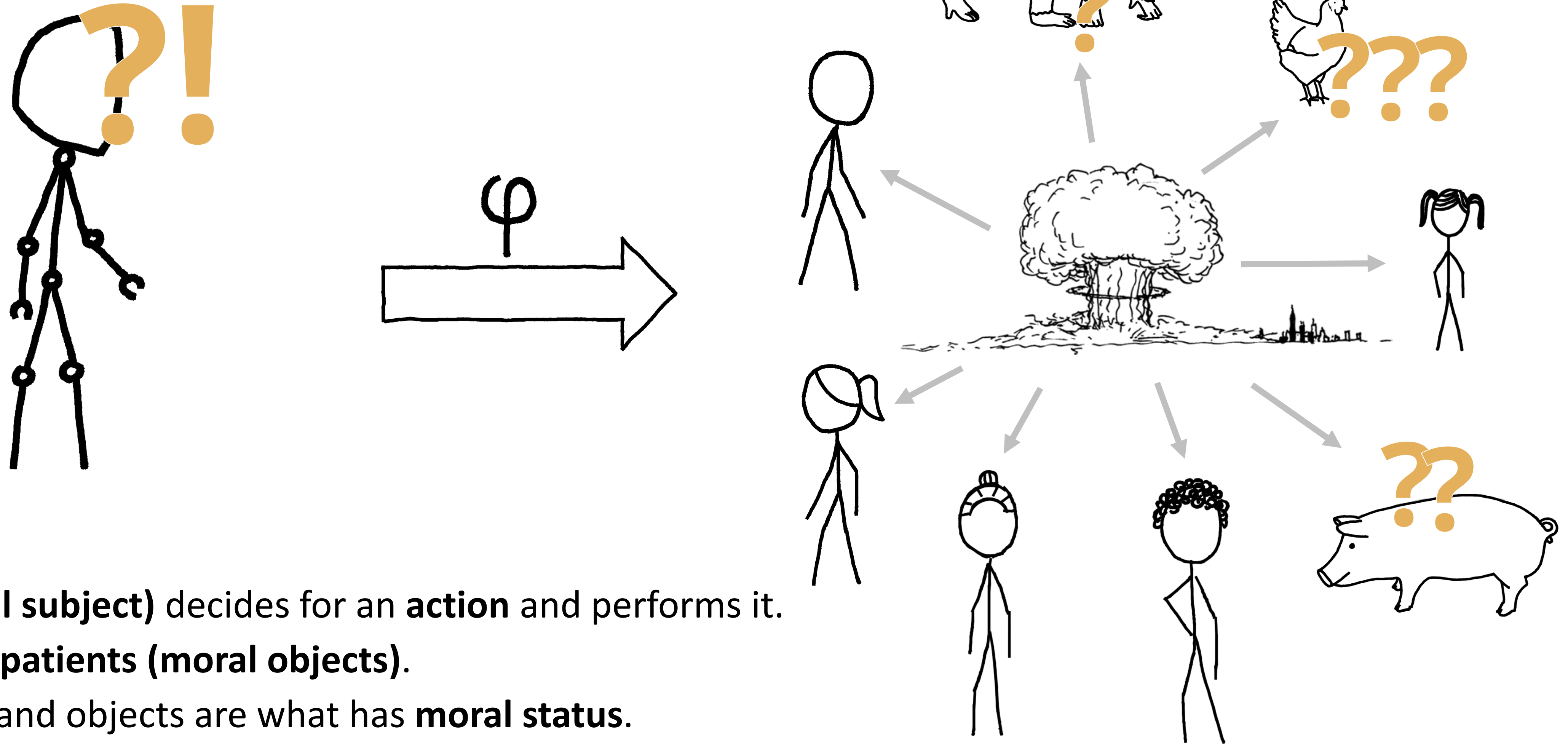


## THE SCOPE OF MORALITY

- The **autonomous lawn mower** shredded a ladybug.  
*That was not right of it!*
- The **autonomous cars** ran over the group of kids instead of a cat.  
*It should have done otherwise!*
- The **lethal autonomous weapon system** destroyed that city for no reason.  
*That was wrong of it!*
- The **care robot** forbade that old lady to watch TV.  
*That was a morally impermissible action.*



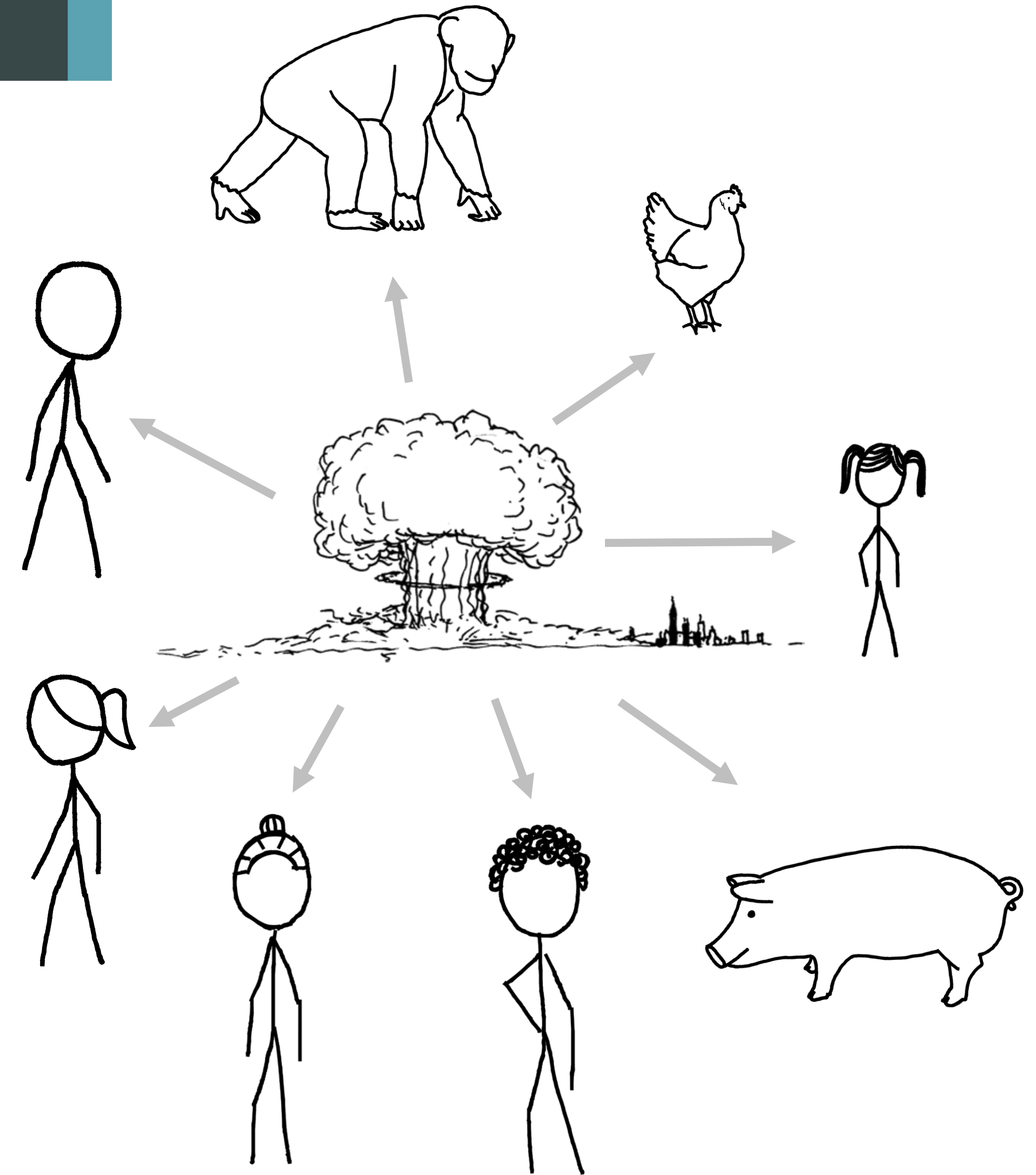
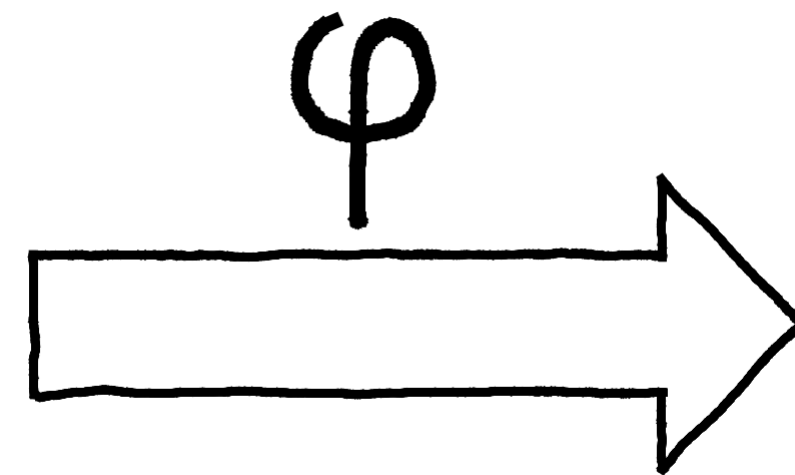
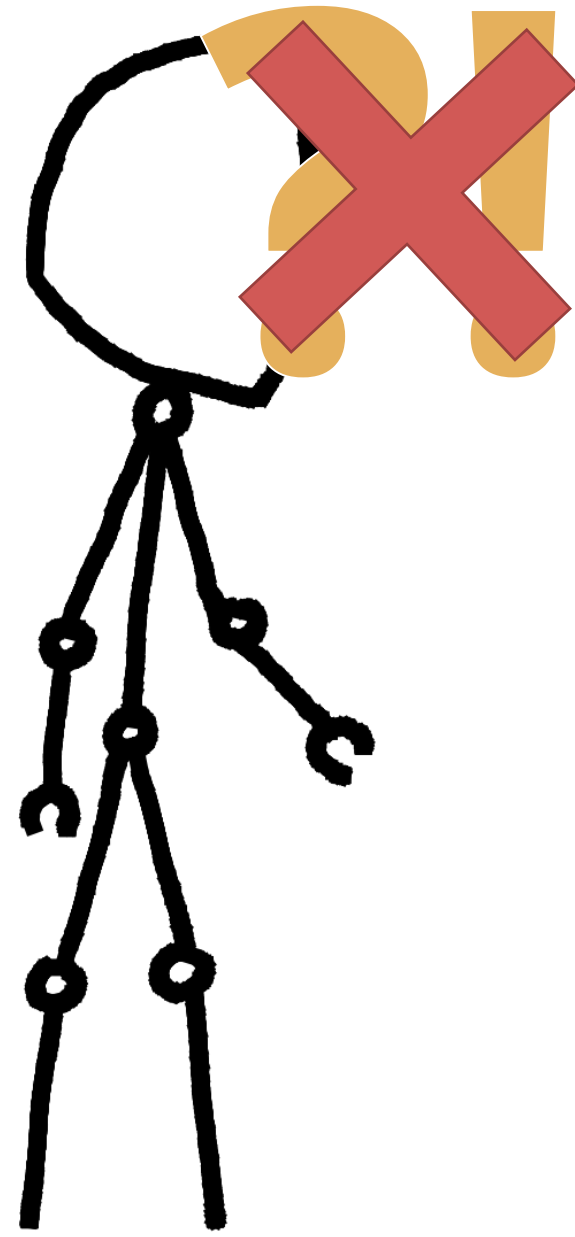
## THE SCOPE OF MORALITY



An **agent (moral subject)** decides for an **action** and performs it.  
This may affect **patients (moral objects)**.  
Moral subjects and objects are what has **moral status**.

