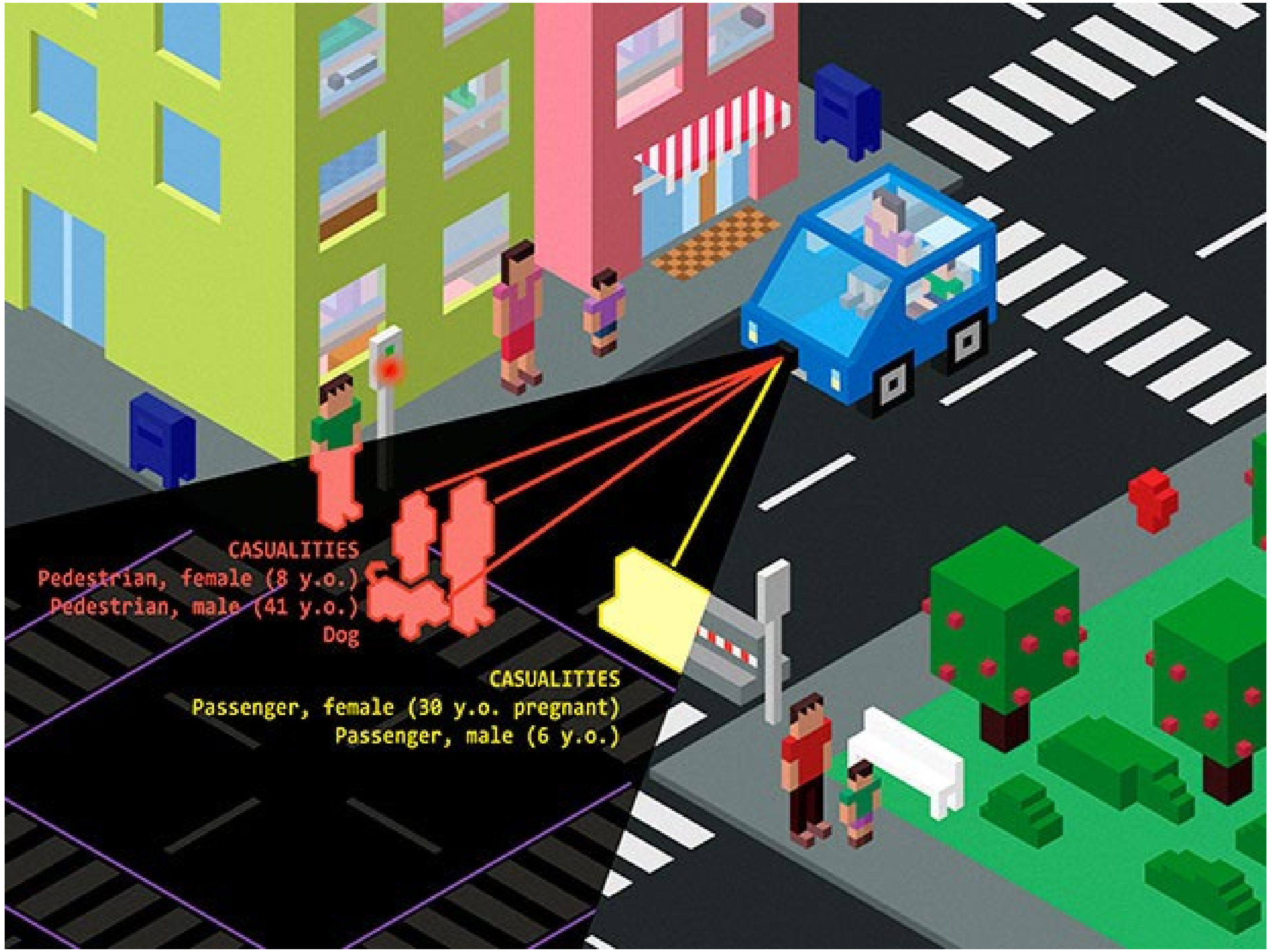
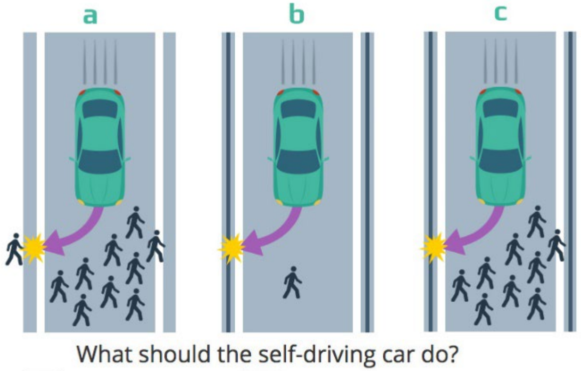
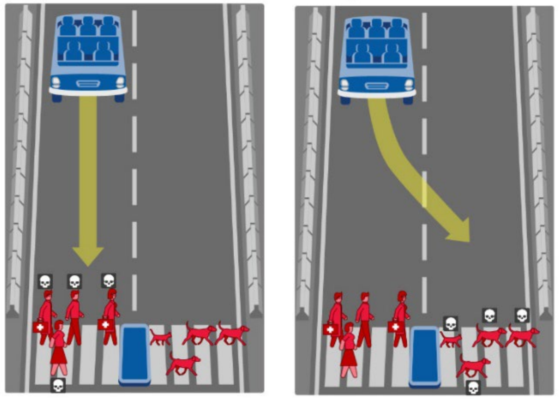

CASUALITIES
Pedestrian, female (8 y.o.)
Pedestrian, male (41 y.o.)
Dog
CASUALITIES
Passenger, female (30 y.o. pregnant)
Passenger, male (6 y.o.)





