Two value paradigms for human-AI interaction

P.J. Nickel, P&E Group, p.j.nickel@tue.nl

2025-09-25 5ARC0

Paradigm 1: Meaningful Human Control

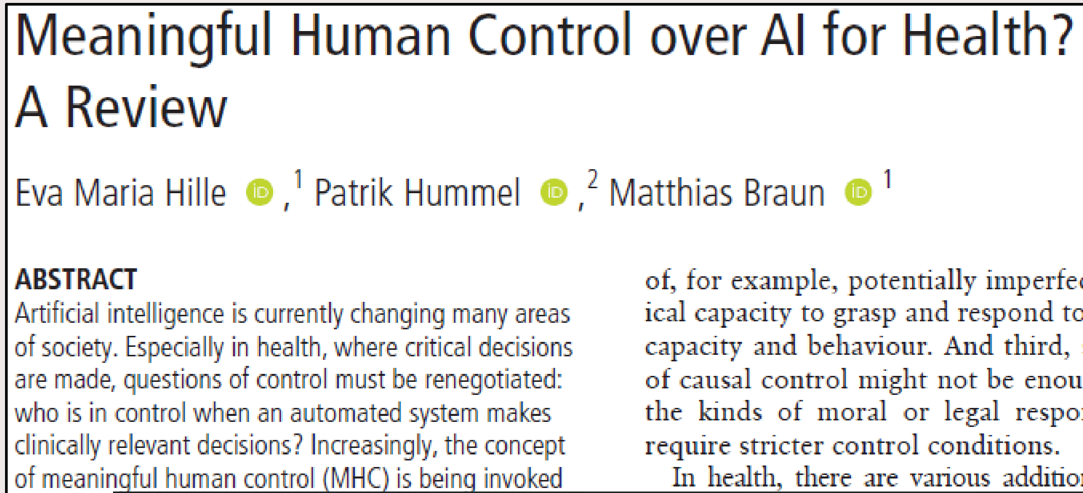
Background: why people care about control and responsibility

- One of the central concerns about a (semi-)autonomous system interacting with humans is that nobody can be reasonably held responsible for bad outcomes caused by the system. This has been called a ***responsibility gap***.
- In order to be held responsible, an entity must satisfy the following conditions:
  - Foreknowledge (“knew or should have known”)
  - Control/ causal influence
  - Freedom (“could have done otherwise”)

“Meaningful human control”

“humans not computers and their algorithms should ultimately remain in control of, and thus morally responsible for, relevant decisions” (Santoni di Sio & van den Hoven 2018).
This is more than merely a “human in the loop”.

“MHC is still an open concept that needs interpretation” (Hille, Hummel, & Braun 2023).



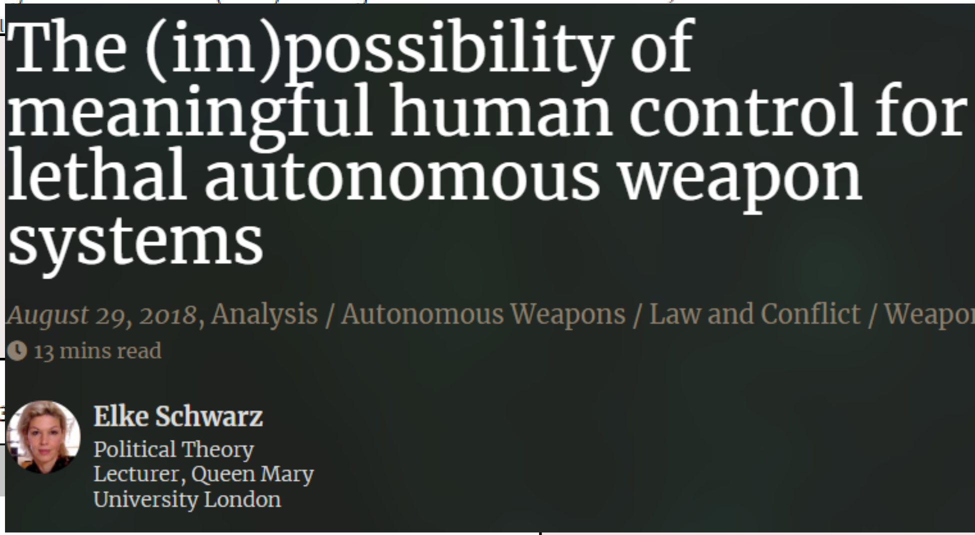
Meaningful Human Control over AI for Health? A Review

Eva Maria Hille [1], Patrik Hummel [2] Matthias Braun [1]

ABSTRACT
Artificial intelligence is currently changing many areas of society. Especially in health, where critical decisions are made, questions of control must be renegotiated: who is in control when an automated system makes clinically relevant decisions? Increasingly, the concept of meaningful human control (MHC) is being invoked for this pu

of, for example, potentially imperfe ical capacity to grasp and respond to capacity and behaviour. And third, of causal control might not be enou the kinds of moral or legal respo require stricter control conditions.

In health, there are various additio



The (im)possibility of meaningful human control for lethal autonomous weapon systems

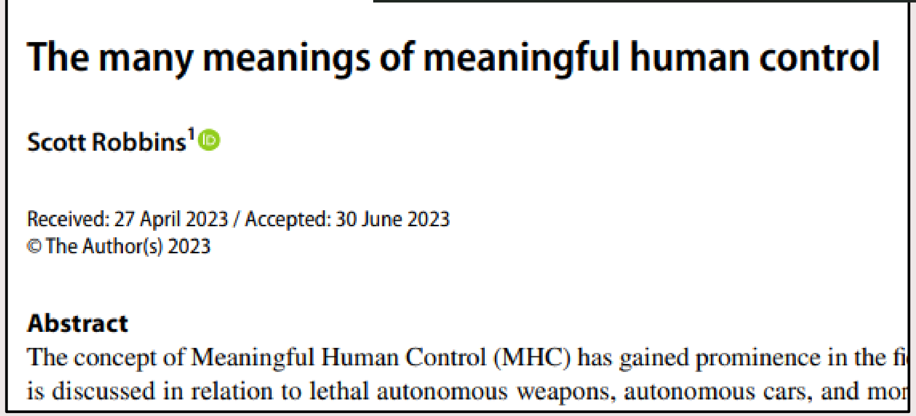
August 29, 2018, Analysis / Autonomous Weapons / Law and Conflict / Weapo

13 mins read

Elke Schwarz
Political Theory
Lecturer, Queen Mary
University London

AI and Ethics
https://doi.org/10.1007/s43681-023-003

ORIGINAL RESEARCH



The many meanings of meaningful human control

Scott Robbins[1]

Received: 27 April 2023 / Accepted: 30 June 2023
© The Author(s) 2023

Abstract
The concept of Meaningful Human Control (MHC) has gained prominence in the fi is discussed in relation to lethal autonomous weapons, autonomous cars, and mo