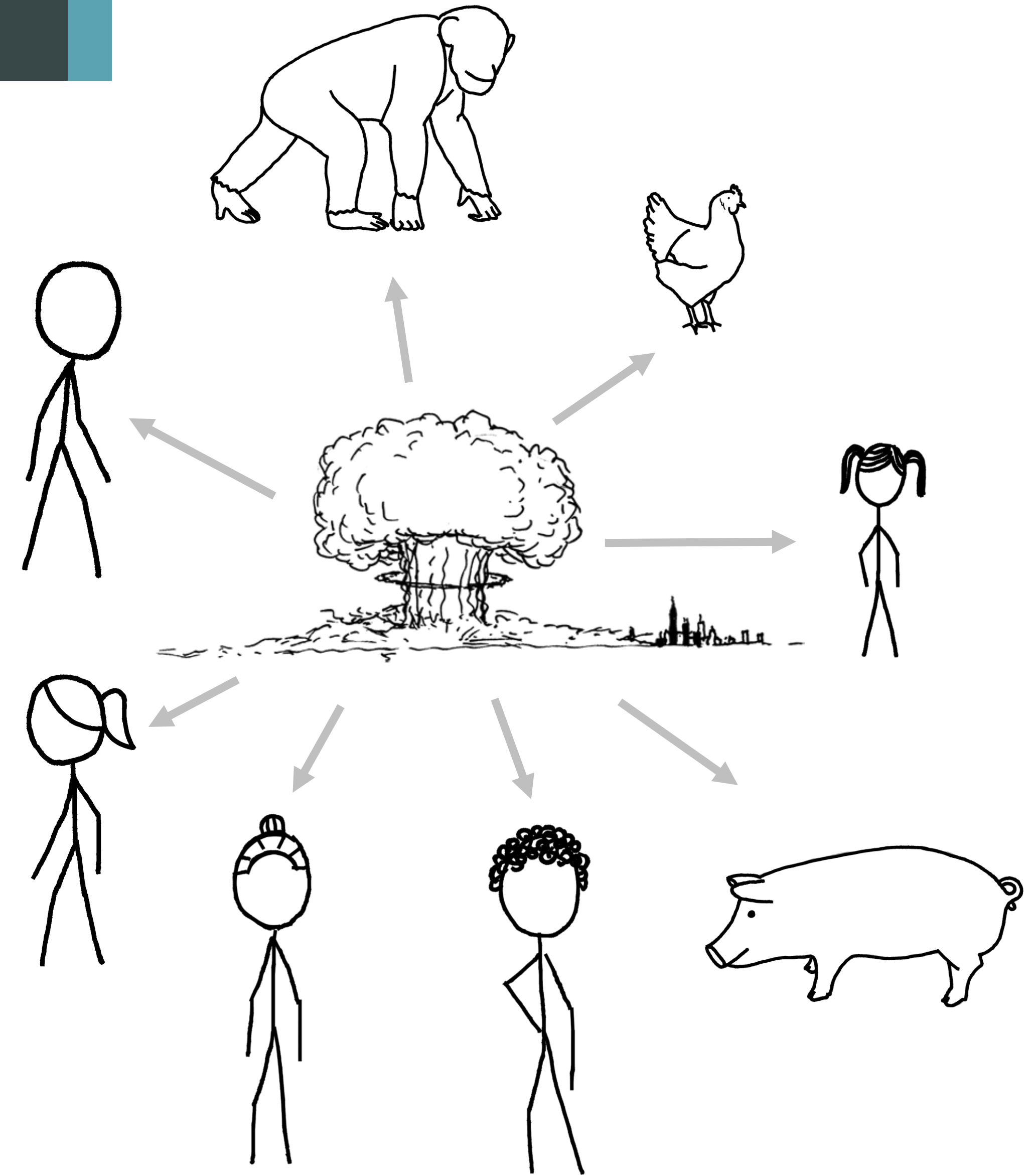
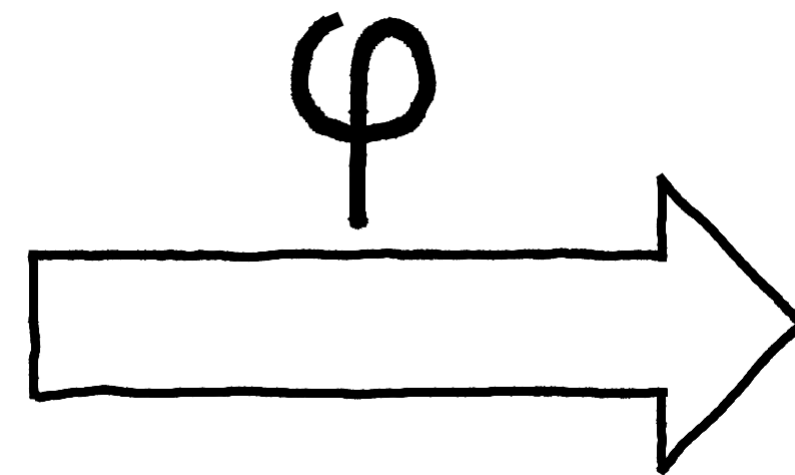
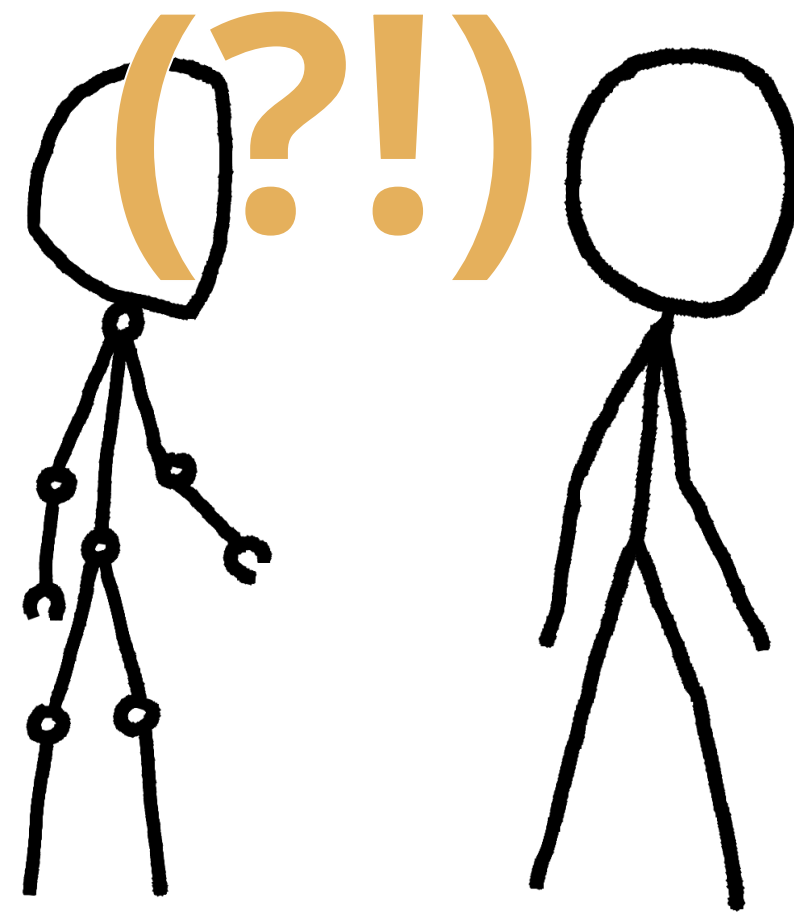


# THE SCOPE OF MORALITY



- Autonomous systems are not agents.
- Is the basic question behind machine ethics ill-posed?

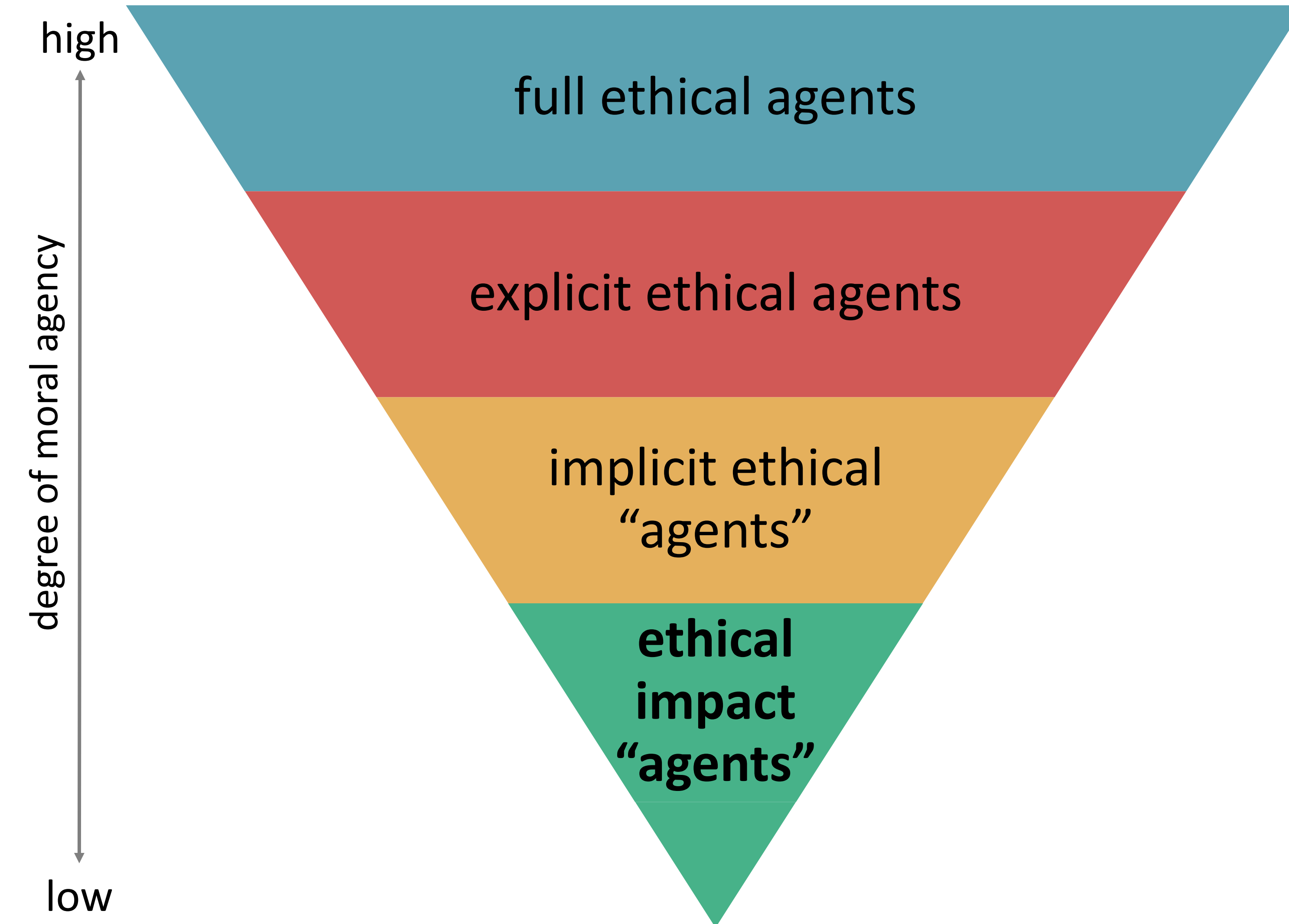
## THE SCOPE OF MORALITY



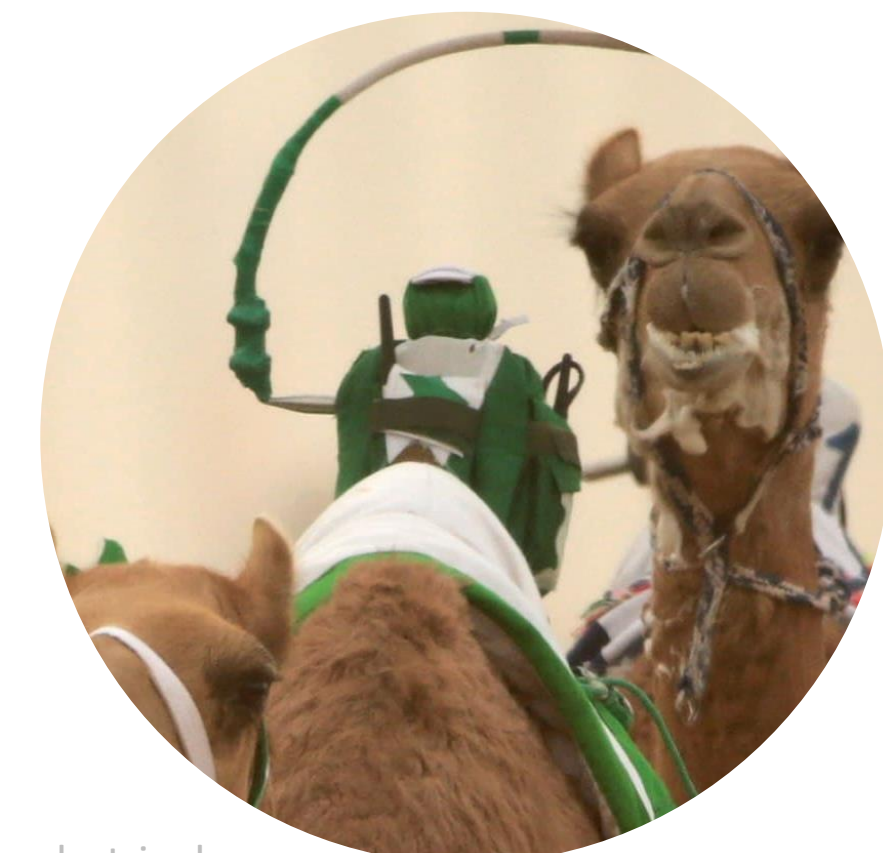
- Even if today's autonomous systems are not moral agents, they still produce some morally relevant output depending on their design. So, we should look for (morally) better designs. While doing that, it could be helpful to pretend them to be moral agents.
- The basic question behind machine ethics is not ill-posed.