Moral machines

Philip J. Nickel, Associate Professor of Philosophy & Ethics

In this lecture

- Review: HCAI and other paradigms of human-AI interaction
- Human-AI interaction
    - Should humans adapt to robots and AI, or should robots and AI be adapted to humans?
- Identify what could be complex for assigning responsibility during an AI-human “handoff” (Mulligan & Nissenbaum 2020, Goldenfein et al. 2020).
- “Moral Machines”: Consider Goodall’s (2014) central argument, and Awad et al.’s (2020) contribution to a solution

Review

- Last session (Thursday 25 Sept) we discussed two value paradigms for human-technology interaction: Meaningful Human Control and Human-Centered AI
- These are contrasted with a Technological Autonomy Ideal

- I’d like you to be equipped to address the question “How should we automate” by thinking about these paradigms.

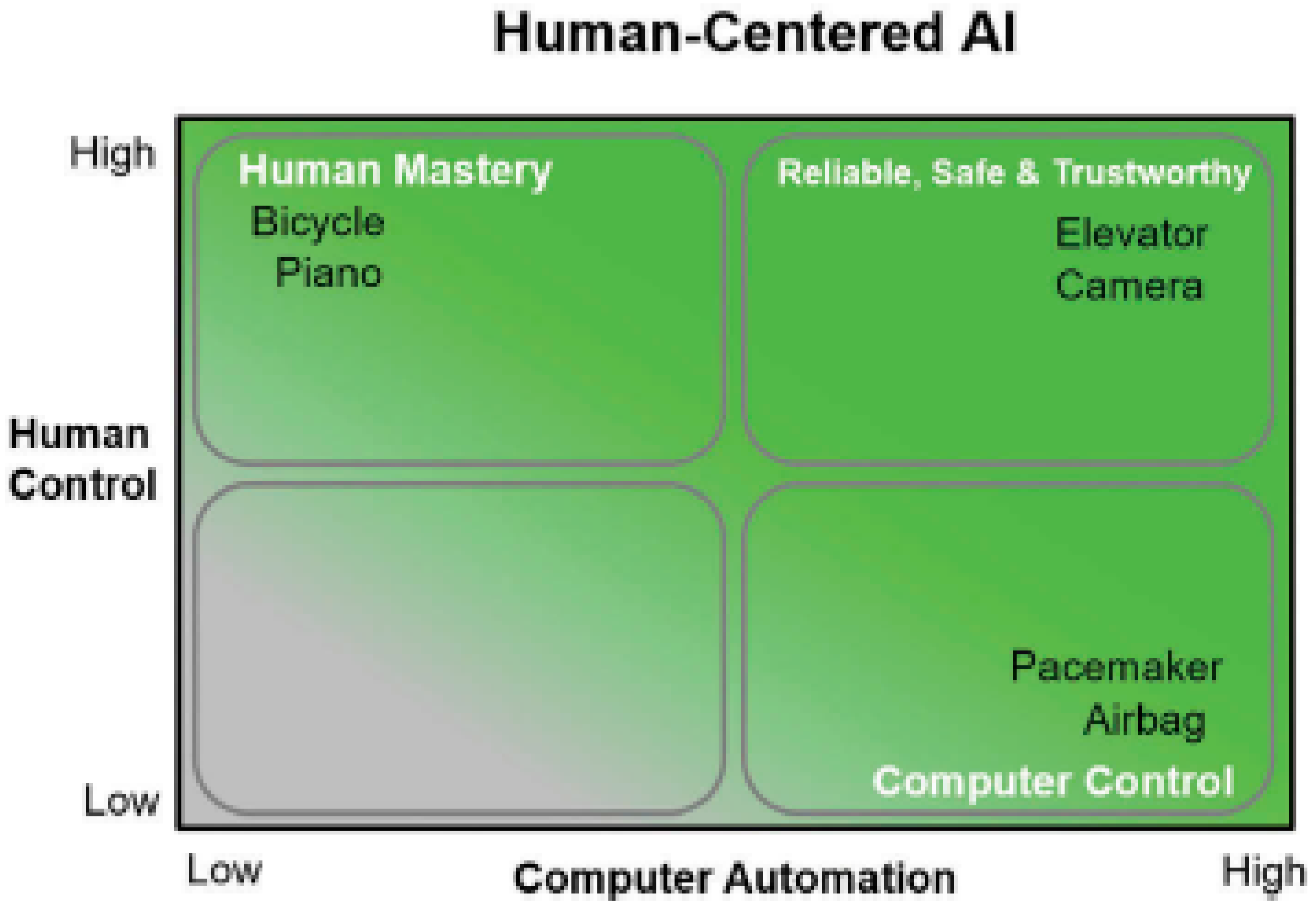


Figure 3. Regions requiring rapid action (high automation, low human control) and human mastery (high human control, low automation).

Additional examples:

- HR screening algorithm
- Warehouse inventory storage & retrieval system
- Smart piano

TAI & HCAI as normative ideals

- We should strive for maximal technological autonomy in technology design.

(the technological autonomy ideal [**TAI**])

Comparable to the view that we should adapt human nature to AI and robots.

- We should strive for high automation **together with** high human control in technology design.

(the Human Centered AI ideal in Shneiderman 2020)

Comparable to the view that we should adapt AI and robots to human nature.

Other ethical arguments regarding TAI vs HCAI (Goldenfein et al. 2020)

TAI

- In practice, TAI will require remote operators to monitor and be ready to take back the tasks, creating invisible, exploitative work (Goldenfein et al 2020).
- TAI entails displacement of tasks and loss of skills (Zoller 2017).
- TAI usually requires a tailored infrastructure, meaning it cannot integrate with existing built environments and would require an unacceptable break with existing practices (Goldenfein et al 2020).