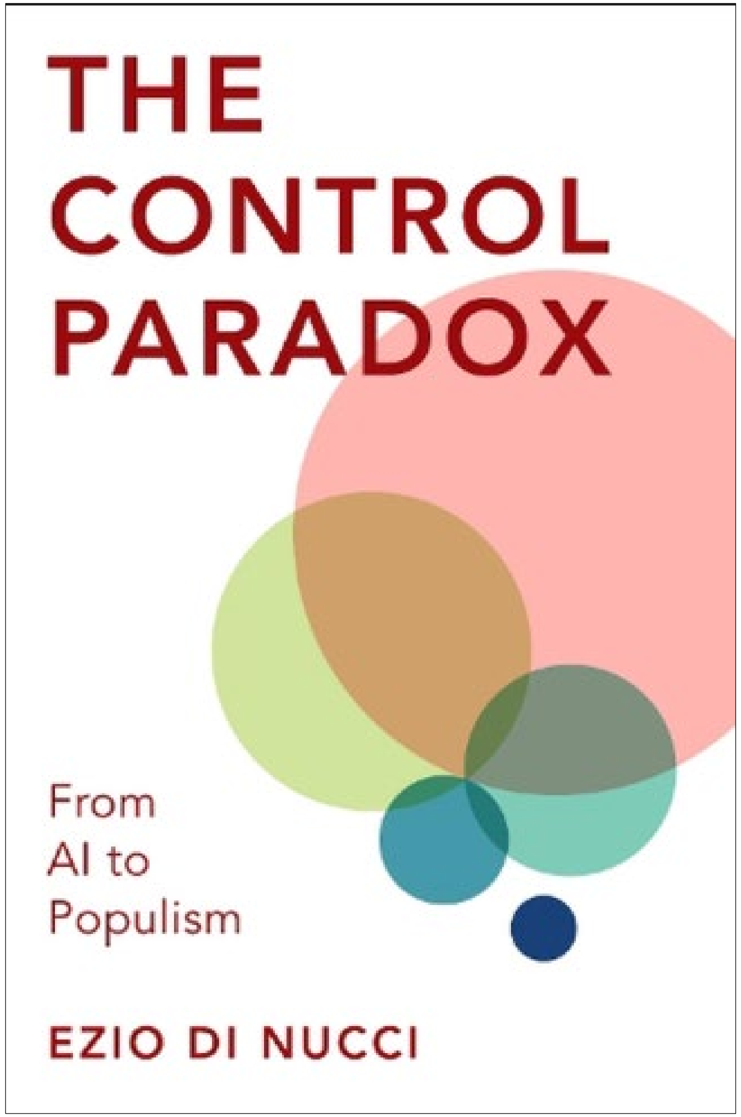
Concepts of control

Guidance control

Regulatory control

Being in control vs. controlling

Relationship of responsibility and control


THE CONTROL PARADOX

From AI to Populism

EZIO DI NUCCI

MHC

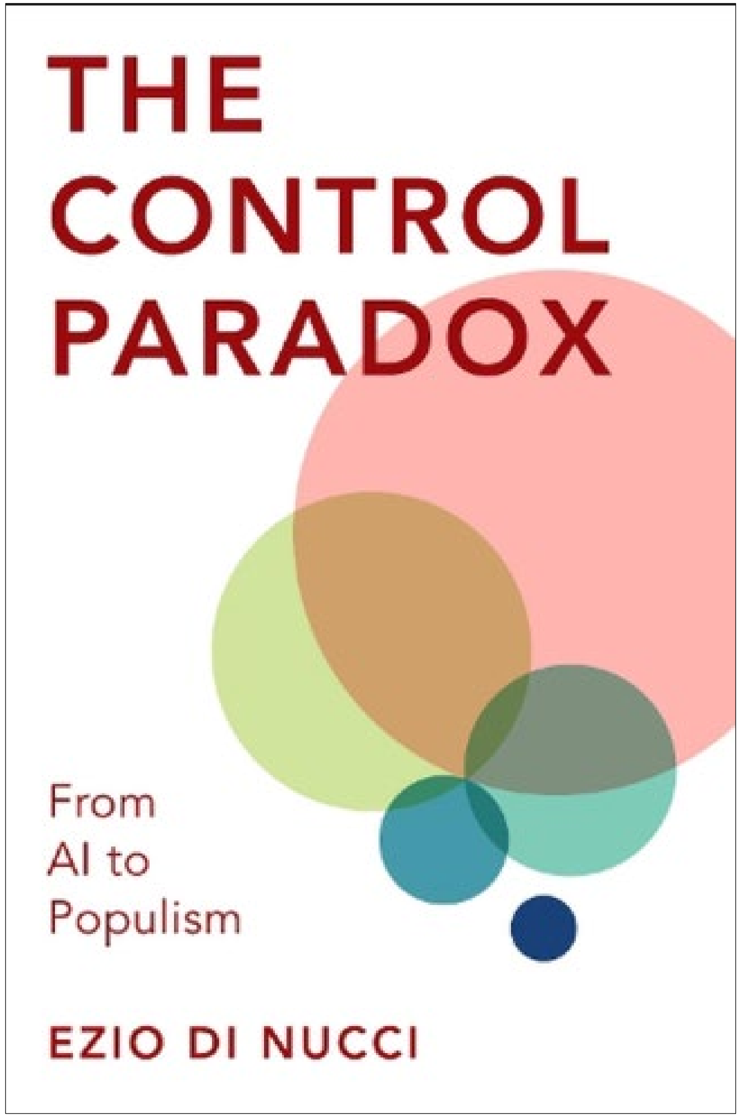
Concepts of control

Guidance control

Regulatory control

Being in control vs. controlling

Relationship of responsibility and control



THE CONTROL PARADOX

From AI to Populism

EZIO DI NUCCI

Two dimensions of control (Fischer & Ravizza)

Guidance control (the basis of “meaningful human control” in Santoni di Sio & van den Hoven 2018).

- Reason-responsiveness
- Each outcome is attributable to an agent

Regulative control

- The agent could have acted differently (freedom)

Two dimensions of control (Fischer & Ravizza)

MHC

Guidance control (the basis of “meaningful human control” in Santoni di Sio & van den Hoven 2018).