# ARTIFICIAL MORAL AGENTS

Adapted from James H. Moor (in “The nature, importance, and difficulty of machine ethics”, *IEEE Intelligent Systems* 21 (4): 18 – 21 .2006),  
[https://philosophynow.org/issues/72/Four\\_Kinds\\_of\\_Ethical\\_Robots](https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots)  
(a 2009 summary of his 2006)



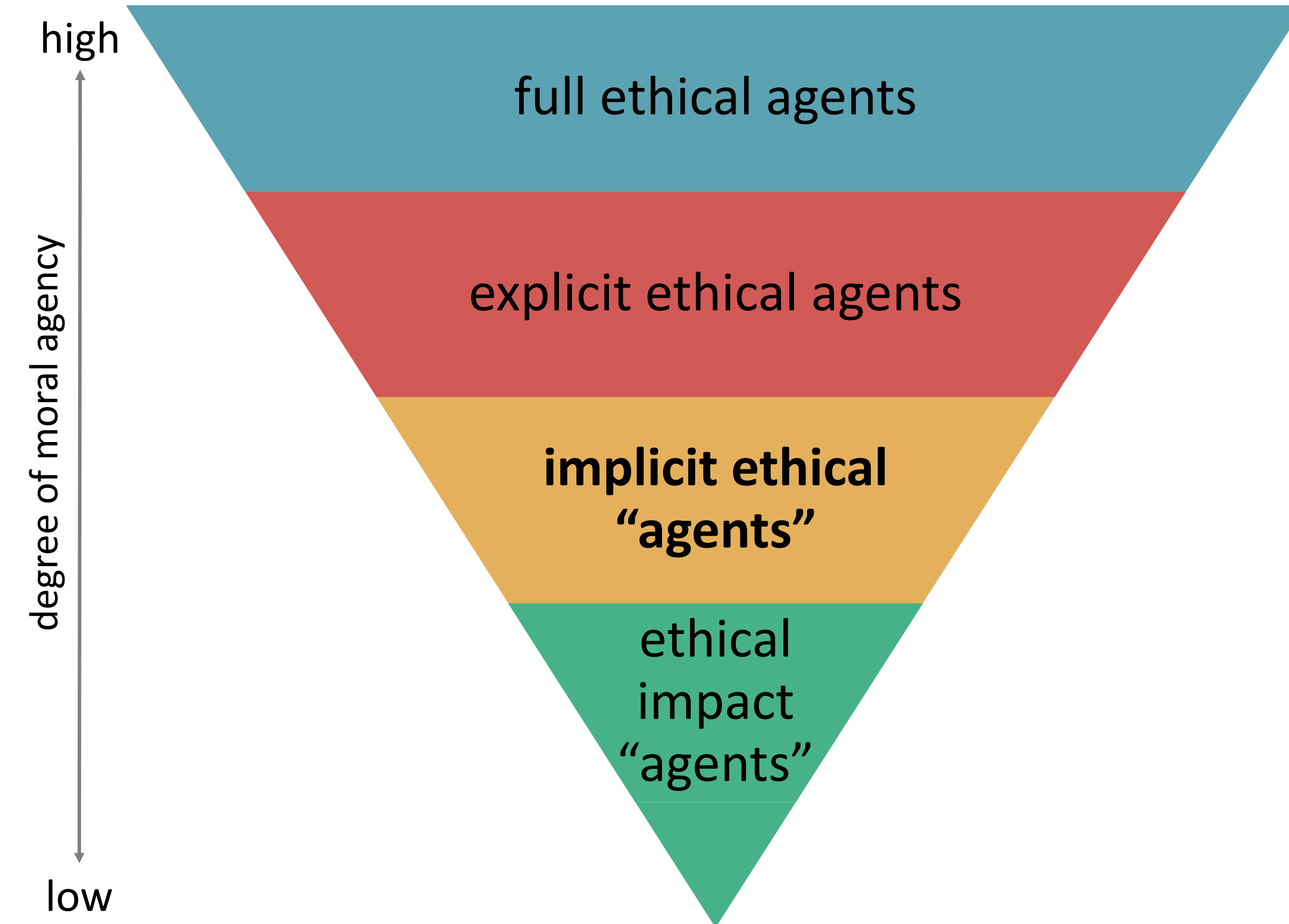
- agents which actions have ethical consequences whether intended or not
- *Examples:*
  - almost any robot
  - an alarm clock
  - robotic camel jockeys



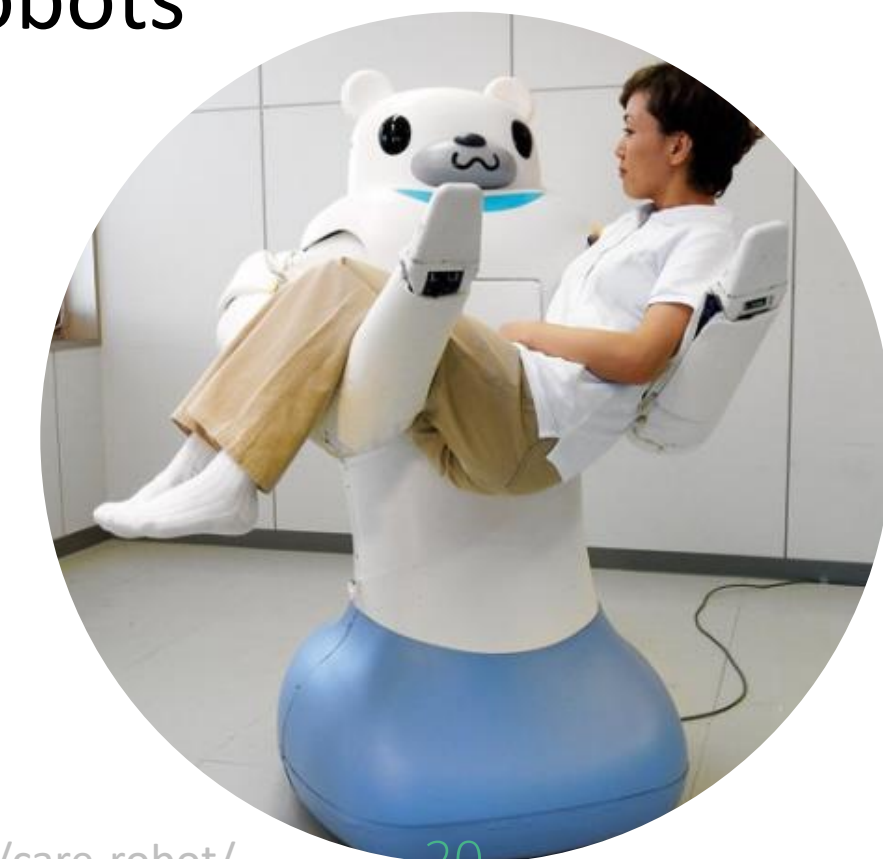


# ARTIFICIAL MORAL AGENTS

Adapted from James H. Moor (in “The nature, importance, and difficulty of machine ethics”, *IEEE Intelligent Systems* 21 (4): 18 – 21 .2006),  
[https://philosophynow.org/issues/72/Four\\_Kinds\\_of\\_Ethical\\_Robots](https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots)  
(a 2009 summary of his 2006)

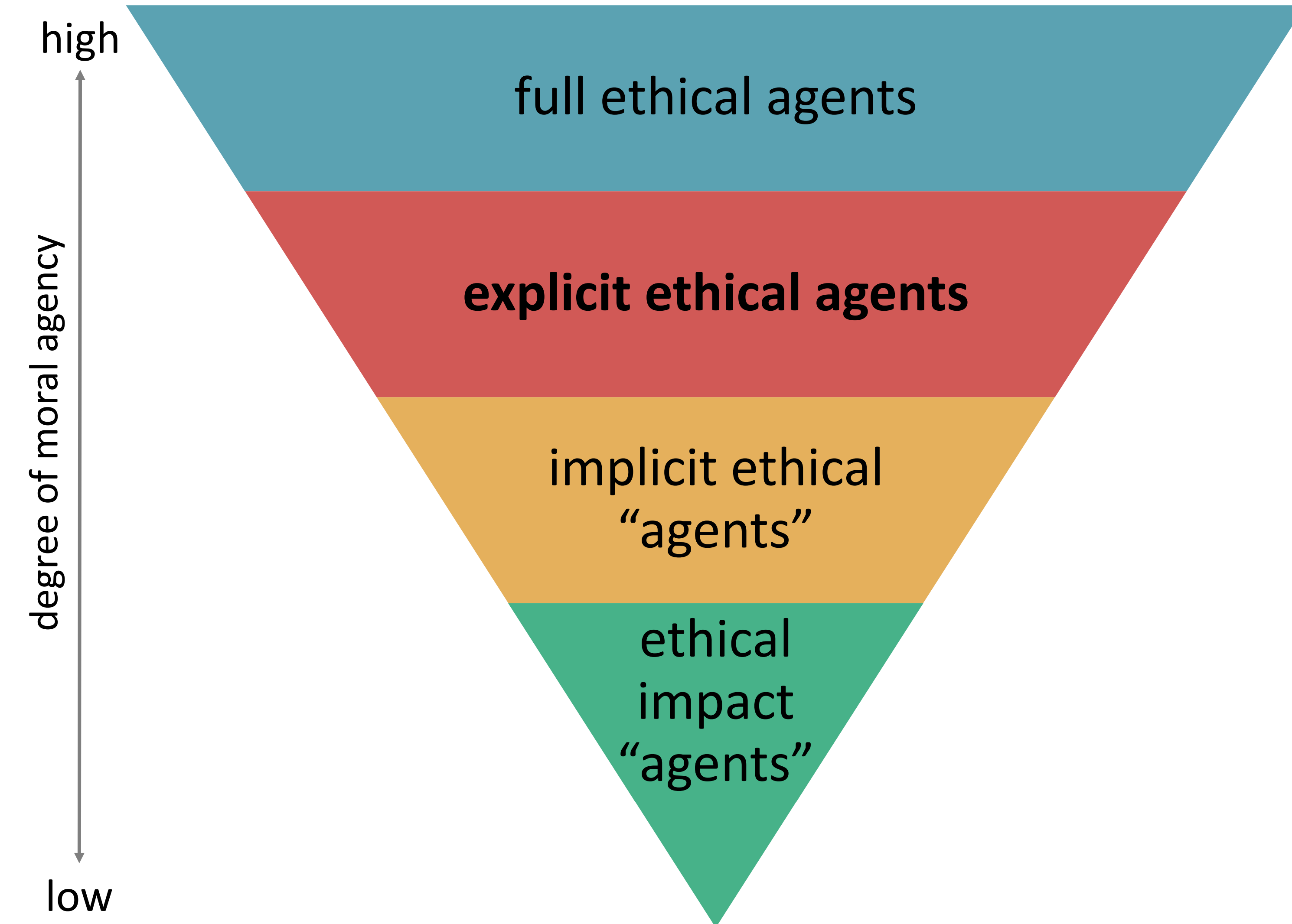


- agents that have ethical considerations built into (i.e. implicit in) their design
  - typically safety or security considerations
- *Examples:*
  - planes with warning devices
  - primitive care robots



# ARTIFICIAL MORAL AGENTS

Adapted from James H. Moor (in “The nature, importance, and difficulty of machine ethics”, *IEEE Intelligent Systems* 21 (4): 18 – 21 .2006),  
[https://philosophynow.org/issues/72/Four\\_Kinds\\_of\\_Ethical\\_Robots](https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots)  
(a 2009 summary of his 2006)

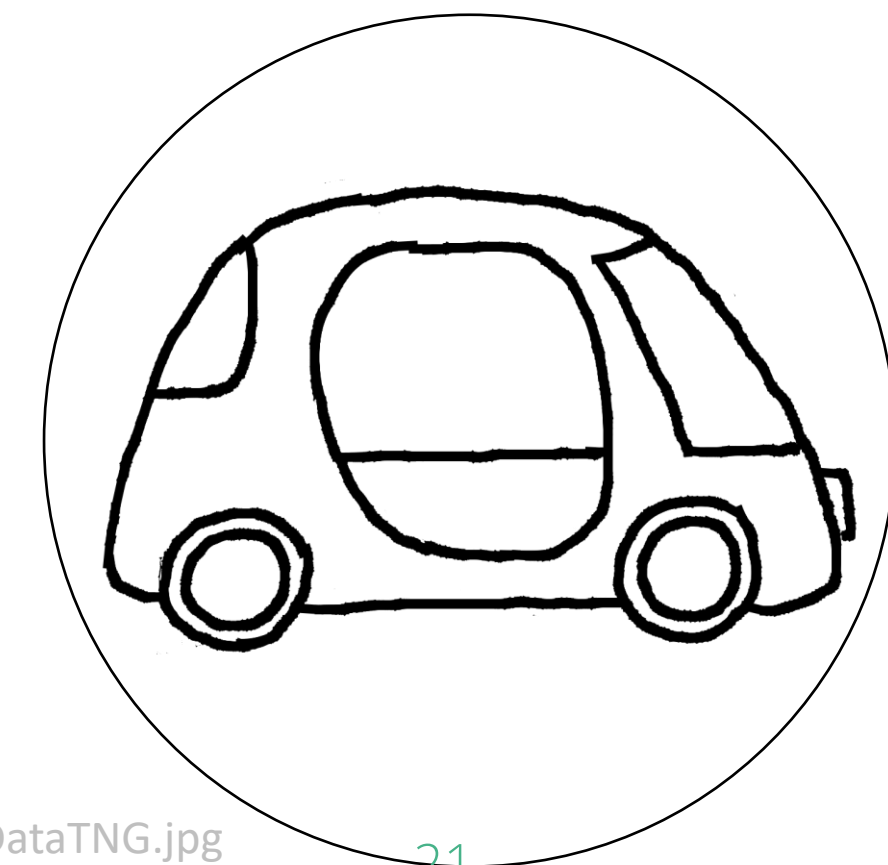


- agents that can identify and process ethical information about a variety of situations and infer what should be done