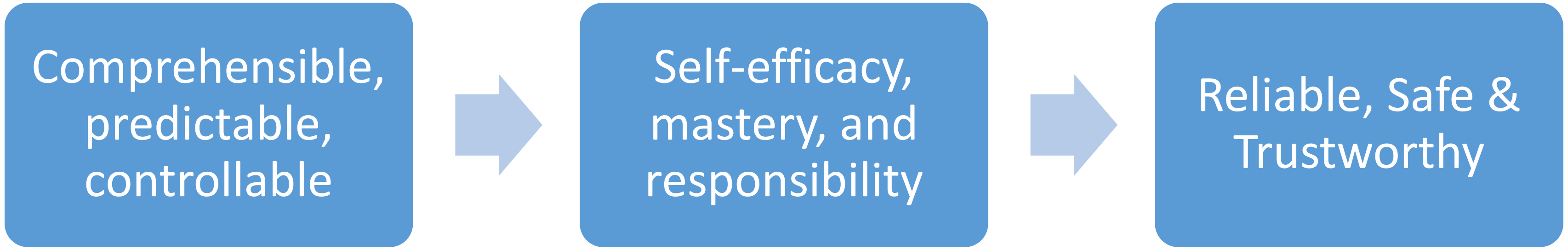
HCAI

- In practice, HCAI entails technology handoffs and questions about responsibility and liability between the human user and AI (Goldenfein et al 2020).

A “handoff” means a moment at which primary control transfers from one entity to another (for example, from a human to a drone’s automated mode, or vice versa).

Schneiderman’s claims in favor of HCAI

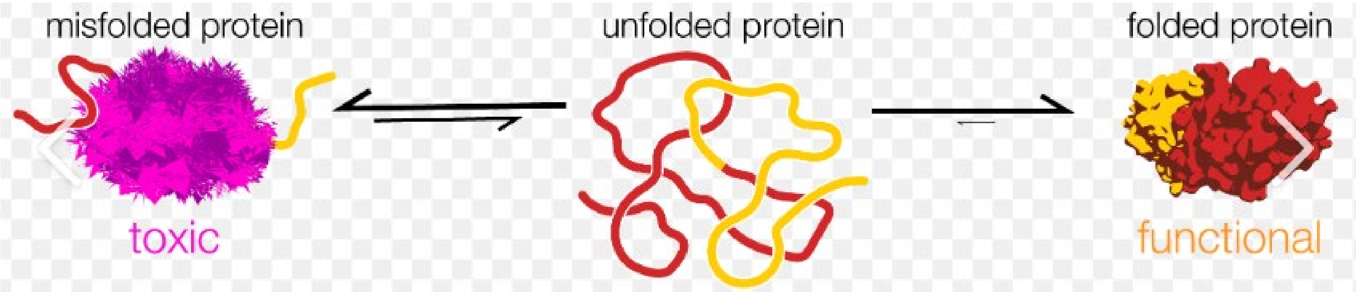
- *According to Schneiderman, HCAI promotes the values of reliability, safety, and trustworthiness.*
  - This is an empirical claim. To test it, we would need to have a clear baseline of fully automated systems for comparison. The isolated cases (e.g., Boeing 737-300 Max) mentioned by Schneiderman are not sufficient.
  - Are there a priori reasons why HCAI would be expected to have higher realization of these values?


Comprehensible, predictable, controllable
Self-efficacy, mastery, and responsibility
Reliable, Safe & Trustworthy

Schneiderman’s picture of HCAI’s relation to values

Apparent counterexample to HCAI

- “While AlphaFold can assume significant tasks previously done by human scientists (i.e., determining protein structures) this should positively impact, or at least have a neutral effect, on task integrity if it allows scientists to re-focus their work efforts on other important aspects of their broader goal of curing diseases. However, there remain risks to AI being used in this way. Continuing with this example, if scientists have trained for many years to do the experimental work that AlphaFold can now do more quickly and accurately, this generates significant risks for their ability to exercise their full capacities, demonstrate their mastery, and utilise the skills they have invested years in developing to reach their full potential” (Bankins & Formosa 2023).



Zooming out: value paradigms and AI

The claims of MHC, HCAI and related value paradigms

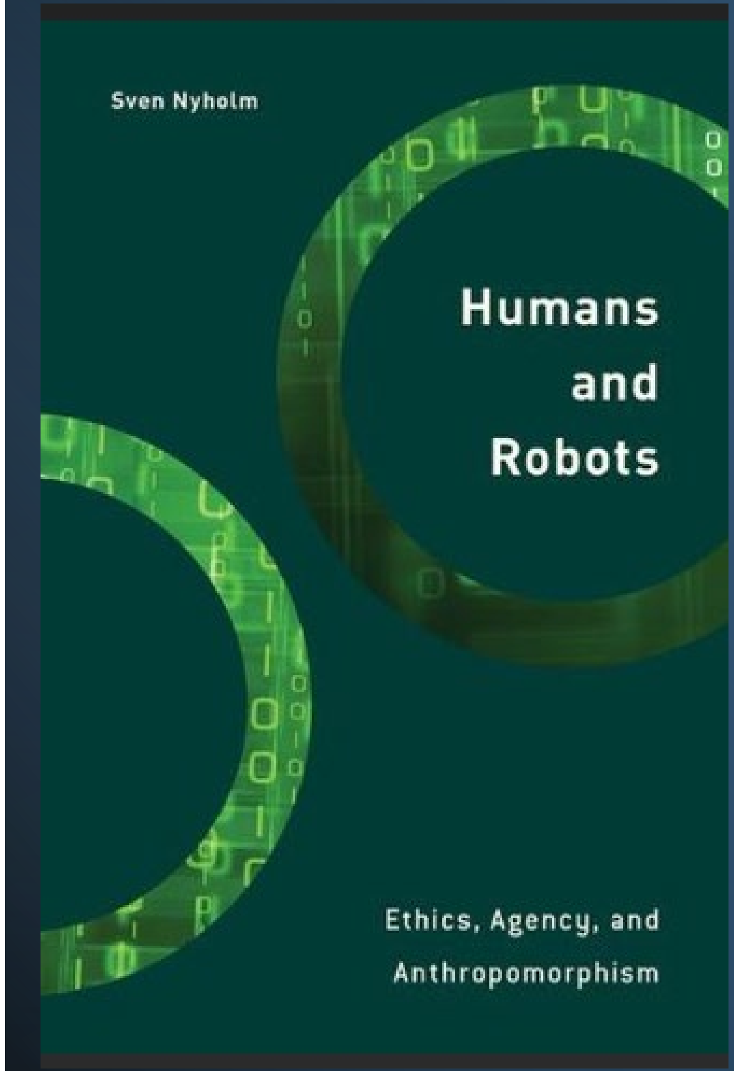
High levels of automation can be combined with human contributions, in a way that allows us to achieve the following values in AI and robotics:

- Responsibility (MHC)
- Trustworthiness (HCAI)
- Amplification of human work (Bankins & Formosa 2023)
- “Hybrid Intelligence”: Synergy of human and machine intelligence (Akata et al. 2020)

I: two modes of human-AI interaction

Nyholm’s argument (2020)

- Human nature is such that human interactions with AI and robots cause harm and risk.
- We should avoid harm and risk.
- Therefore, we should adapt AI and robots to human nature, or we should adapt human nature to AI and robots.



Nyholm’s argument (2020)

- Human nature is such that human interactions with AI and robots cause harm and risk.
- We should avoid harm and risk.
- Therefore, we should adapt AI and robots to human nature, or we should adapt human nature to AI and robots.

→ Which of the four principles does this relate to?

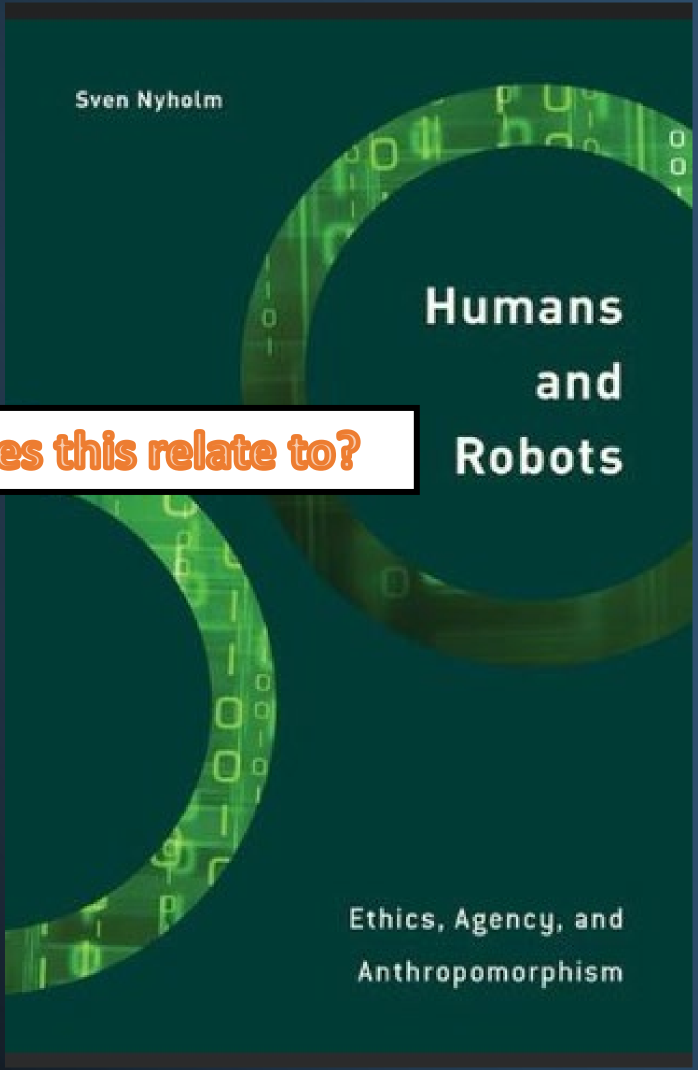
Nyholm’s argument (2020)

- Human nature is such that human interactions with AI and robots cause harm and risk.
- We should avoid harm and risk.
- Therefore, we should adapt AI and robots to human nature, or we should adapt human nature to AI and robots.

→ Which of the four principles does this relate to?

Examples:

- A robot is made to look, speak, and react like a human.
- A human follows a robot around as it packs delivery orders, fixing the problems that it causes or cannot deal with.
- A human engages in extensive prompt engineering to get quality natural language outputs from ChatGPT.



Some cases supporting premise one

“Human nature is such that human interactions with AI and robots cause harm and risk.”

- Uber’s self-driving car kills a pedestrian.
- A person walks into a manufacturing robot’s arm radius and is severely injured.
- An AI-based algorithm causes people to be wrongly accused of insurance fraud.

4 conditions on acceptable risk:

van de Poel & Royakkers (pp. 228-234):

1. Do the benefits outweigh the costs?
2. Availability of alternatives
3. Fair distribution of risks and benefits
4. Informed consent

Adapting Nyholm’s argument

(1) Human nature and the nature of AI are such that in some cases, without adaptation of one to the other, human interactions with AI and robots cause **unacceptable** harm and risk.

(2) We should avoid **unacceptable** harm and risk in those cases.

(3) Therefore, in those cases we should adapt AI and robots to human nature, or we should adapt human nature to AI and robots.

…

(N) In general, we should adapt AI and robots to human nature in such cases.

Nyholm 2020: Two options in "mixed traffic":

Option 1: make self-driving cars drive like humans

Option 2: seek means of making humans drive more like self-driving cars

What would be examples?

Nyholm’s conclusion, implications

“We should adapt AI and robots to human nature, or we should adapt human nature to AI and robots.” For example, we might choose between:

Assisted driving

Vs.

Fully automated vehicles

II: AI-human handoffs

Suppose we take a more human-centered approach…

When AI and humans interact, this can involve “handoffs” in which tasks are transferred from human to robots or vice versa.