- Reason-responsiveness (tracking in MHC)
- Each outcome is attributable to an agent (tracing in MHC)

Regulative control

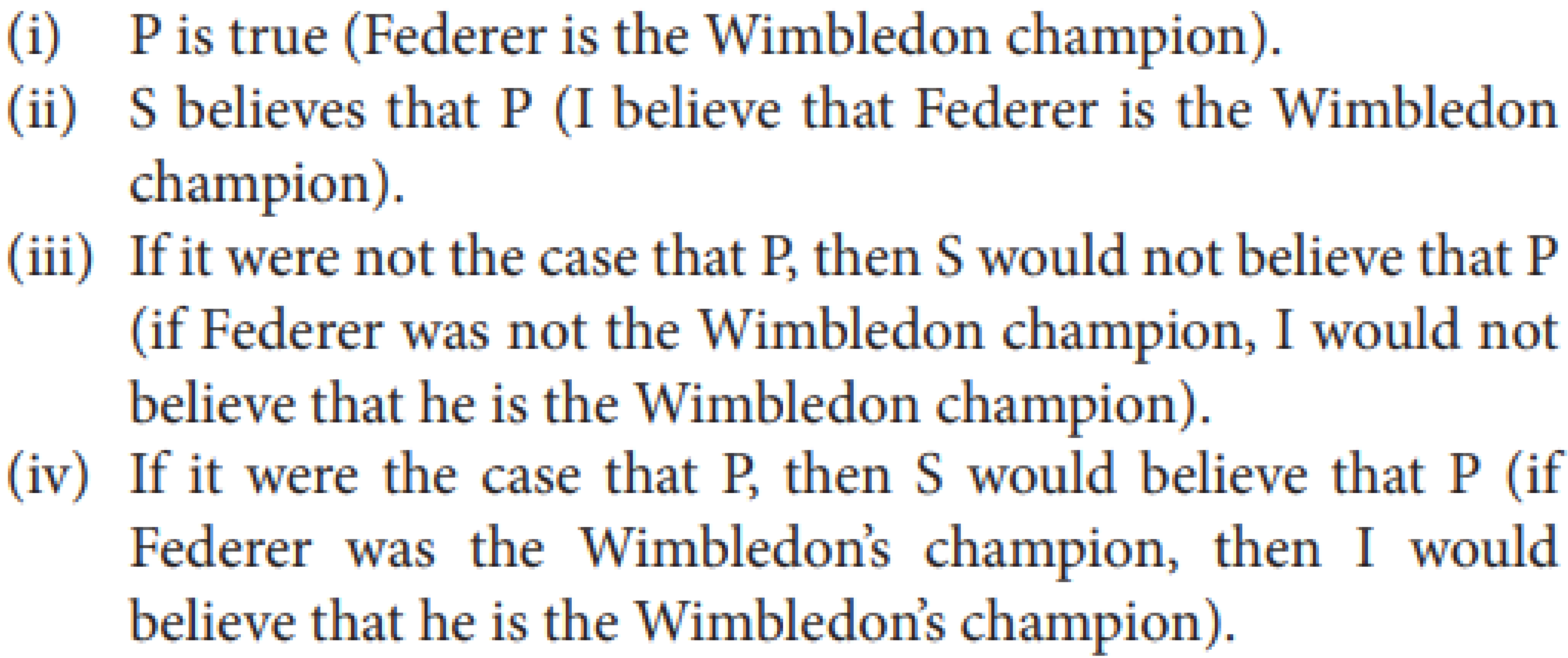
- The agent could have acted differently (freedom --- *not* part of MHC)

Meaningful human control

- Meaningful human control means guidance control.
- Regulative control (deeper freedom of action) is not required.
- Meaningful human control has two aspects: tracking and tracing. These are meant to “unpack” the idea of guidance control (and thus of MHC as a whole).

(Santoni de Sio & van den Hoven 2019)

Tracking condition



(i) P is true (Federer is the Wimbledon champion).

(ii) S believes that P (I believe that Federer is the Wimbledon champion).

(iii) If it were not the case that P, then S would not believe that P (if Federer was not the Wimbledon champion, I would not believe that he is the Wimbledon champion).

(iv) If it were the case that P, then S would believe that P (if Federer was the Wimbledon’s champion, then I would believe that he is the Wimbledon’s champion).

What does tracking mean for a drone?

Tracing condition

- “ownership” of an action or decision (Fischer & Ravizza)
- To make sense of “tracing” we need to think of ownership/ causal traceability in relation to automation cases, some of which are complex.

What does tracing mean for a drone?

Automation and responsibility

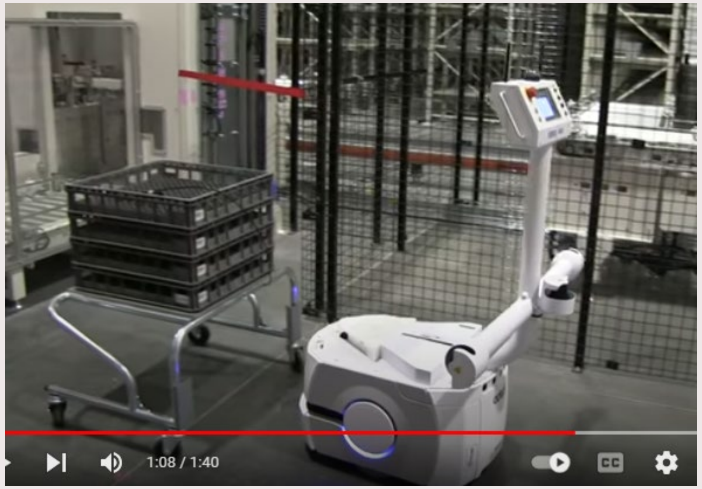
automation

full automation

human in the loop

human must override system in some conditions

system must override human in some conditions



https://www.youtube.com/watch?v=f0P1Ikyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