- can work out reasonable resolutions in case of conflicting principles

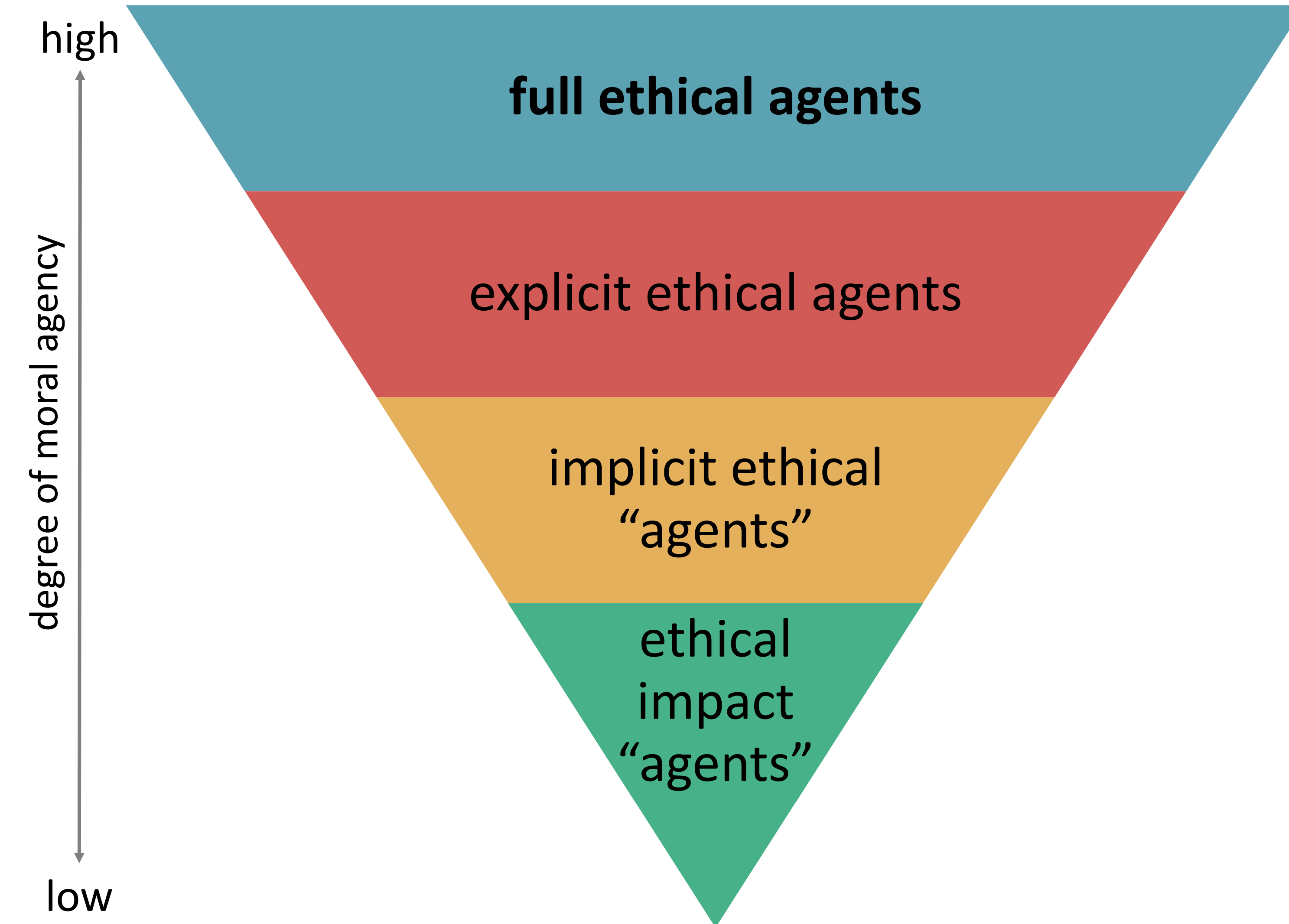
- *Examples:*

- autonomous cars
- sophisticated robots in Sci-Fi movies



# ARTIFICIAL MORAL AGENTS

Adapted from James H. Moor (in “The nature, importance, and difficulty of machine ethics”, *IEEE Intelligent Systems* 21 (4): 18 – 21 .2006),  
[https://philosophynow.org/issues/72/Four\\_Kinds\\_of\\_Ethical\\_Robots](https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots)  
(a 2009 summary of his 2006)



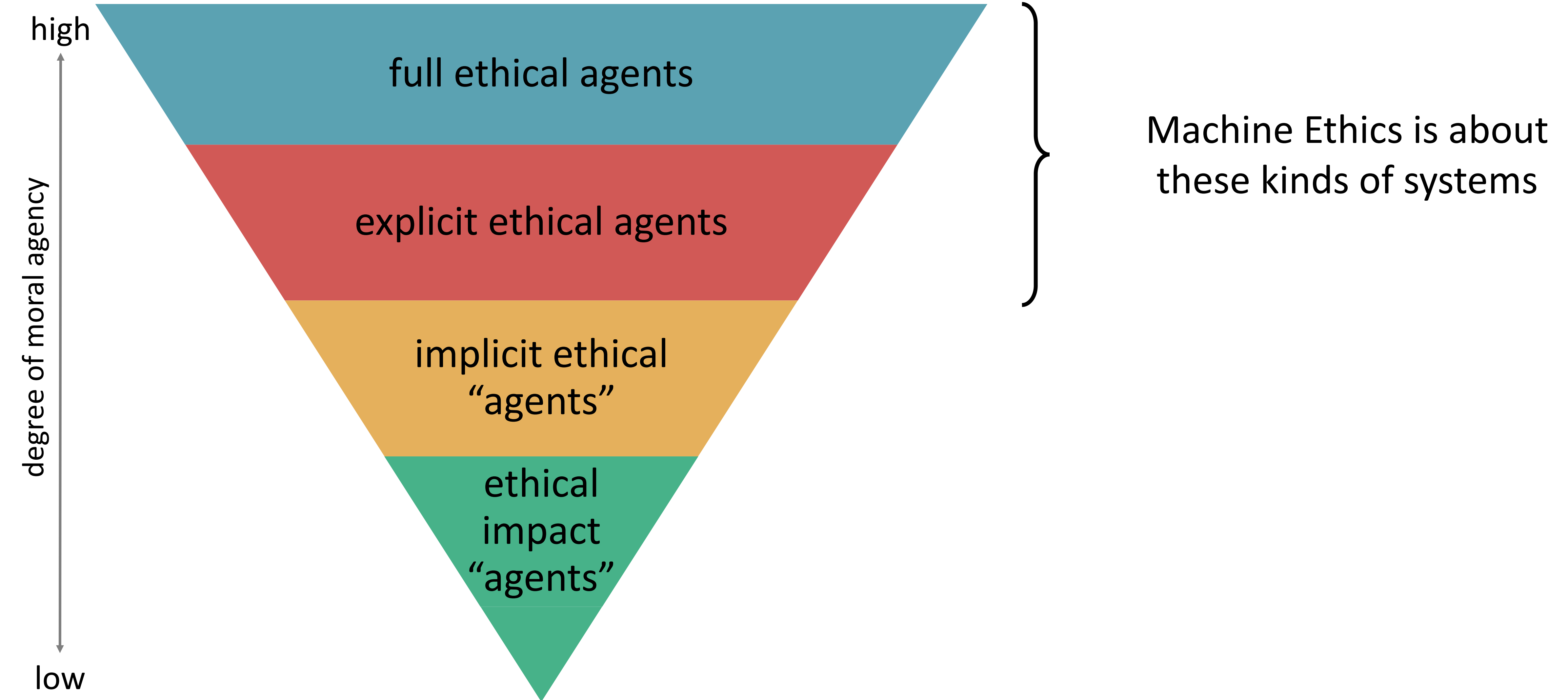
- like full explicit agents, but with central metaphysical features like
  - consciousness
  - intentionality
  - free will (?)
- *Examples:*
  - normal human adults
  - Roy Batty from *Bladerunner*





# ARTIFICIAL MORAL AGENTS

Adapted from James H. Moor (in “The nature, importance, and difficulty of machine ethics”, *IEEE Intelligent Systems* 21 (4): 18 – 21 .2006),  
[https://philosophynow.org/issues/72/Four\\_Kinds\\_of\\_Ethical\\_Robots](https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots)  
(a 2009 summary of his 2006)

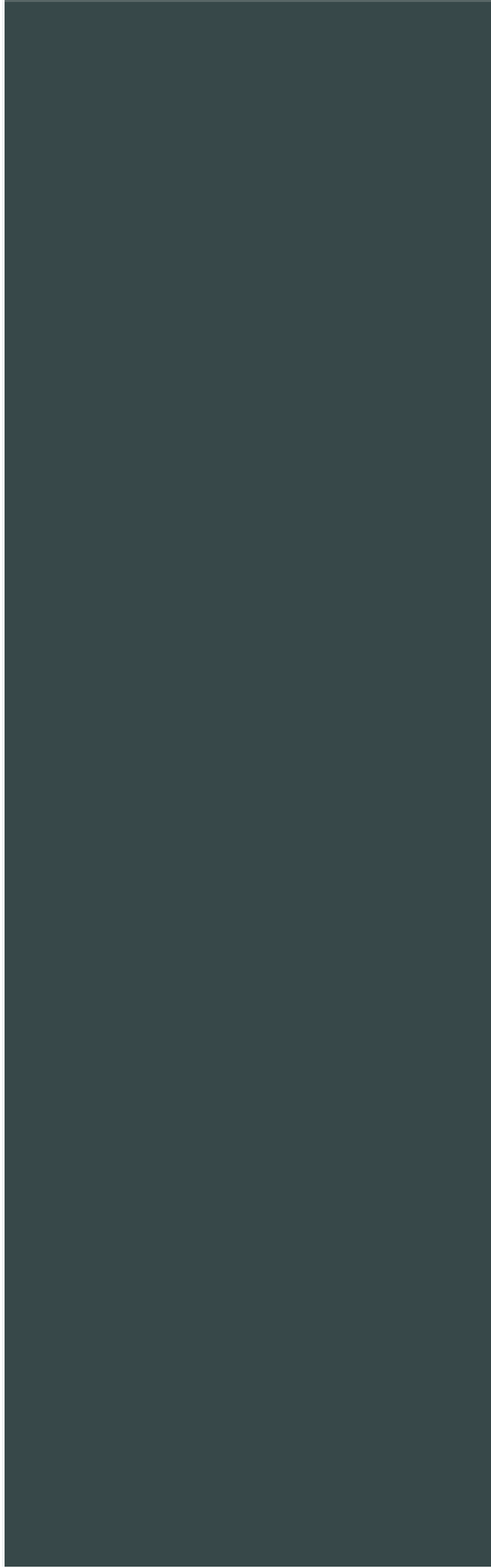
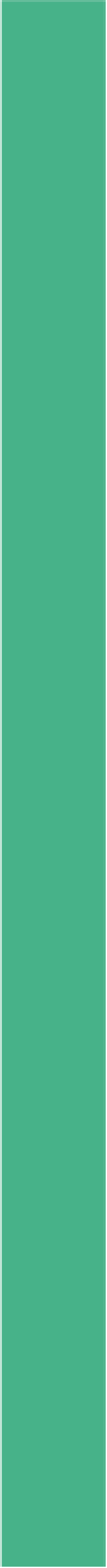
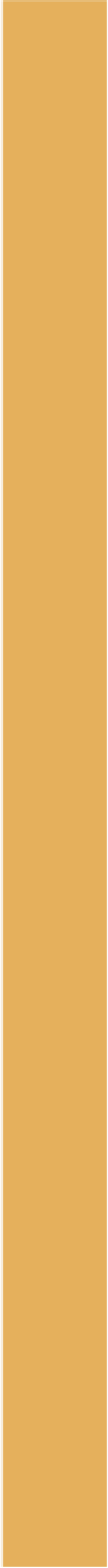
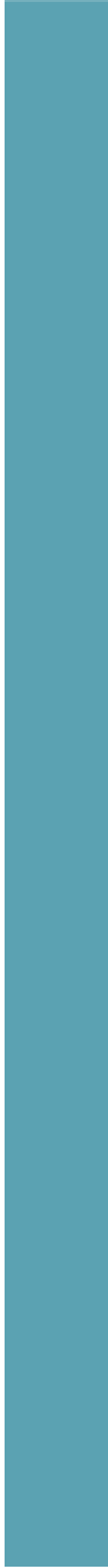




Can autonomous systems be the addressees of moral demands?



Technically not, because they are not moral subjects, but still they produce morally relevant output – and that suffices to be careful about how to implement them.



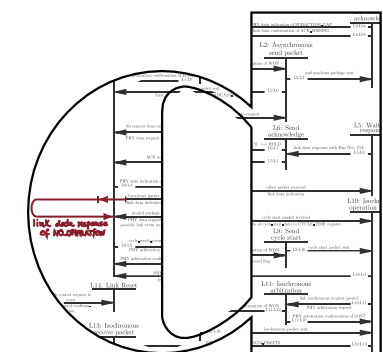




# Ethics for Nerds

An Advanced Course in Computer Science  
Summer Semester 2020

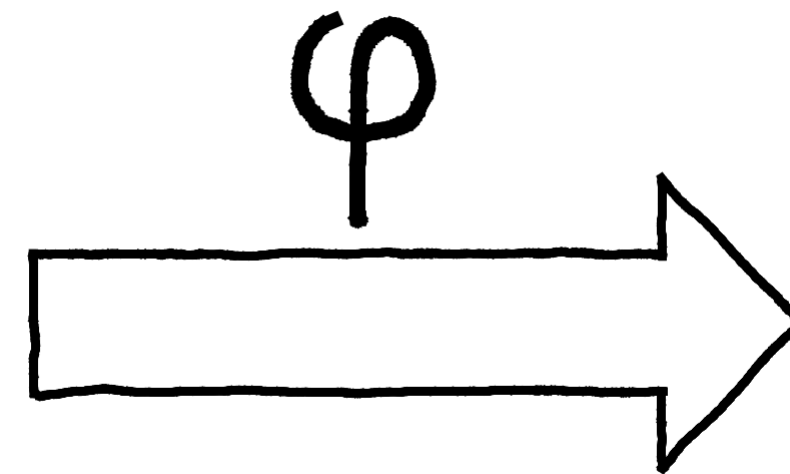
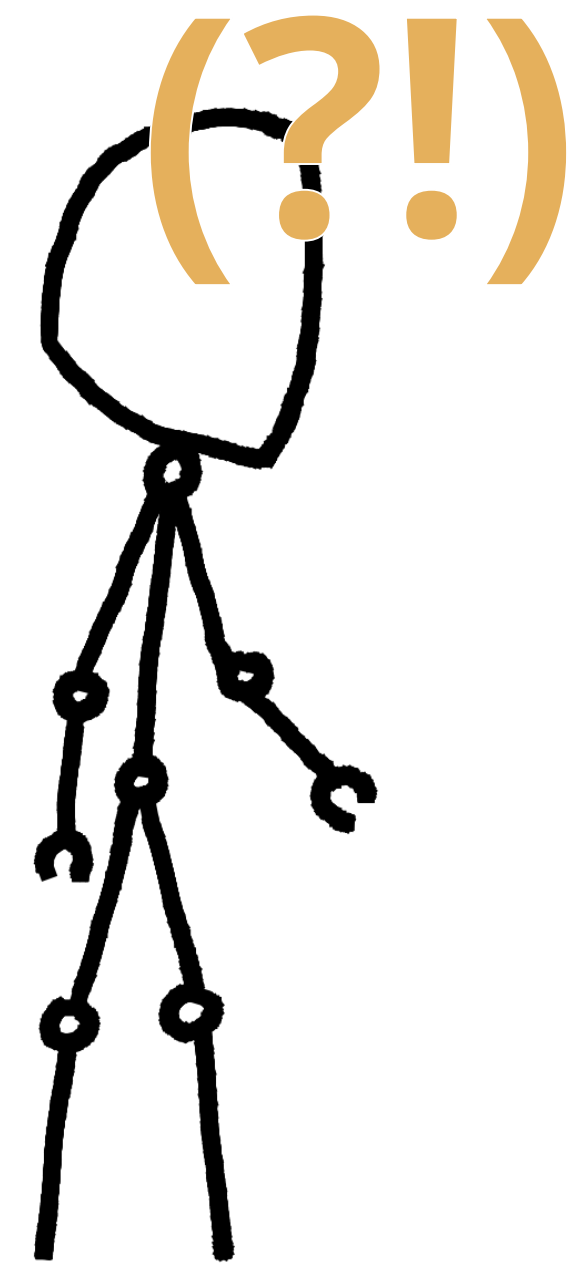
Current Topics C7.3  
Moral Autonomous Systems  
How to implement morality?