Responsibility attribution can be difficult in such cases.

Handoffs in assisted driving

“In Level 3 automation, the driver is not required to remain attentive but must be available to take control of the vehicle within a certain amount of time after the receipt of an alert” (Goodall 2014: 60, referring to NHTSA Road Vehicle Automation Levels).

Task displacement and handoffs

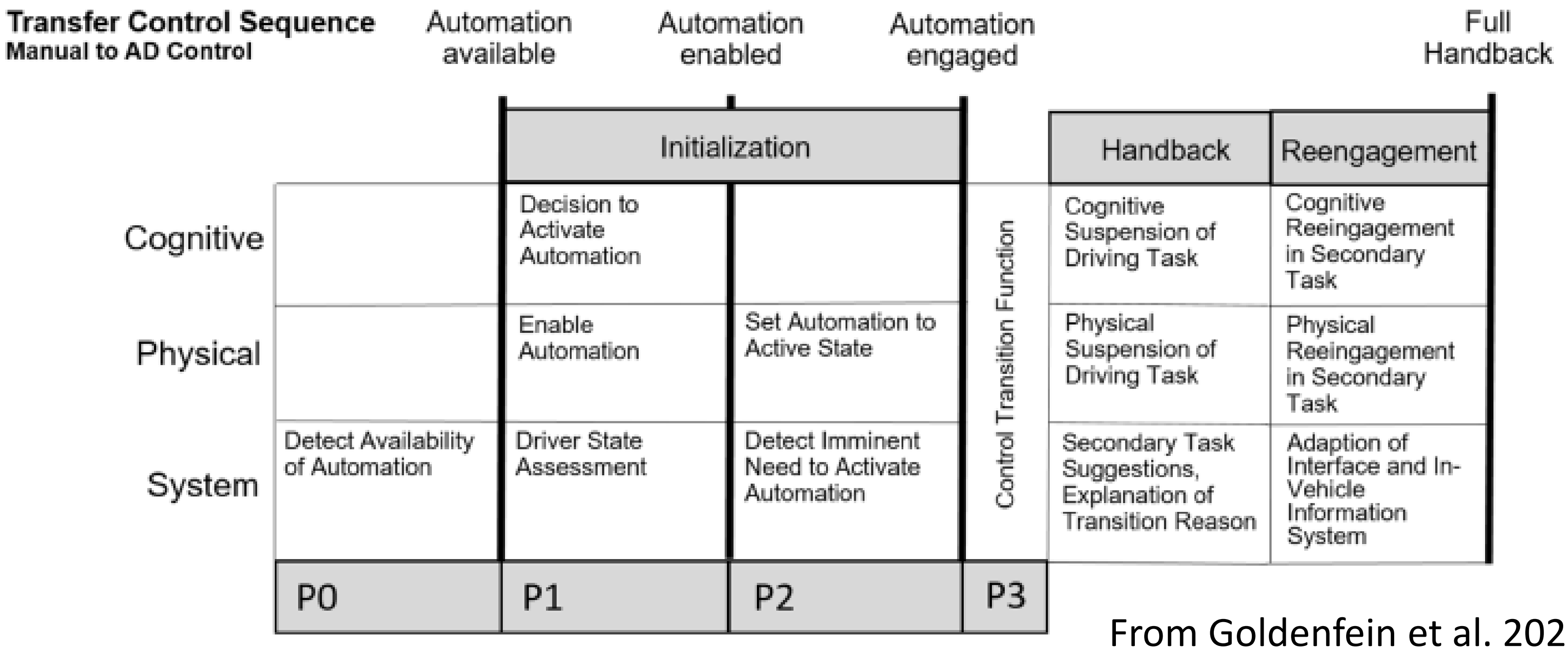
In automation, a mechanical system takes over a task from a human. For example, doing laundry is largely taken over by a machine in Dutch houses. Putting an item on a calendar can be largely taken over by Siri. Let’s call this *task displacement*.

In cases of task displacement where humans and automation cooperate on tasks, *handoffs and handbacks* occur when control shifts from human to automation or vice versa (Goldenfein et al. 2020).

Responsibility and handoffs

1. During handoffs and handbacks (i.e., from driver control to vehicle control of braking), it is difficult to determine who or what is in control of a task.
2. Responsibility for a task depends on an attribution of control (definition)
3. Therefore, during handoffs it is difficult to determine responsibility for a task.

Table 2: Handback Taxonomy (SAE 3114 Phases on the Bottom, Additional Phases on Top)



Transfer Control Sequence
Manual to AD Control

Automation available | Automation enabled | Automation engaged | Full Handback

		Initialization			Handback	Reengagement
Cognitive		Decision to Activate Automation		Control Transition Function	Cognitive Suspension of Driving Task	Cognitive Reengagement in Secondary Task
Physical		Enable Automation	Set Automation to Active State		Physical Suspension of Driving Task	Physical Reengagement in Secondary Task
System	Detect Availability of Automation	Driver State Assessment	Detect Imminent Need to Activate Automation		Secondary Task Suggestions, Explanation of Transition Reason	Adaption of Interface and In-Vehicle Information System
	P0	P1	P2	P3		

From Goldenfein et al. 2020: 854

Part III: Moral “decisions” in autonomous machines

See Goodall 2014 & Awad et al. 2020

Now suppose we go for a more autonomous level of automation…

Machine ethics (def)