Automation and responsibility

automation

full automation

human in the loop

human must override system in some conditions

system must override human in some conditions

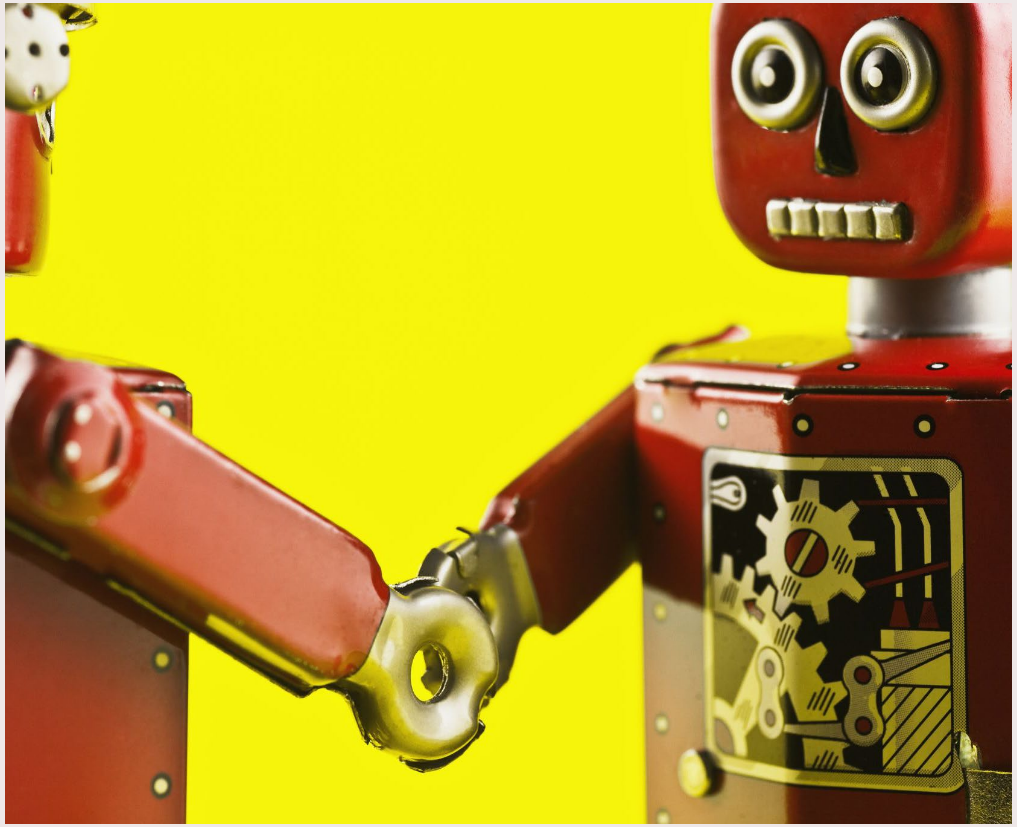
https://www.youtube.com/watch?v=f0P1Ikyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

Can a machine be responsible?

Apparently not: responsibility contains a freedom condition that machines do not satisfy.

In addition, there is a “retribution gap” (Danaher 2016): it does not make sense to blame or punish a machine.

One reason this might matter is that a machine cannot feel guilt, suffer, or correct its own behavior in response to blame.



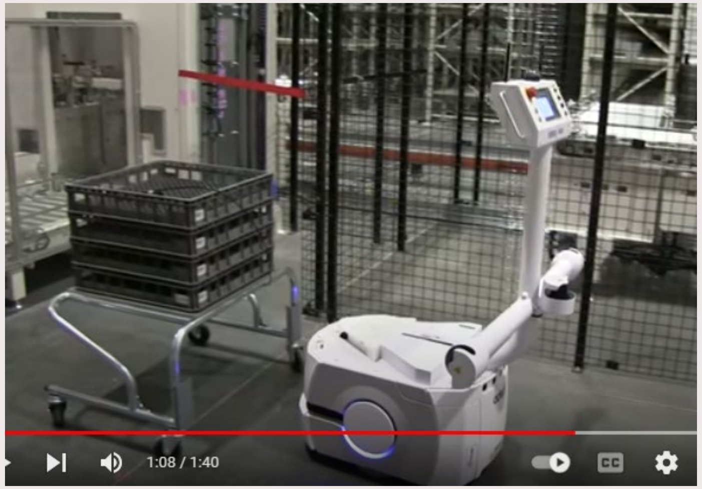
→ Therefore, meaningful *human* control is needed to deal with the responsibility gap.

Responsibility and Meaningful Human Control

- “Delegation of activity to smart machines is compatible with genuine responsibility attributions, even if the machine’s decisions are not free ” (Santoni di Sio & van den Hoven, p. 5).
- Humans can be responsible: “there is a difference between human actions and other natural events, and they thus claim that human agents can be legitimately seen as morally responsible for at least some of their actions” (ibid.).
- The main idea is that when guidance control is satisfied by automation, the humans who built and operate the system can be held responsible; and when it isn’t satisfied, something has gone wrong that should have been prevented during the design process.

Automation and responsibility

- automation
  - full automation
  - human in the loop
    - human must override system in some conditions
    - system must override human in some conditions



https://www.youtube.com/watch?v=f0P1Jkyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

Responsibility of primary and secondary operators

https://www.drivingschoolsupplies.ie/store/p1/dual-controls-for-driving-schools.html

Responsibility and design

“‘many of the choices made by designers can be seen as decisions about what should be delegated to a machine and what should be left to the initiative of human actors.”

Akrich 1992, P. 216

Responsibility and design

“many of the choices made by designers can be seen as decisions about what should be delegated to a machine and what should be left to the initiative of human actors.”

Akrich 1992, P. 216

Image: https://industrialscripts.com/stage-directions/

Automation and responsibility

automation

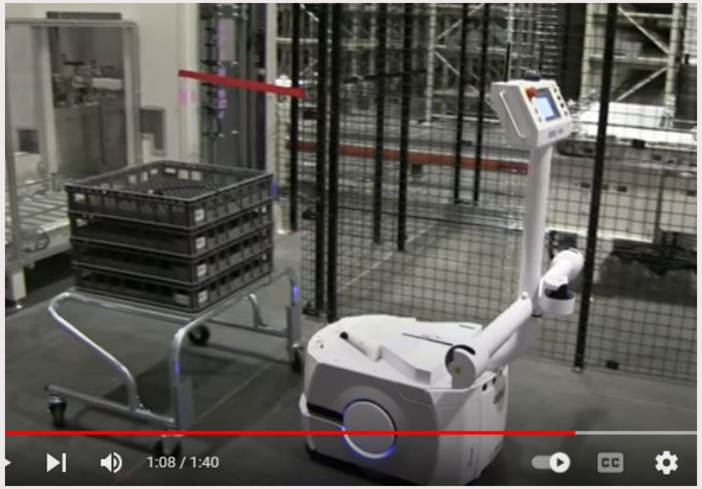
full automation

human in the loop

human must override system in some conditions

system must override human in some conditions

Scripts



https://www.youtube.com/watch?v=f0P1Jkyz8To
https://www.cnet.com/pictures/the-increasingly-autonomous-robots-of-war-pictures/

Solutions to the problem of responsibility and failure conditions: assigning forward-looking responsibility more strictly