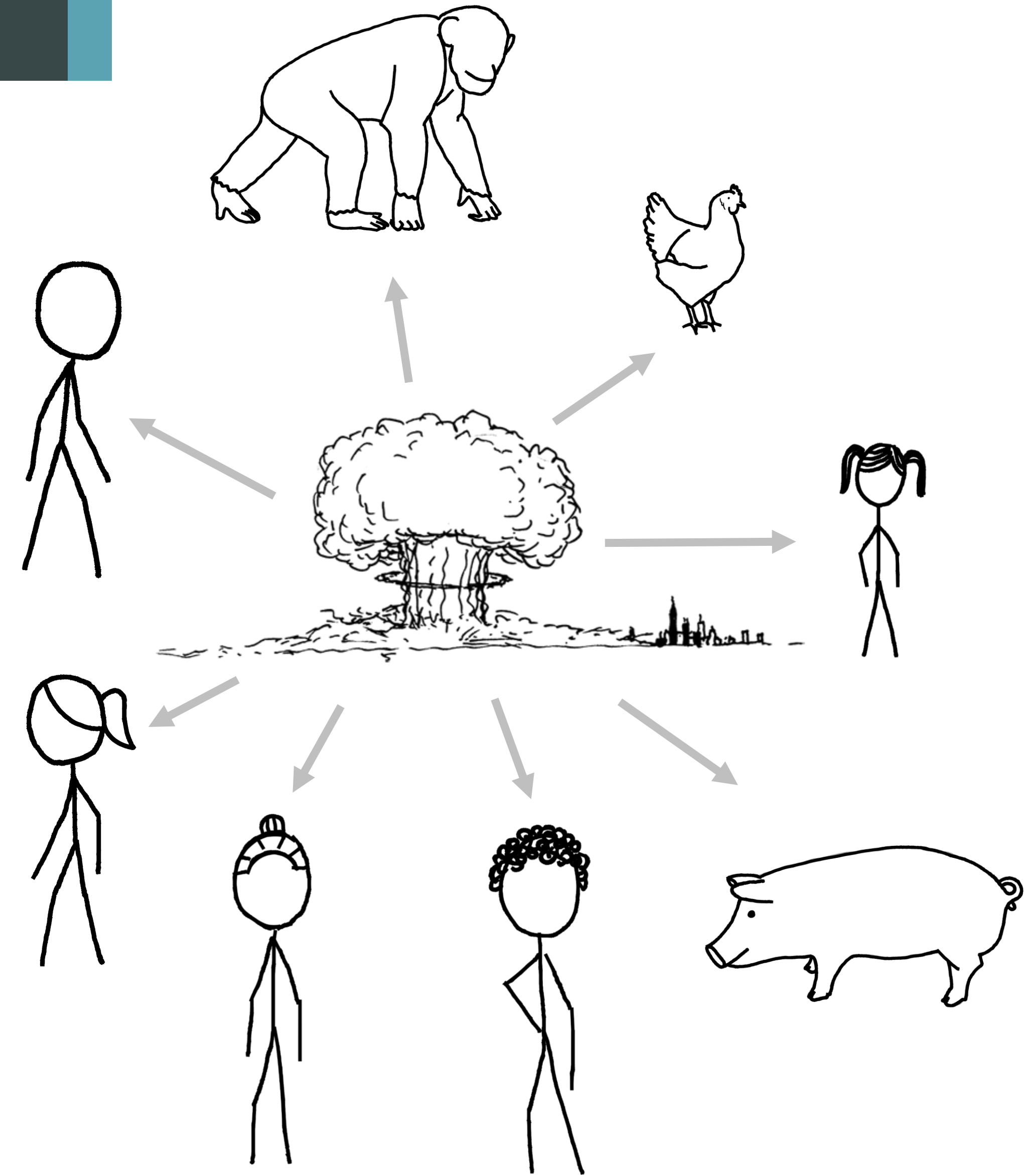
Prof. Holger Hermanns,  
Kevin Baum, Sarah Sterz

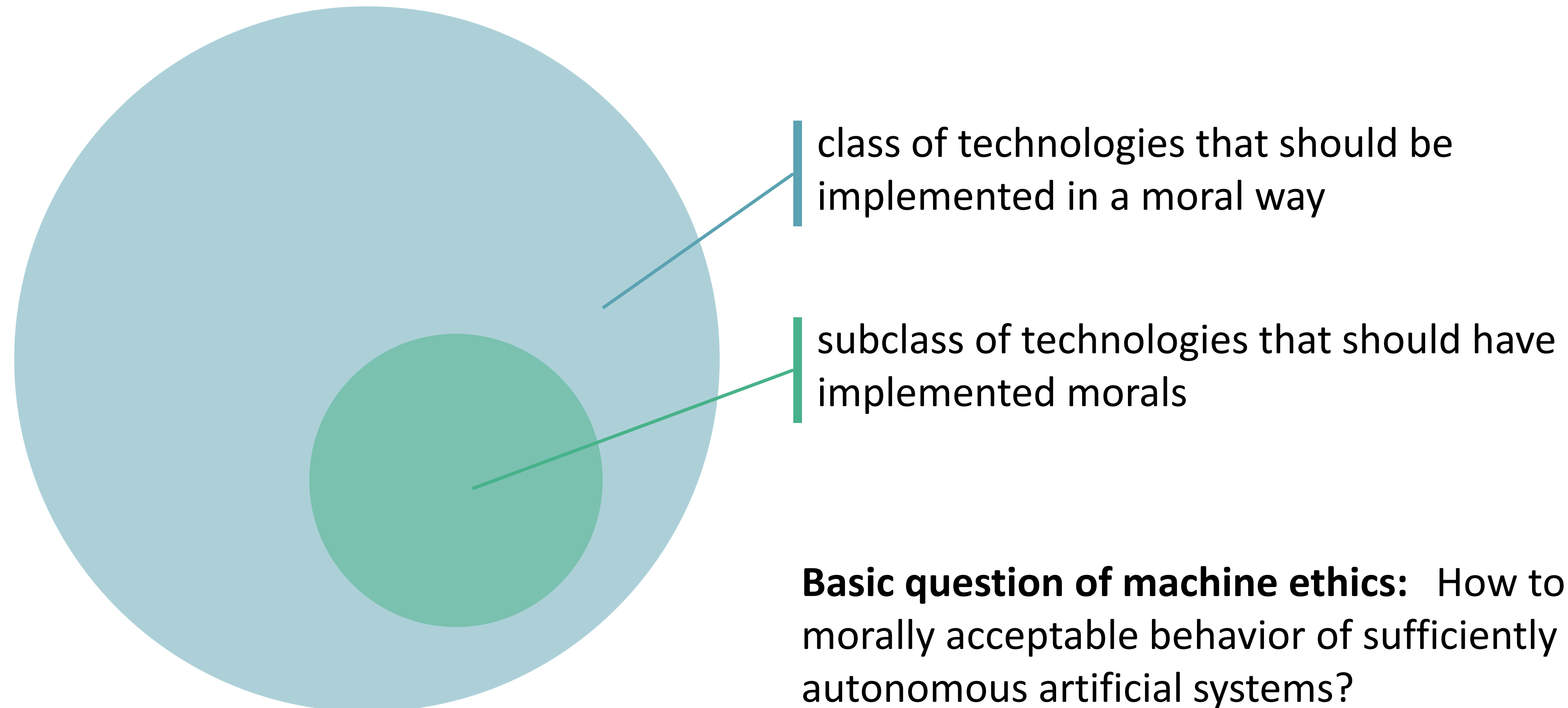


UNIVERSITÄT  
DES  
SAARLANDES  
26



↑  
that's you and you should care  
about what you implement  
and how you do that

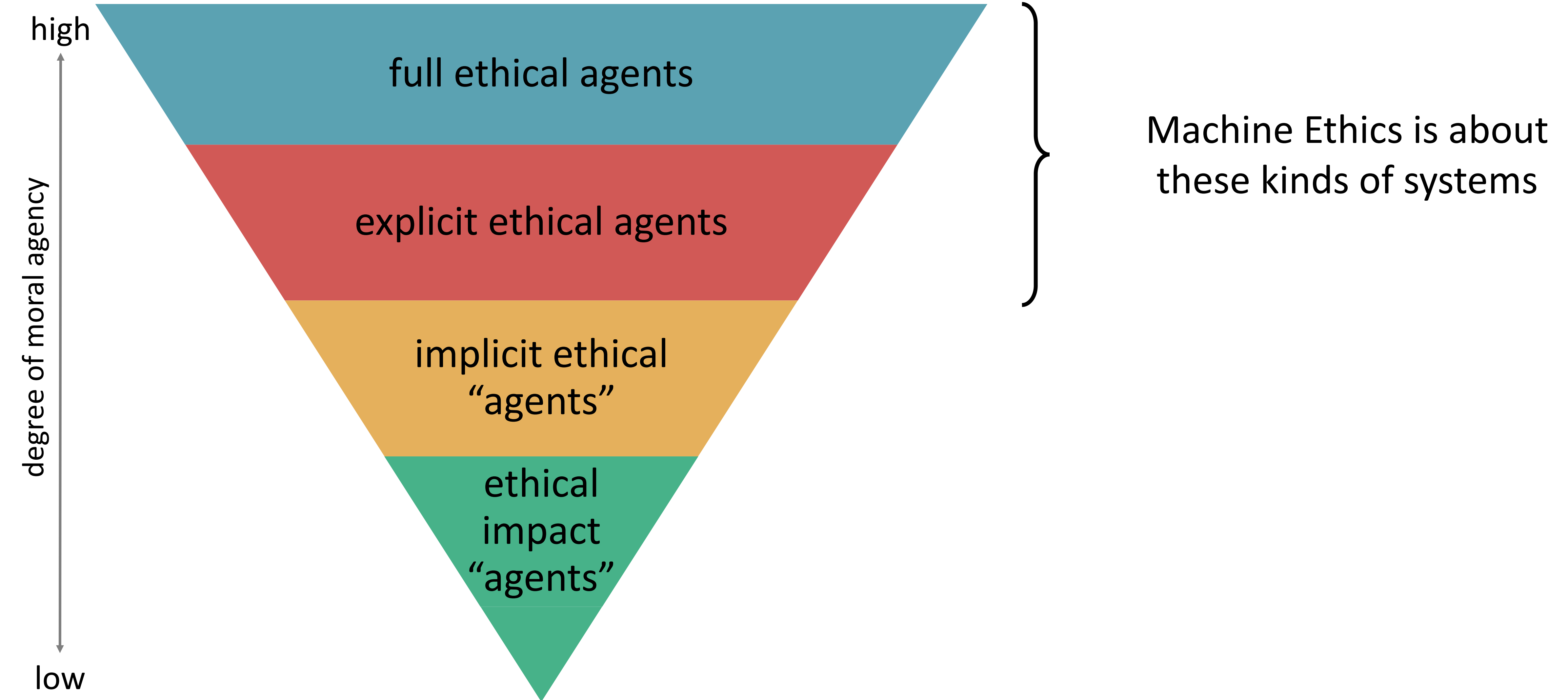






# MACHINE ETHICS

Adapted from James H. Moor (in “The nature, importance, and difficulty of machine ethics”, *IEEE Intelligent Systems* 21 (4): 18 – 21 .2006),  
[https://philosophynow.org/issues/72/Four\\_Kinds\\_of\\_Ethical\\_Robots](https://philosophynow.org/issues/72/Four_Kinds_of_Ethical_Robots)  
(a 2009 summary of his 2006)



### Example (Historically Impactful & Useful ‘Toy’)

Isaac Asimov’s three Laws of Robotics, originally from his story “Runaround” (1942), the short story collection „I, Robot“ (1950)