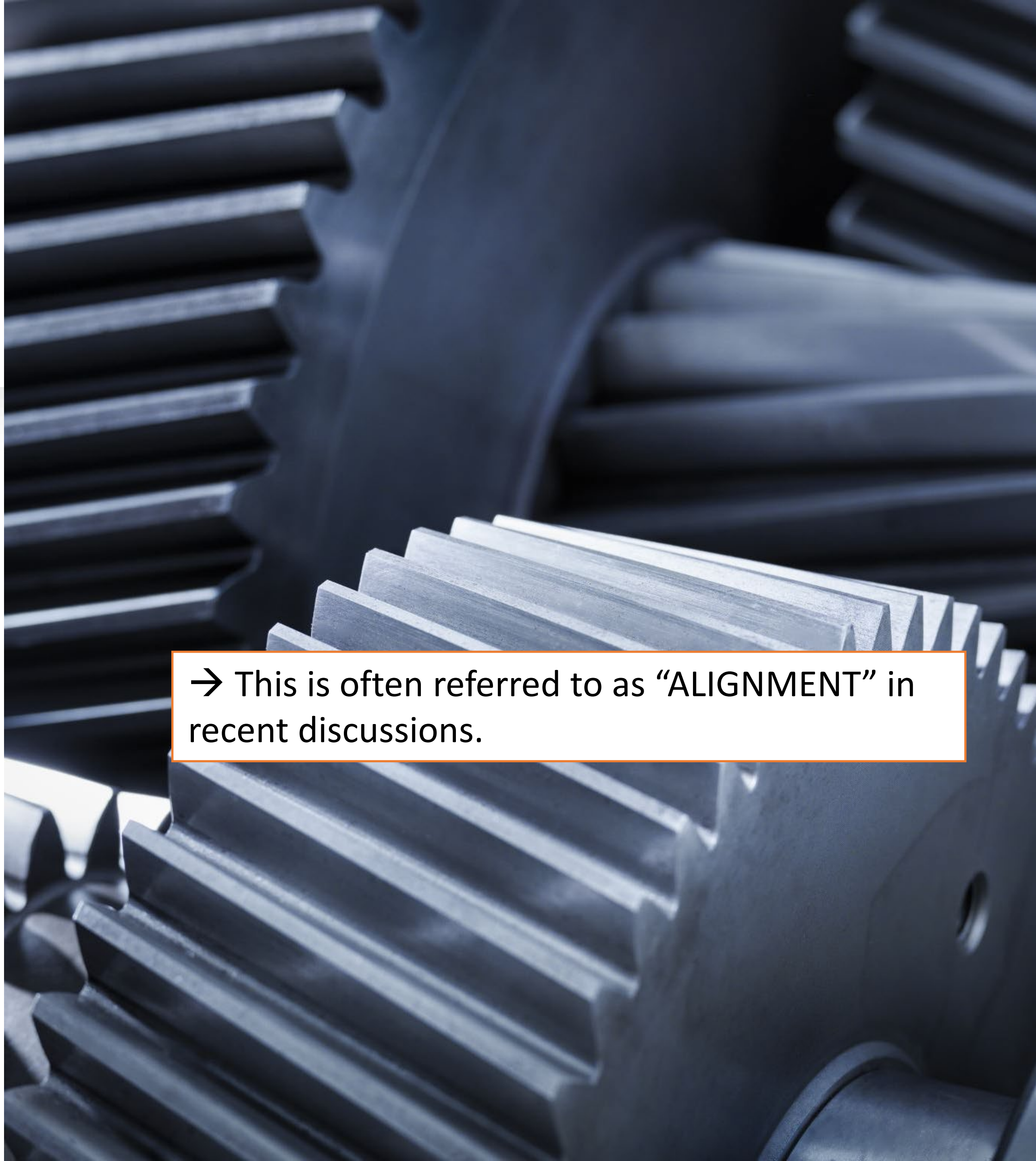
Machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. (Anderson and Anderson 2007: 15)

AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency. (Dignum 2018: 1, 2)

Machine ethics (def)

Machine ethics is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. (Anderson and Anderson 2007: 15)

AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency. (Dignum 2018: 1, 2)



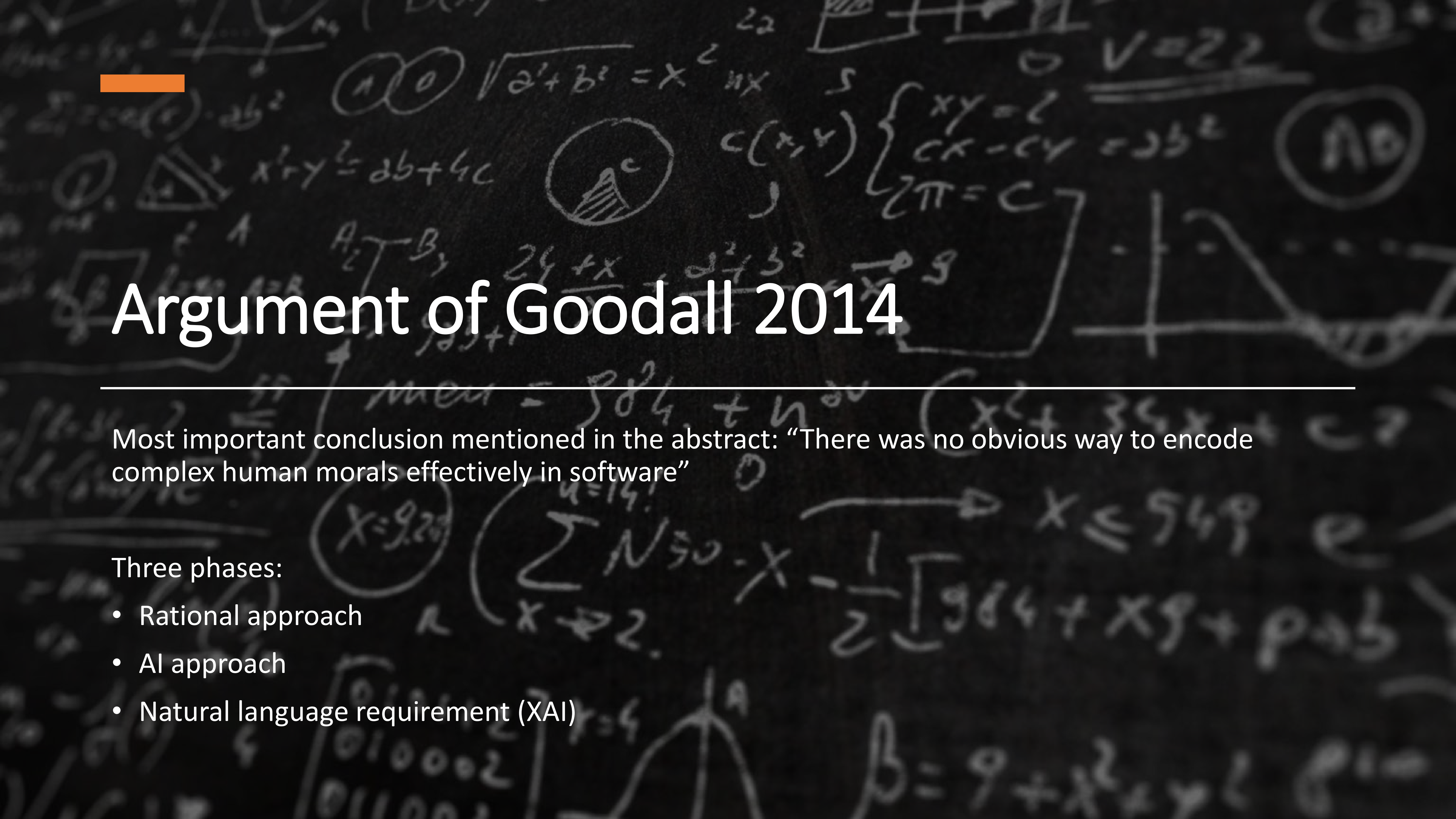
→ This is often referred to as “ALIGNMENT” in recent discussions.