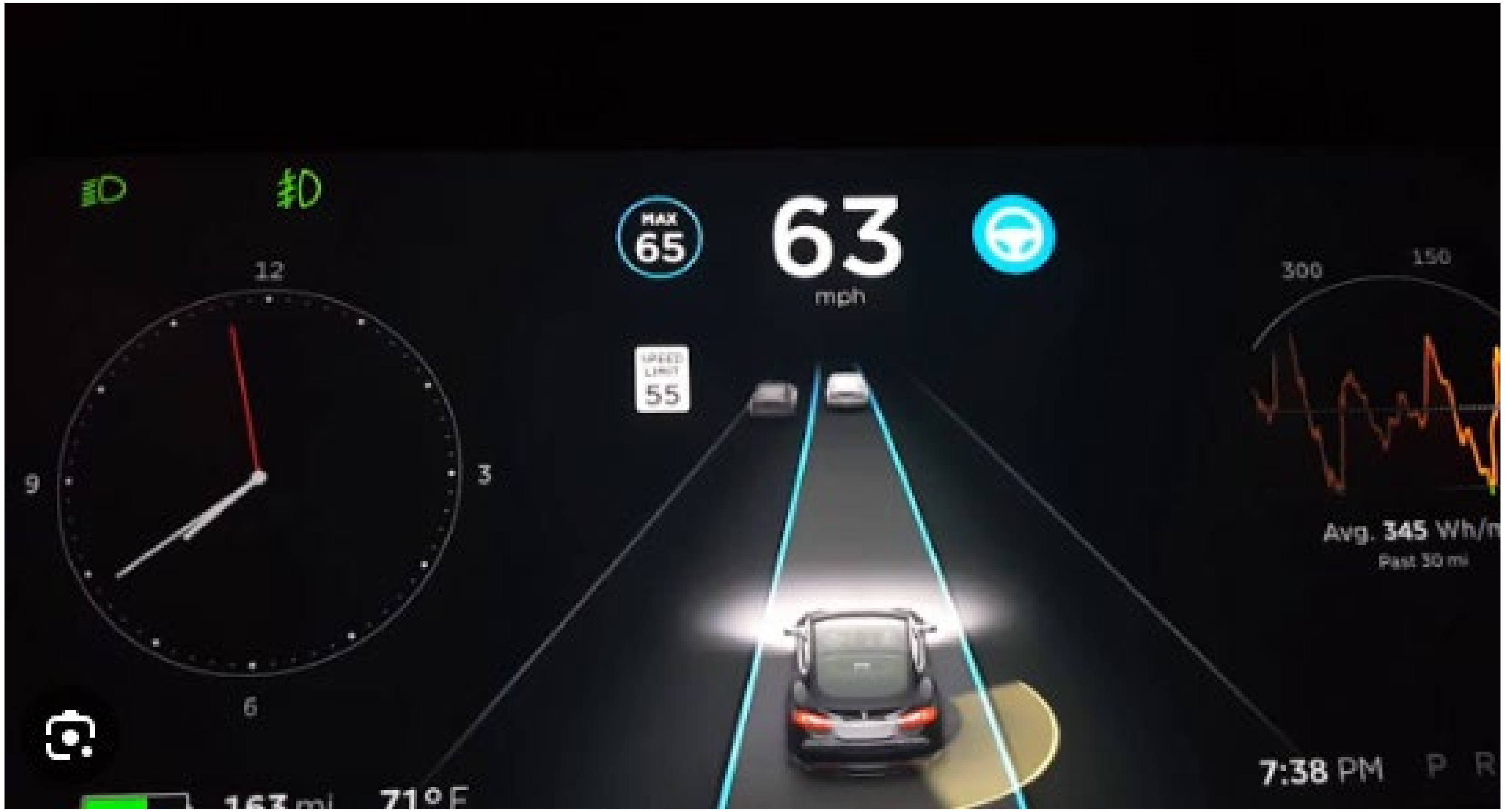
Can we write scripts more clearly and broadly to anticipate more failure conditions?

Can we insulate the automated system from potential failure conditions?

Can we enforce the scripts?

Implications for designers

“Designing for satisfying the tracing condition means ensuring that different human agents along the chain are technically and psychologically capable of complying with their tasks and are well aware of their responsibility for the behavior of the autonomous system.” (Santoni di Sio & van den Hoven 2018).



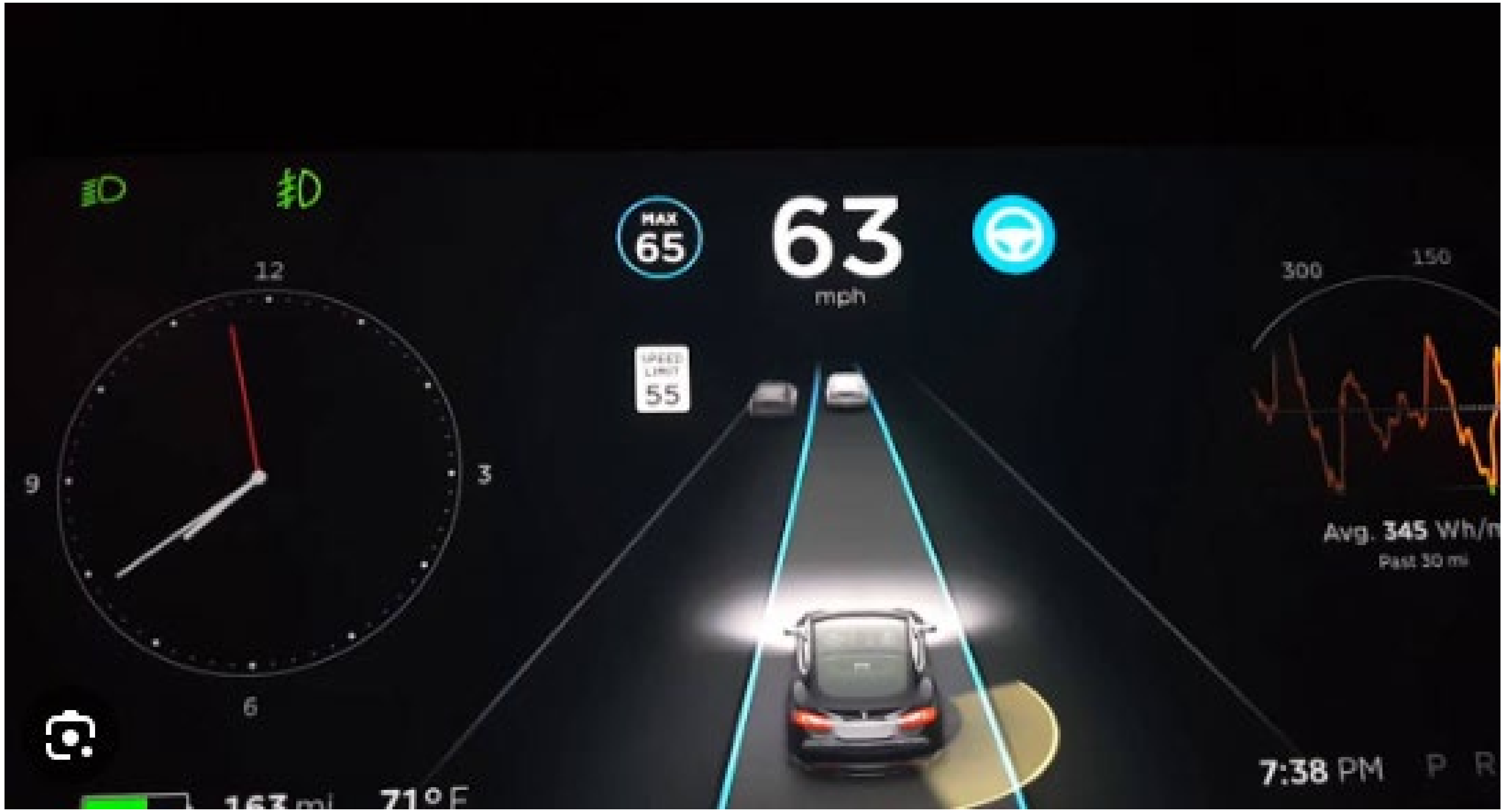
A Tesla driver has guidance control of the vehicle.