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



## FIRST THOUGHTS ON RESTRICTED BEHAVIOR

### Example (Historically Impactful & Useful ‘Toy’)

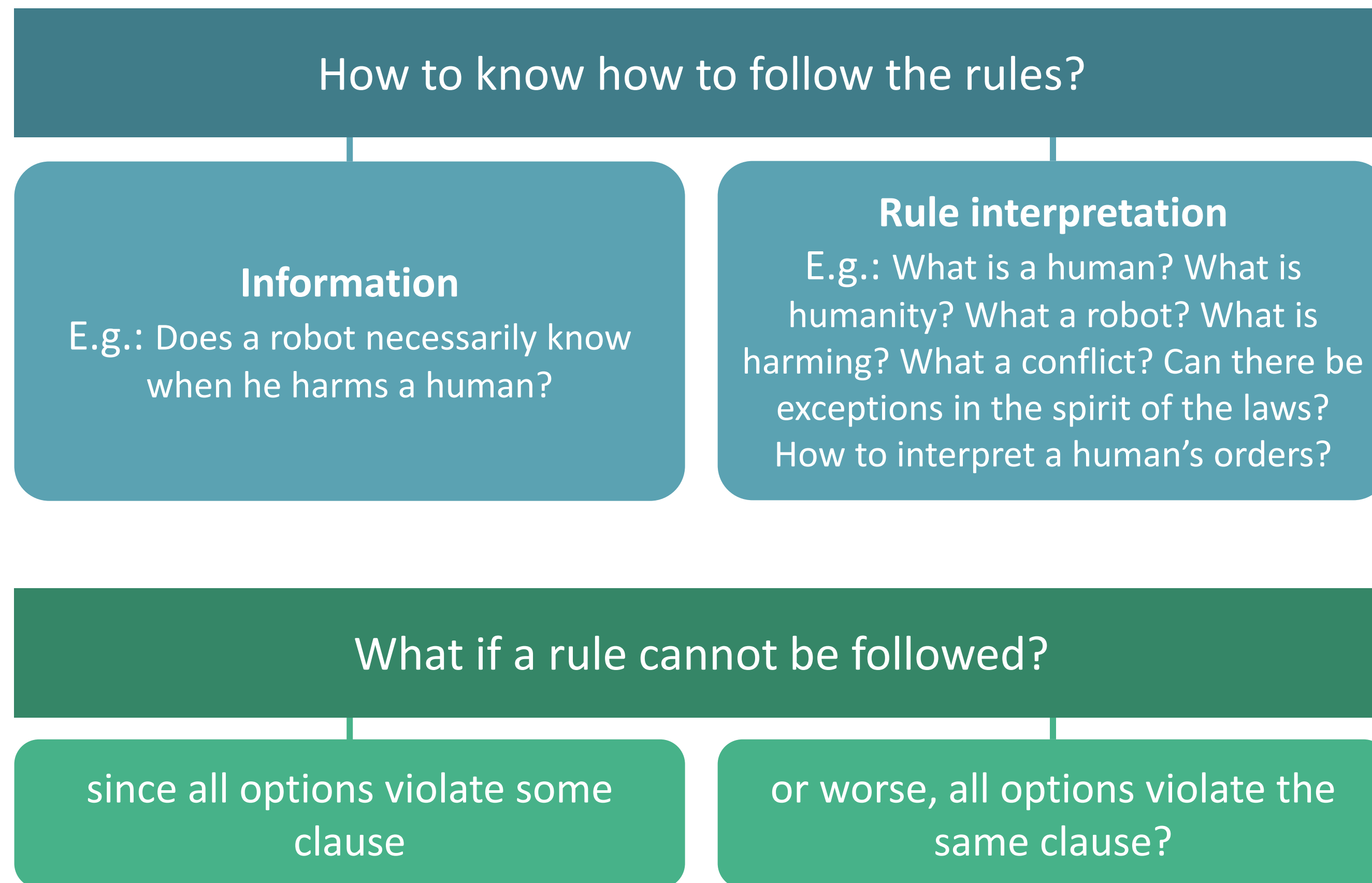
Zeroth law added in “The Robots of Empire” (1983)

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
1. A robot may not injure a human being or, through inaction, allow a human being to come to harm, **except he would otherwise conflict with the Zeroth Law.**
2. A robot must obey the orders given it by human beings except where such orders would conflict with the Zeroth or First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the Zeroth, First or Second Laws.





## Problems

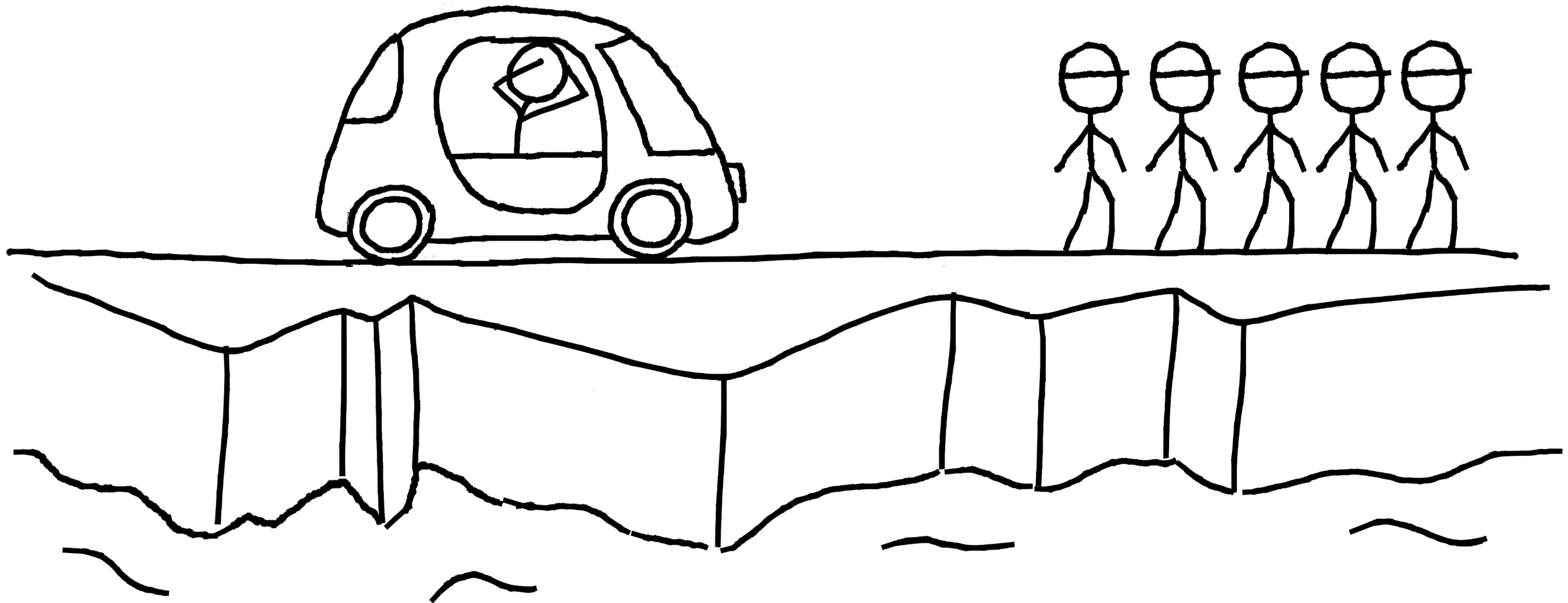


0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
1. A robot may not injure a human being or, through inaction, allow a human being to come to harm, except he would otherwise conflict with the Zeroth Law.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the Zeroth or First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the Zeroth, First or Second Laws.



**Our standard example:** What should the car do in this situation?

→ We have several possibilities to answer that question.



# MACHINE ETHICS: THE CENTRAL PROBLEM

## Possibility 1: Answer in a Confident Way

Really? You can come up with the correct moral theory? Philosophers try that for hundreds and thousands of years...

Find the right, or at least a highly plausible moral theory and then implement this theory in the cars.

How to implement ethical theories? How to even formalize them properly? How to algorithmize them?

Even if we know/can do all that: Do autonomous systems have the necessary capabilities for information acquisition and information processing? Are they fit for the theories?



## Possibility 2: Dodge the Problem

a) We cannot decide what is right in the above scenario, so just follow the rules (don't leave your track).

That doesn't avoid the problem. Even if there were strict rules, it would come down to pure fatalism: "Stuff happens, live with it!"

b) We cannot decide that, so just choose randomly.

That doesn't avoid the problem. It is just giving up: "Stuff happens randomly, live with it!"

## Possibility 2: Dodge the Problem

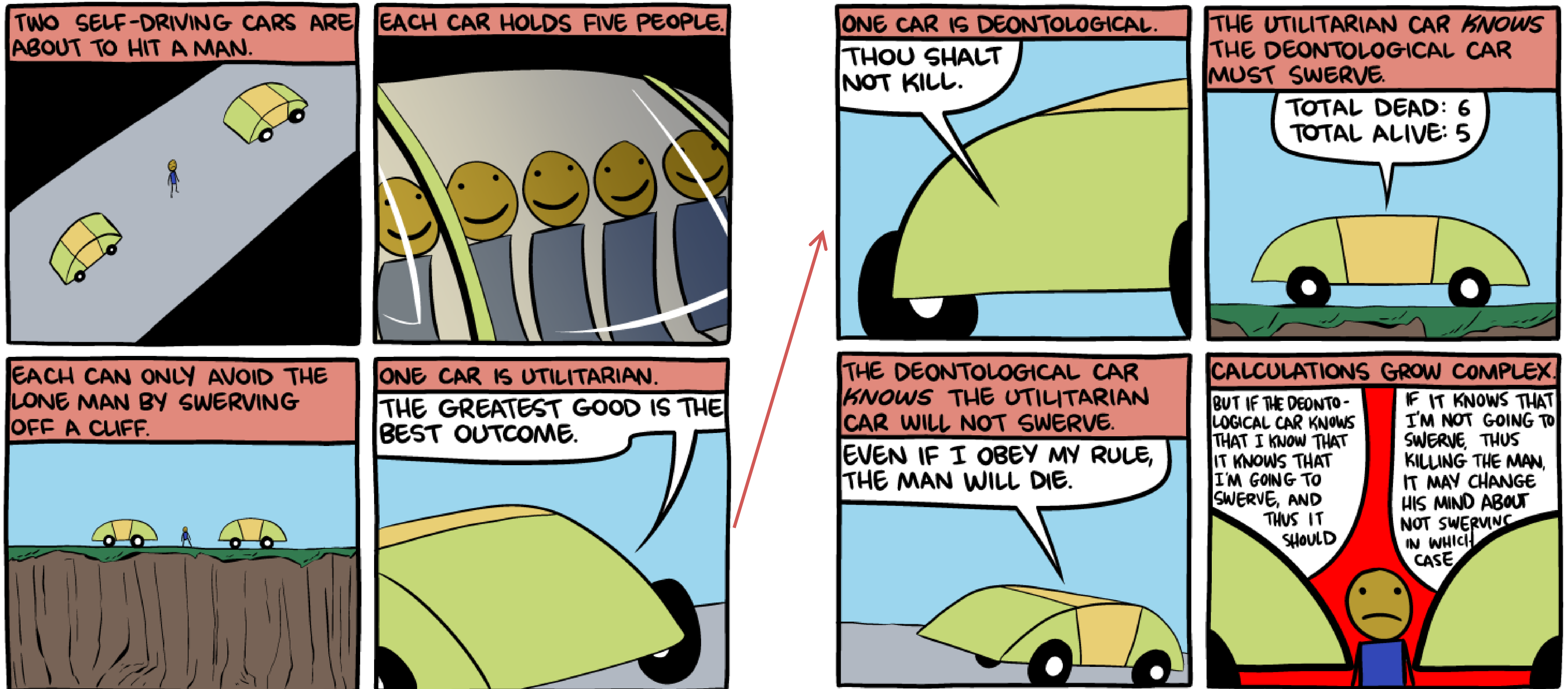
- c) Let users pick the moral character of their cars out of a set of ethically permissible characters (for instance, let them decide some trolley cases when they order the car, so some will get consequentialist cars, some deontologist cars, and so on – possibly reflecting their owners moral character).

Again: How to formalize, algorithmize,  
implement ethical theories in the first place?