Asimov’s law of robotics

First Law—A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law—A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law—A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



Argument of Goodall 2014

Most important conclusion mentioned in the abstract: “There was no obvious way to encode complex human morals effectively in software”

Three phases:

- Rational approach
- AI approach
- Natural language requirement (XAI)

Goodall: Phases in programming moral machines

1. Rational: Rule-based ethics or outcome-based ethics
    - Goodall: Rule-based ethics is too minimal to determine outcomes in unclassified cases or conflict cases, whereas outcome-based ethics recommends some morally bad decisions.
2. AI-based: Use training data so that AI can learn to make moral decisions.
    - Goodall: This would reproduce human shortcomings. Also, it would lack explainability.
    - Another issue: Where would the training data come from? Whose judgments would be taken into account, and how?
3. Natural language feedback (XAI) allows us to formulate the rules "post-hoc"

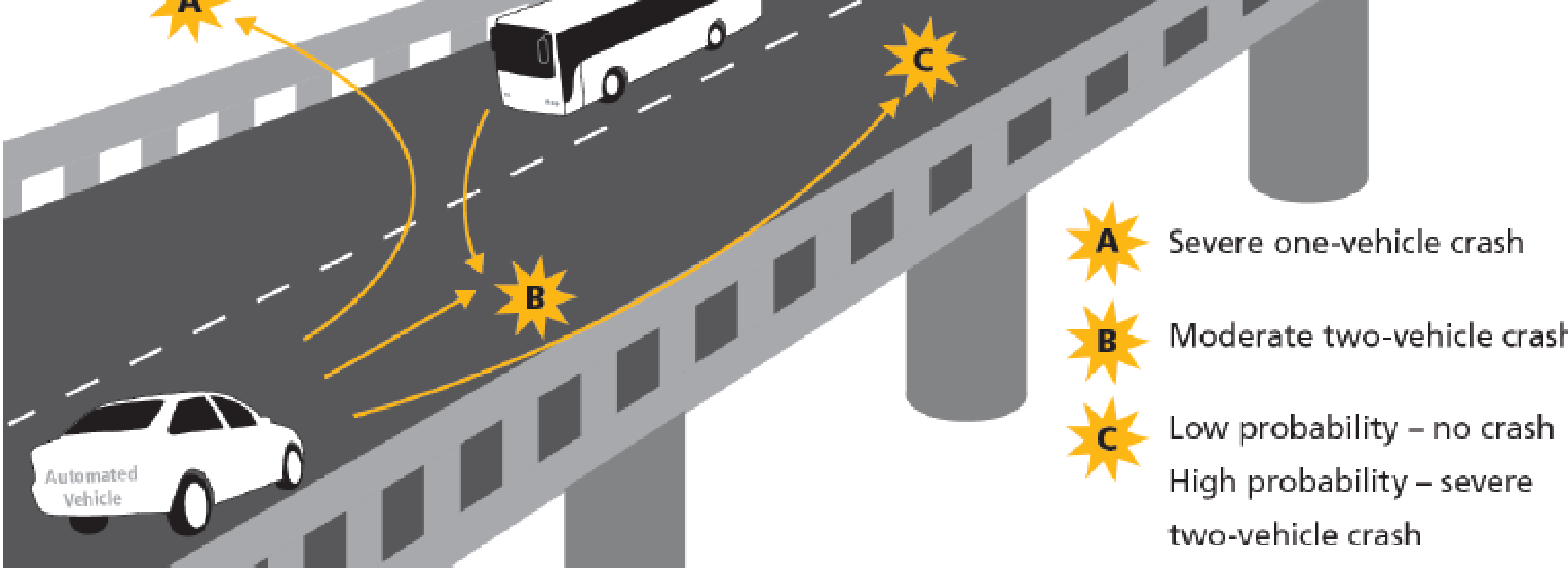

A
B
C
Automated Vehicle
A Severe one-vehicle crash
B Moderate two-vehicle crash
C Low probability – no crash
High probability – severe two-vehicle crash

FIGURE 1 Diagram of three alternative trajectories for an automated vehicle when an oncoming bus suddenly enters the vehicle's lane.