Suppose the Tesla driver sees an animal, but does not change lanes to avoid it.

Counterfactually, had she tried to change lanes, the autopilot would have blocked her.

Is the Tesla driver responsible for the animal’s death, even though automation made sure that she couldn’t do otherwise?

MHC predicts yes.



Reverse case: Autopilot has guidance control of the vehicle.

Suppose as a foreseeable consequence of its design, autopilot does not change lanes to avoid an animal.

Suppose that had Autopilot tried to change lanes, the human operator would have blocked it.

Are the designers of Autopilot responsible for the animal's death, even though the human operator made sure that it couldn't do otherwise?

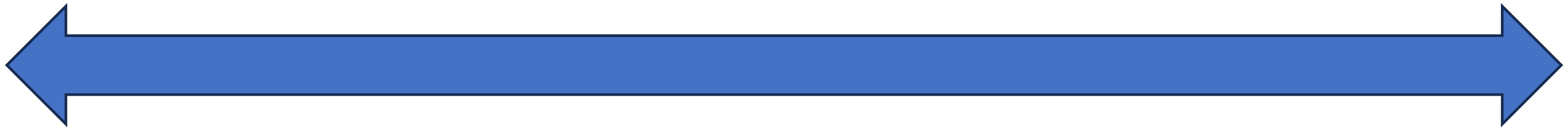
MHC predicts yes.

Background: why people care about control and responsibility

- One of the central concerns about a (semi-)autonomous system interacting with humans is that nobody can be reasonably held responsible for bad outcomes caused by the system. This has been called a ***responsibility gap***.
- In order to be held responsible, an entity must satisfy the following conditions:
  - Foreknowledge ("knew or should have known") → guidance control, tracking (MHC)
  - Control/ causal influence → guidance control, tracing (MHC)
  - Freedom ("could have done otherwise") → regulative control (MHC)

Paradigm 2: Human Centered AI

Levels of automation:
a one-dimensional model



Complete human control

Autonomous mechanical control

Levels of automation

- Suggests a tradeoff between human control and automation.

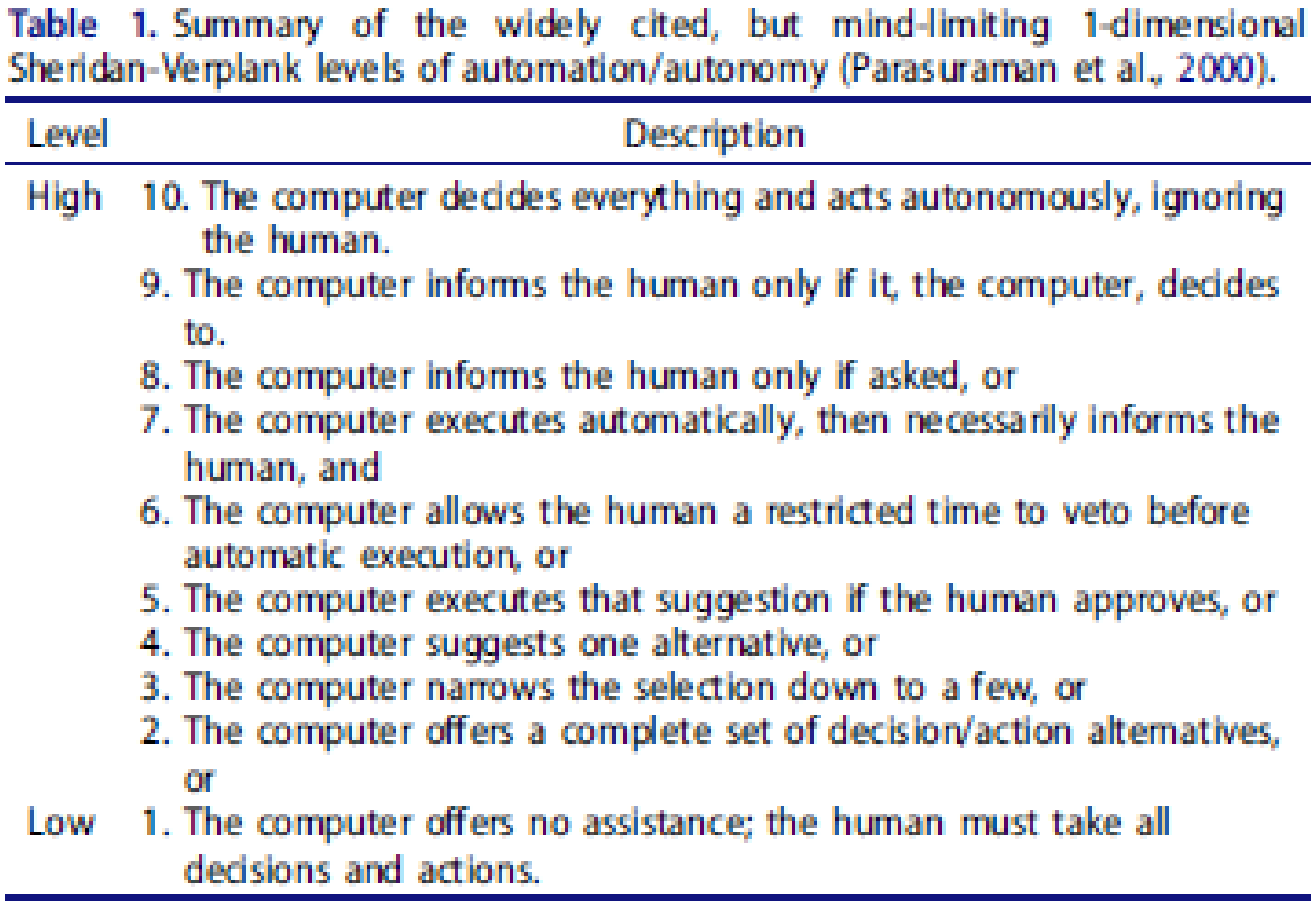
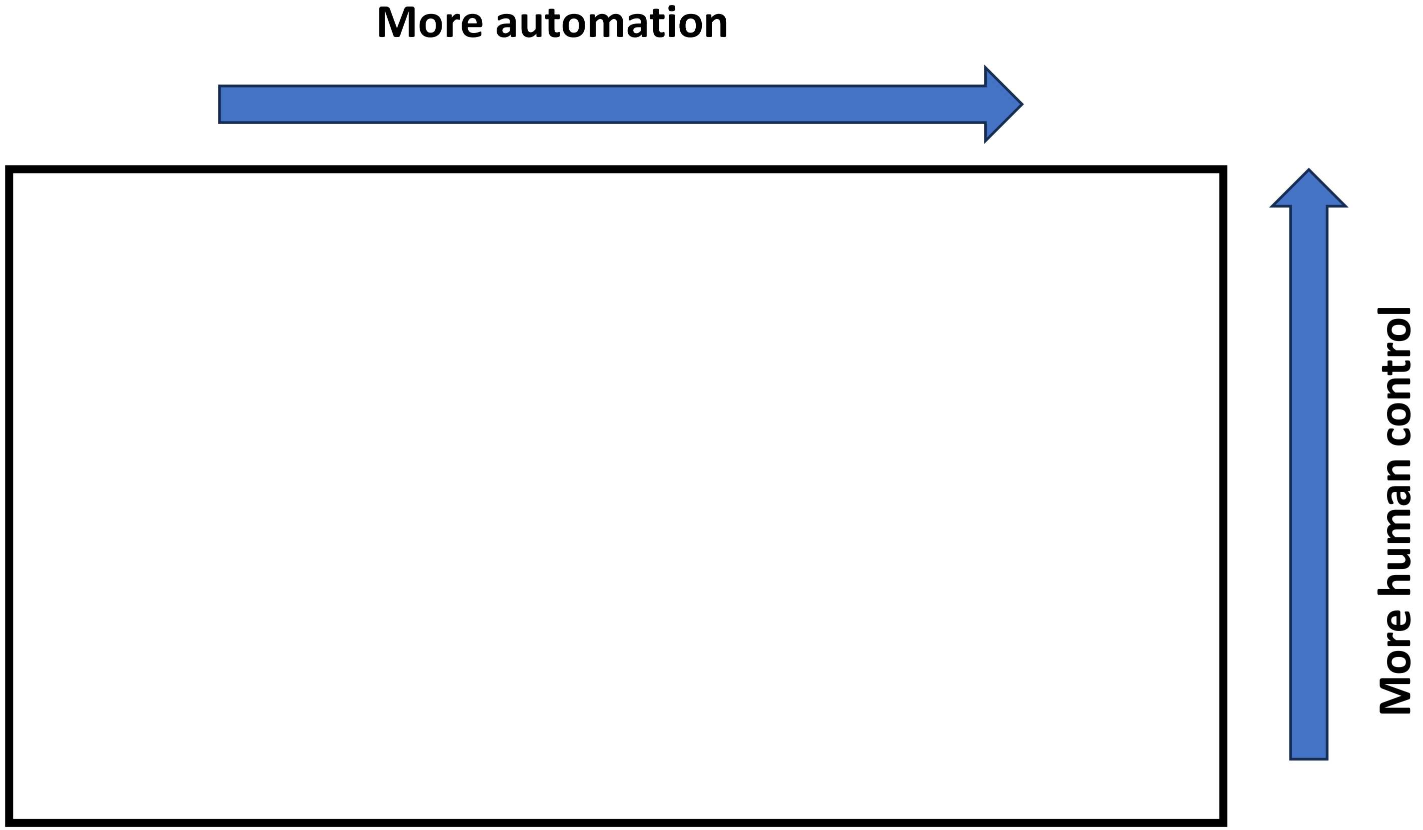