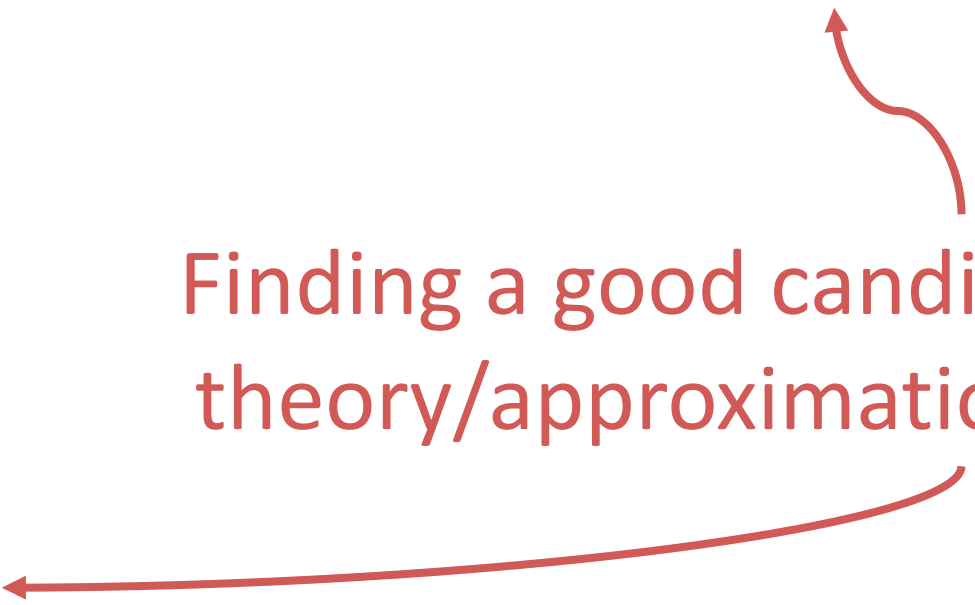
And consider the following:



# MACHINE ETHICS: THE CENTRAL PROBLEM



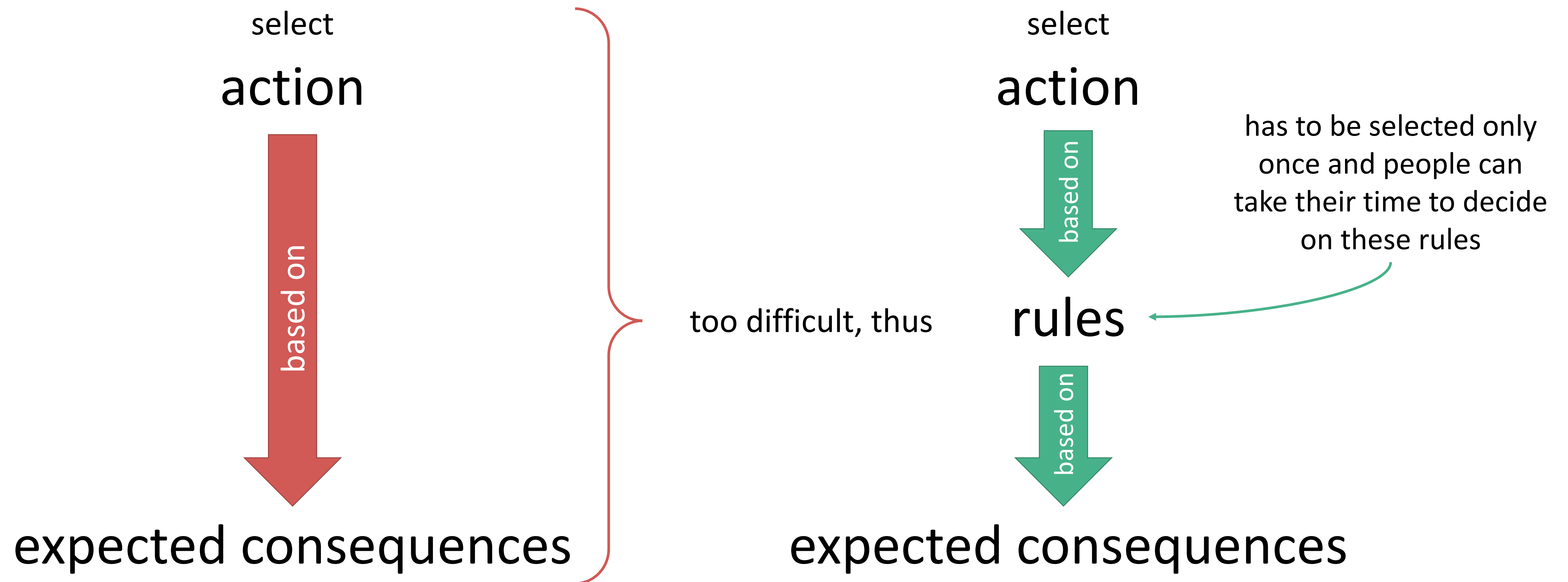
## Possibility 2: Dodge the Problem

- d) Maybe normative theories are too sophisticated for such a simple autonomous system: Too much moral sensitivity and general deliberation abilities are required. Then we need a normative sub-theory for a class of very restricted agents of this kind.
  - e) Use a reasonable approximation of a suitable normative theory, e.g. a restricted form of rule consequentialism.
- Finding a good candidate for such a sub-theory/approximation is very hard, too.
- 

But still something has to be done...

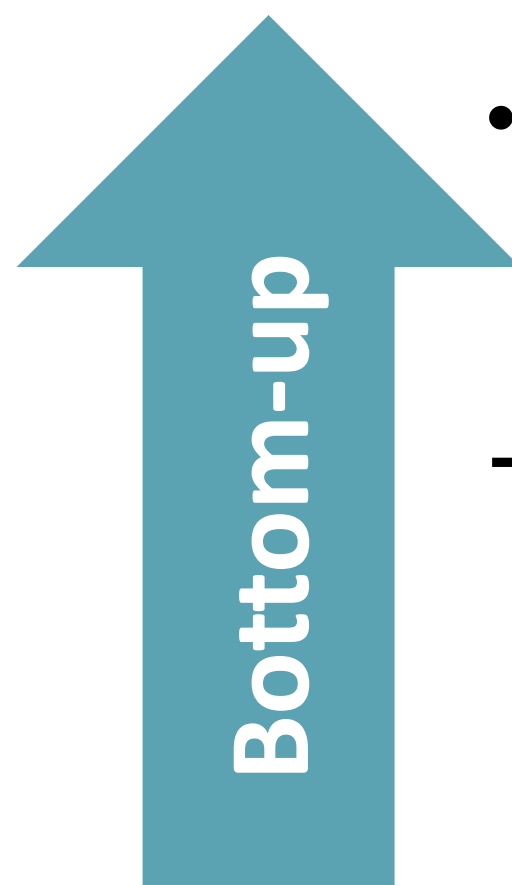
# RULE CONSEQUENTIALISM

**Idea:** select set of rules that (of all sets of rules) will result in the best world

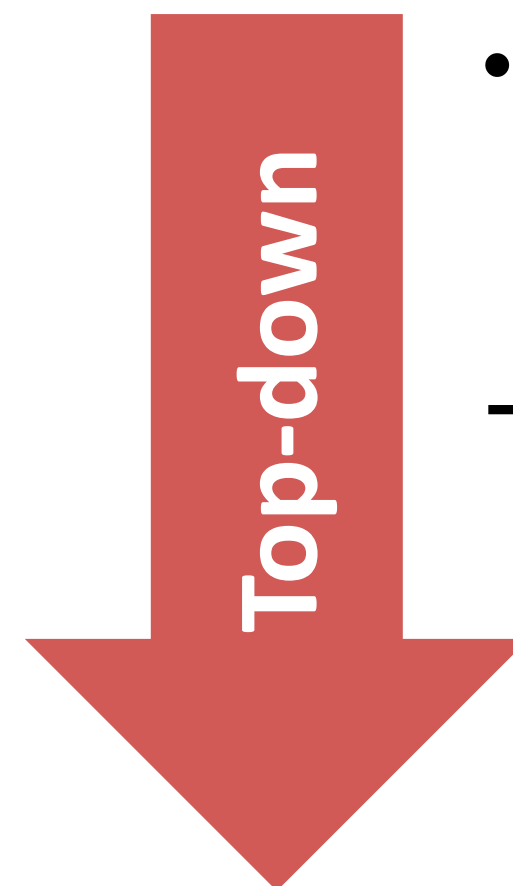


## A Useful Distinction

*Top-down ('deductive')* vs. *bottom-up ('inductive')* approaches  
Roughly:

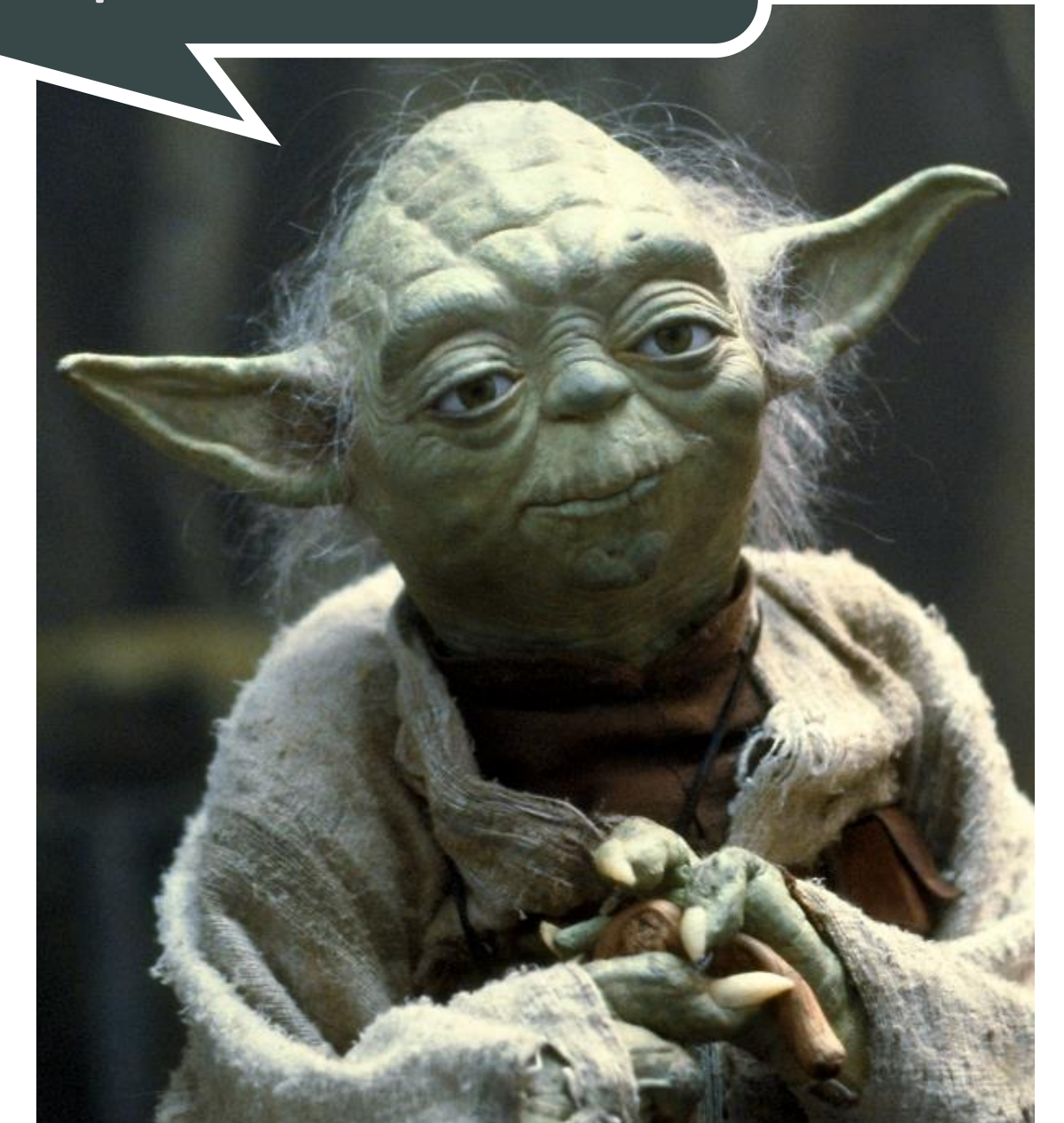


- derive rule from paradigmatic examples
- perceive situation, compare to previously seen situations, and decide what to do



- give rules explicitly
- perceive the situation, compare with rules, and decide what to do

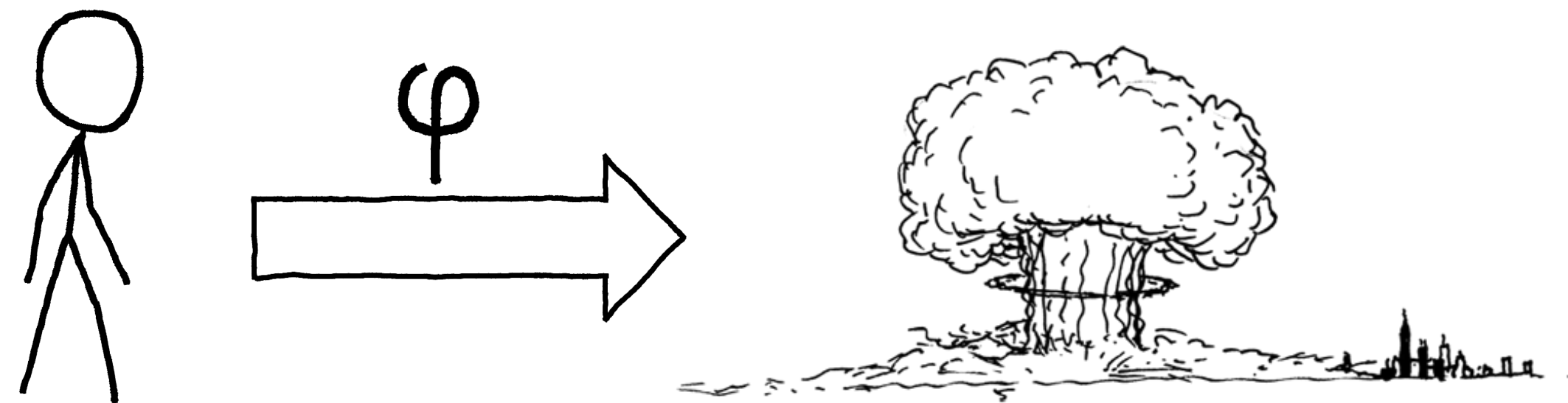
Hard to answer this question is.



*What kind of approach do Jedi need?*



## Normative Theories – Bottom-up or top-down?



An **agent** performs an **act**. An **act** has **consequences**.