Moral aspect of crash decisions in mixed traffic

Moral Machines project

Why ethics rather than direct regulation (Awad et al. 2020, p. 49)?

- Black Box machine learning makes the principles opaque
- Technology changes faster than can be regulated
- Source of errors hard to trace to the technology or the source data

“All these factors make it especially
challenging to regulate the negative externalities
created by intelligent machines,
and to turn them into moral
machines. And if the ethics of machine
behavior are not sorted out soon, it is
likely that societal push-back will drastically
slow down the adoption of intelligent
machines.”

Argument for crowd-sourcing AV ethics (Awad et al. 2020, p. 50)

1. SITL ("society in the loop") is necessary for a dynamic consensus on the ethics of intelligent machines.
2. We need such a consensus if we wish to pursue the benefits of AVs and other full automation.
3. (We should wish to pursue the benefits of AVs and other full automation.)
4. Therefore, we should pursue SITL.

Some results of Moral Machines project

People prefer:

- Sparing the lawful over the unlawful
- Sparing humans over pets
- Sparing the greater number over the fewer
- Sparing younger humans over older

People are silent on some issues:

- Sparing vehicle occupants over non-occupants
- Killing as a means to saving lives
- Sparing the cautious over the uncautious

Where Awad et al think the crowdsourcing fits in

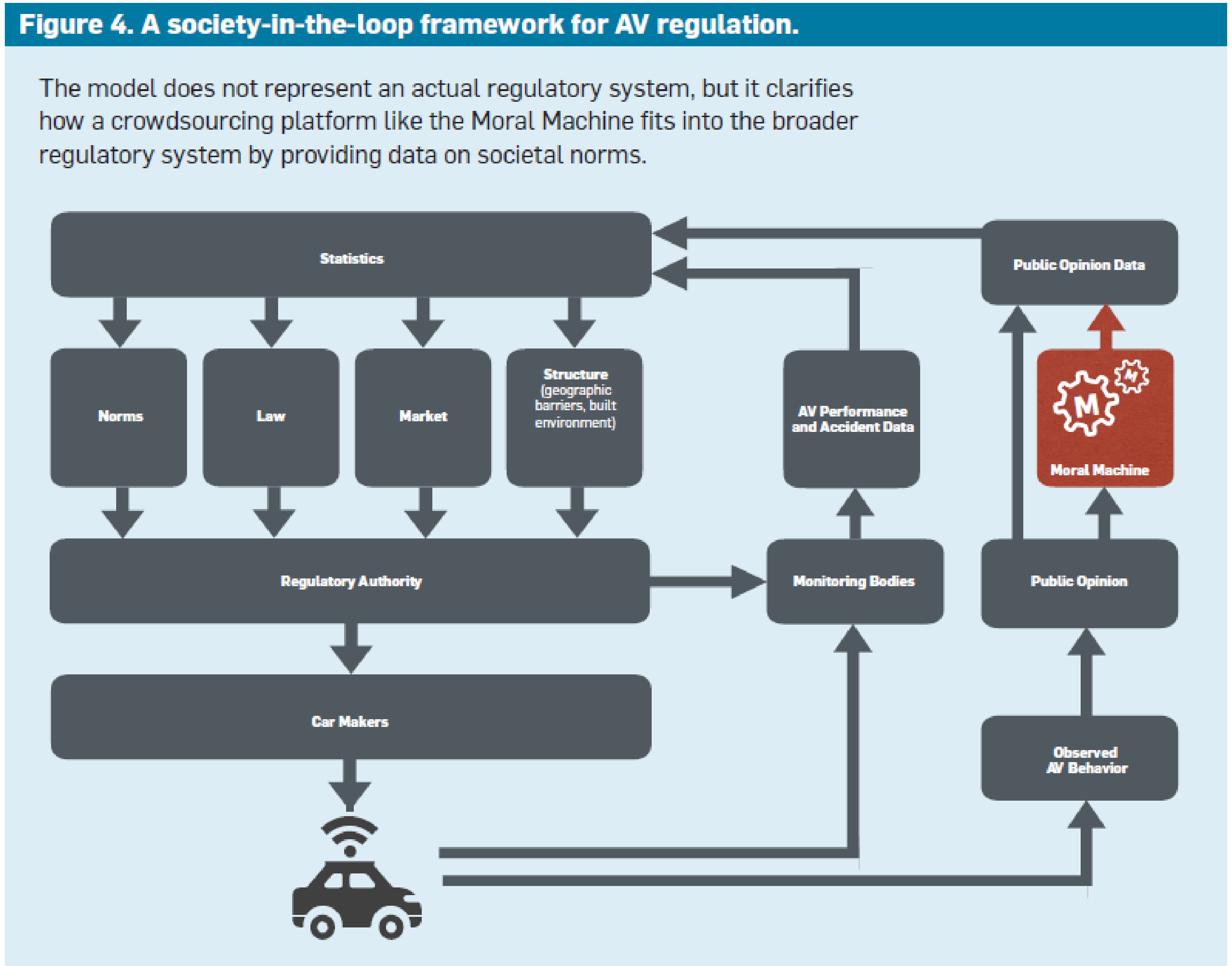


Figure 4. A society-in-the-loop framework for AV regulation.

The model does not represent an actual regulatory system, but it clarifies how a crowdsourcing platform like the Moral Machine fits into the broader regulatory system by providing data on societal norms.

Summing up

- There are broadly two ways that we can implement human-centered AI: one in which humans adapt to AI, and the other in which AI is adapted to humans.
- Human-centered AI may come with handoffs, where we face additional questions of control and responsibility.
- If we choose a more automation-centered approach (autonomous machines), we may face the complicated question of how to program moral aspects of decision making into the automation.

Tutorial today

- We will look at two kinds of argumentation patterns that will be useful for the exam:

- Rights-based argumentation
- Argumentation from necessary conditions