Table 1. Summary of the widely cited, but mind-limiting 1-dimensional Sheridan-Verplank levels of automation/autonomy (Parasuraman et al., 2000).

Level	Description
High	10. The computer decides everything and acts autonomously, ignoring the human.
	9. The computer informs the human only if it, the computer, decides to.
	8. The computer informs the human only if asked, or
	7. The computer executes automatically, then necessarily informs the human, and
	6. The computer allows the human a restricted time to veto before automatic execution, or
	5. The computer executes that suggestion if the human approves, or
	4. The computer suggests one alternative, or
	3. The computer narrows the selection down to a few, or
	2. The computer offers a complete set of decision/action alternatives, or
Low	1. The computer offers no assistance; the human must take all decisions and actions.

A two-dimensional model of automation (Shneiderman 2020)



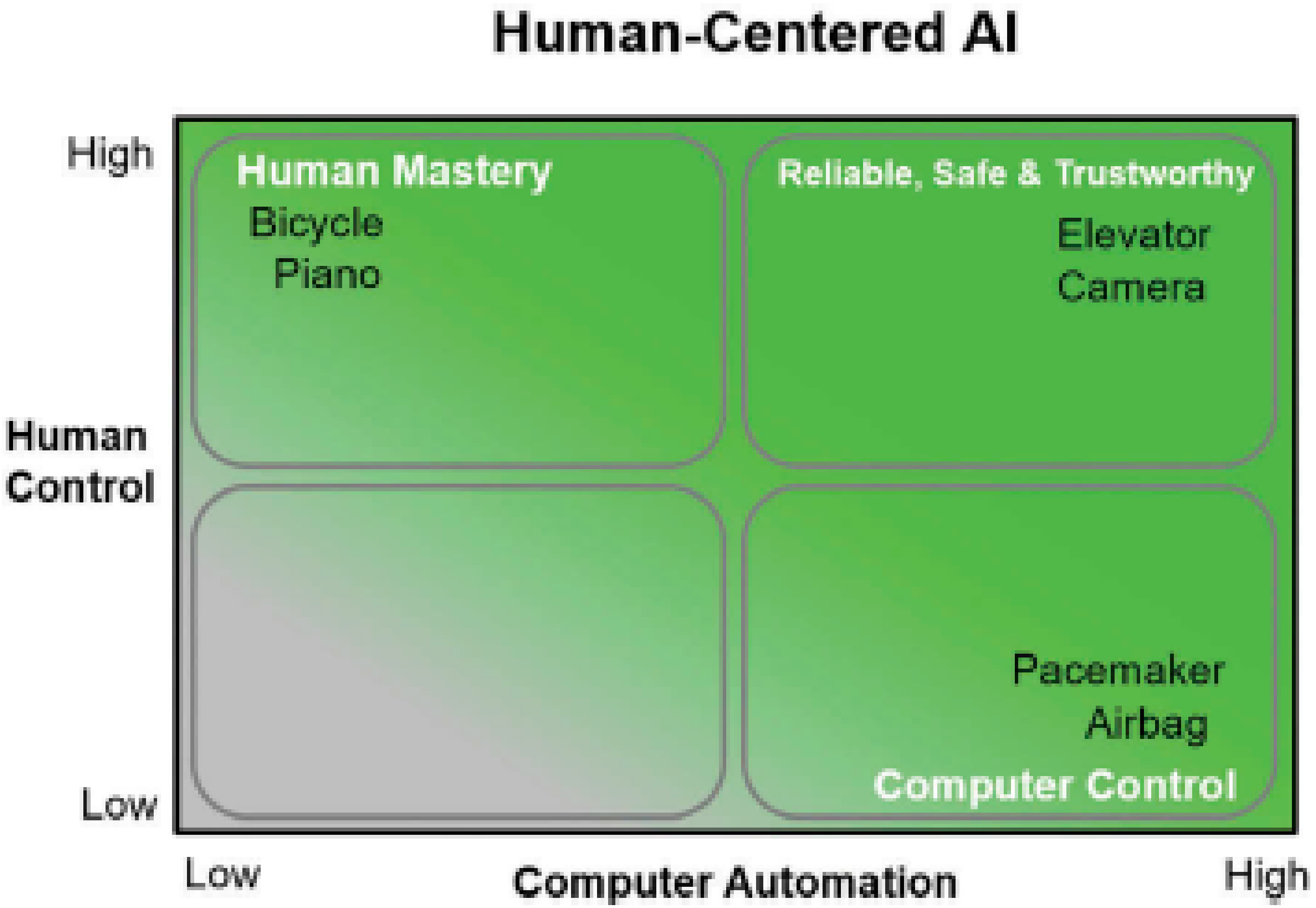


Figure 3. Regions requiring rapid action (high automation, low human control) and human mastery (high human control, low automation).

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.

Shneiderman’s critiques of the TAI

- “Humans have to spend more effort monitoring autonomous computers because they are unsure what it will do, often leading to inferior performance.” (495)
- Deployment of autonomous systems will always fall short of the desired level of autonomy, creating problems.
- “Computers are not teammates, collaborators, or co-active partners… Humans are responsible for actions of the technology assists that they use…” (497)