### Virtue Theories

- basically, this seems to be *bottom-up only*

### Deontological Theories

- system of principles and rules that must be applied properly → top down approach
- application may involve bottom-up aspects, namely in extracting the relevant information from a situation
- then we need a *hybrid system*

### Consequentialist Theories

- top-down module for the calculation and decision
- most likely a bottom-up module for the axiological part
- *hybrid system*

### Example (Historically Impactful & Useful 'Toy')

Isaac Asimov's three Laws of Robotics (actually a misnomer), originally from his story "Runaround" (1942), the short story collection „I, Robot“ (1950)

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



# TOP-DOWN vs. BOTTOM-UP APPROACHES

## Problems: Asimov's Laws of Robotics

How to know how to follow the rules?

### Information

E.g.: Does a robot necessarily know when he harms a human?

### Rule interpretation

E.g.: What is a human? What is humanity? What a robot? What is harming? What a conflict? Can there be exceptions in the spirit of the laws? How to interpret a human's orders?

cannot be solved by direct programming, but rather by some form of learning → bottom-up approach

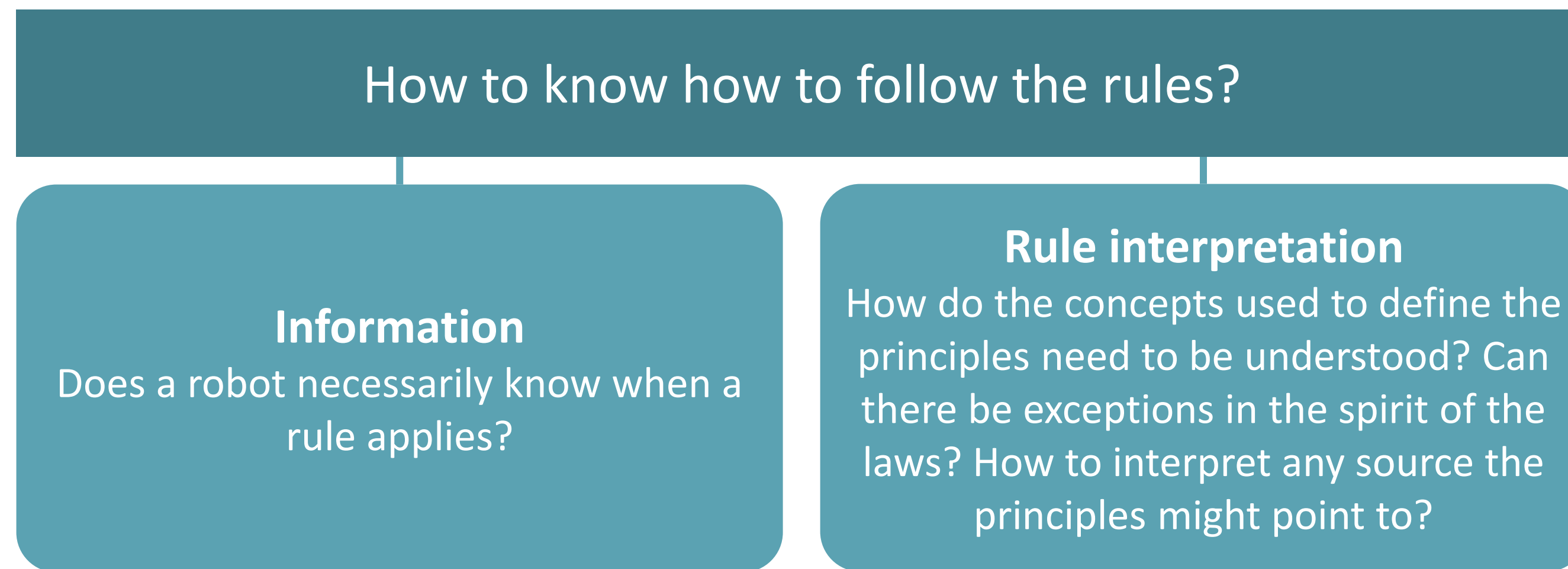
What if a rule cannot be followed?

since all options violate some clause

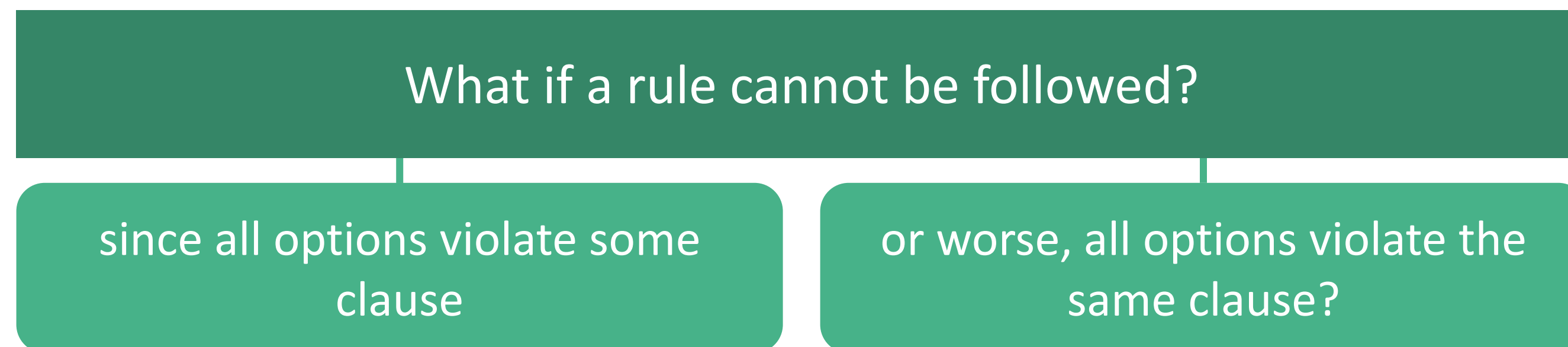
or worse, all options violate the same clause?

Needs serious interdisciplinary work between philosopher, roboticists and computer scientists → unclear which approach

## Problems: Rule Consequentialism



cannot be solved by direct programming, but rather by some form of learning → bottom-up approach



Needs serious interdisciplinary work between philosopher, roboticists and computer scientists → unclear which approach

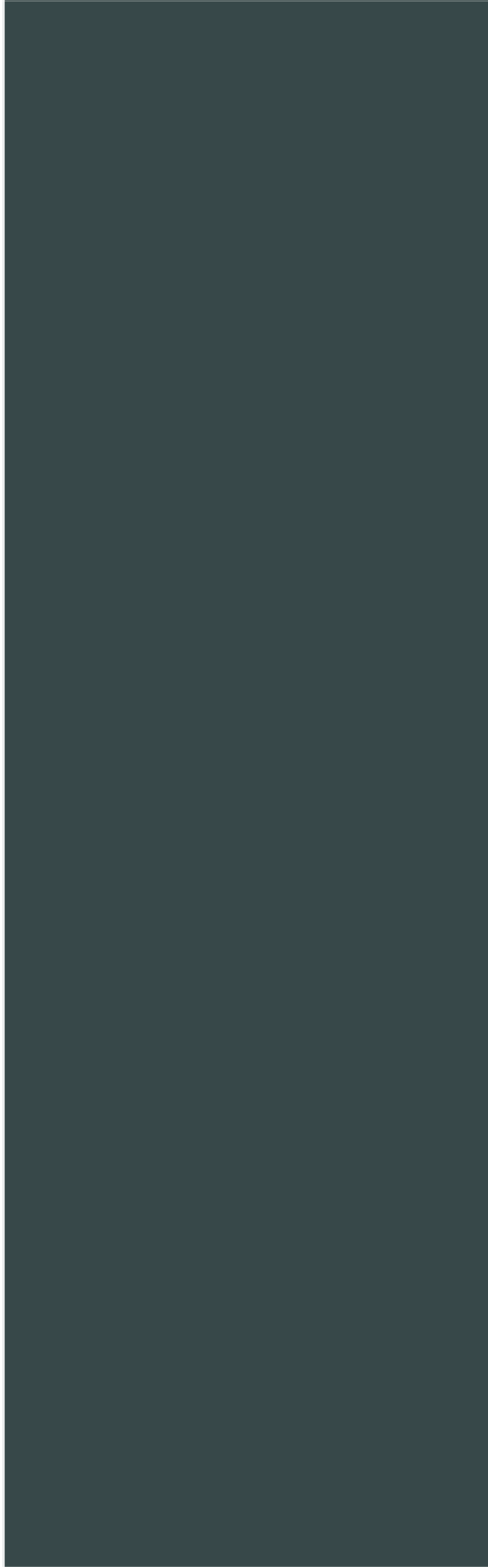
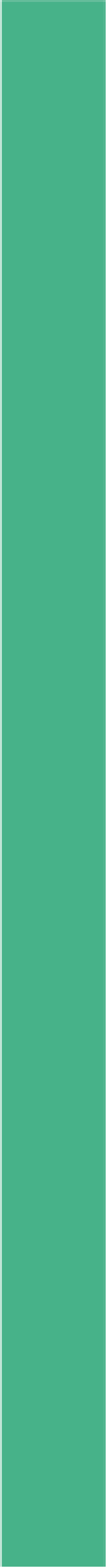
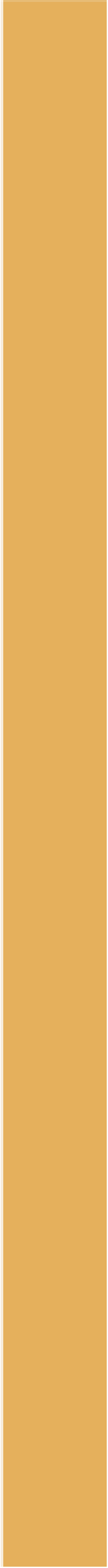


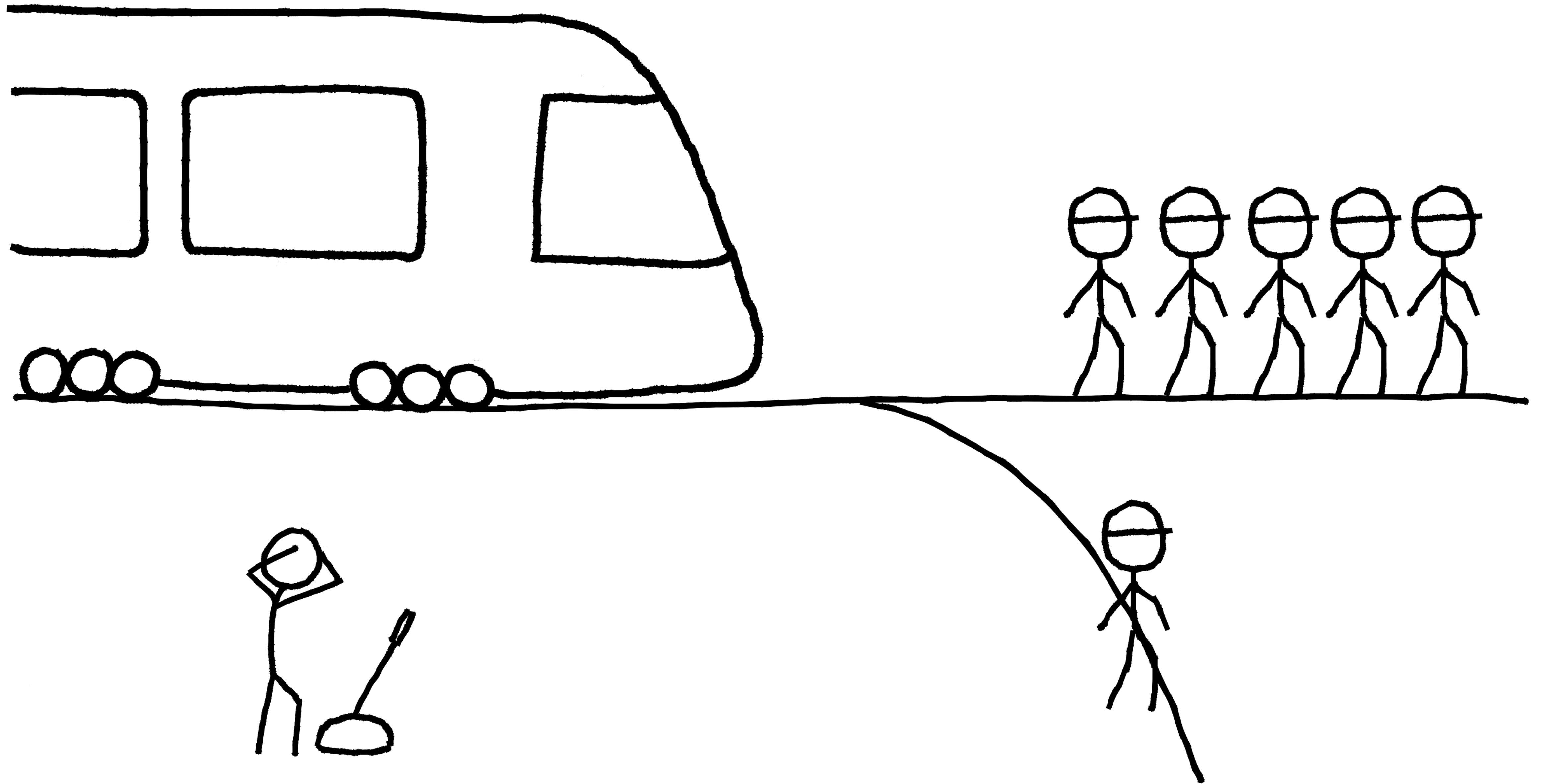


How to implement morality?



Basically, we don't know yet.





# Moral Courage



Be charitable. Reasonably weigh your options and their consequences.

Be courageous. Apply your knowledge of ethics.

Use your reasoning skills. Care!

Be open minded and roll back your beliefs if necessary.

t

h

a

n

k

y

o

u