- Most humans are not interested in high levels of autonomy.
- Avoids programmers’ “algorithmic hubris”

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).

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).

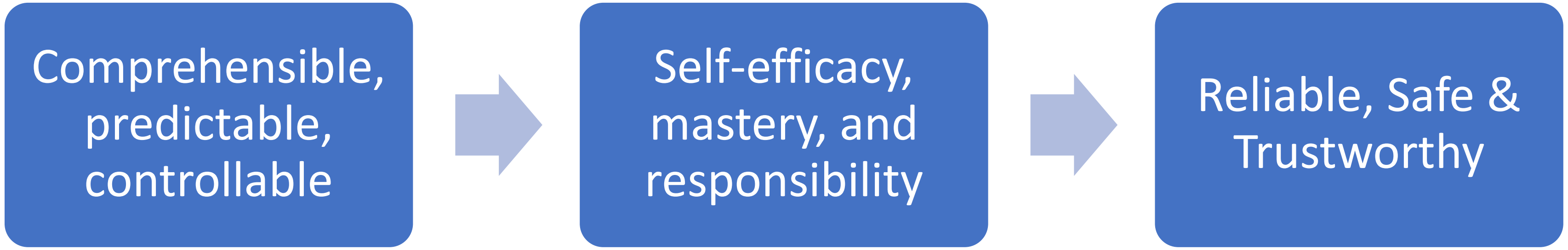
HCAI

- 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

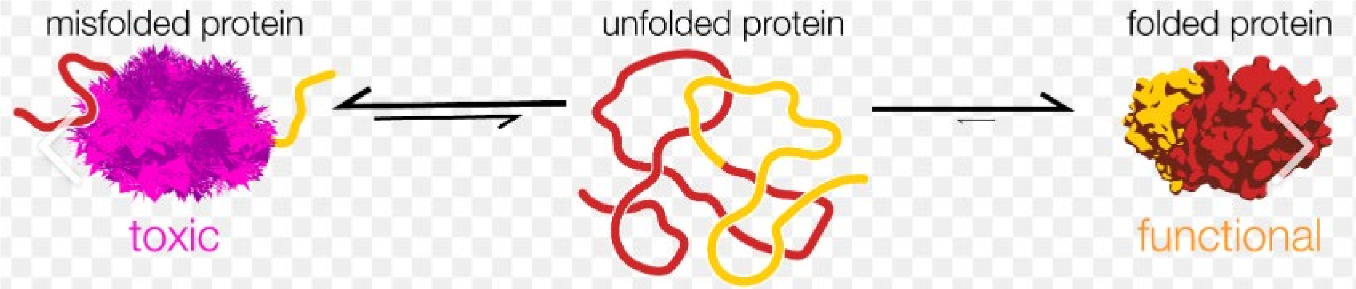
Self-efficacy, mastery, and responsibility

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:

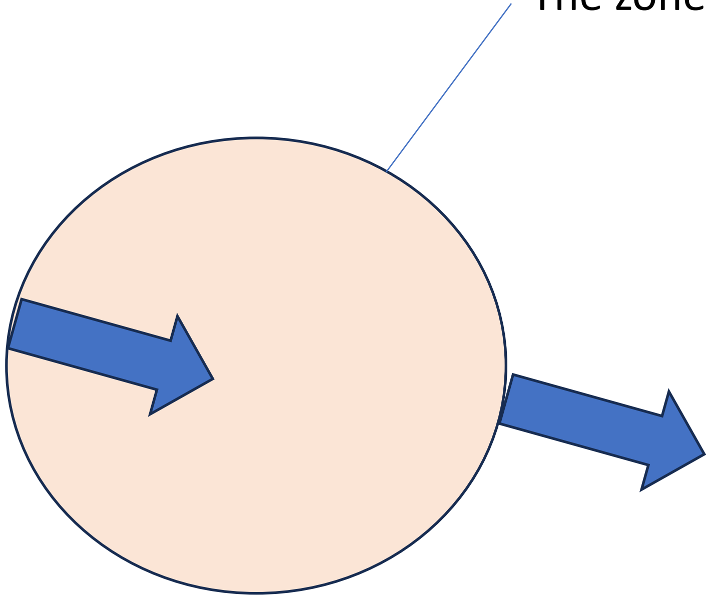
- Responsibility (MHC)
- Trustworthiness (HCAI)
- Amplification of human work (Bankins & Formosa 2023)
- “Hybrid Intelligence”: Synergy of human and machine intelligence (Akata et al. 2020)

Automating *judgment-supporting tasks*

- Suppose a person’s discretion in judgment (e.g., professional, aesthetic, scientific, or moral judgment) is important to their authority and responsibility
- Now consider automation of “judgment-supporting” tasks:
  - Deciding on the parameters of judgment (what scale to use)
  - Filtering the evidence and deciding when one has enough evidence
  - Taking a second look/ checking one’s work
  - Noting consistencies or inconsistencies across time and context
  - Providing an account or explanation of the judgment after the fact
- Here the human retains control. But human discretionary authority can be simultaneously reduced (depending how the social situation constitutes it), along with trust and responsibility.

Implications: discretionary authority in AI-based automation is independent of control

AI applications can increase or decrease the discretionary authority of the professional while keeping control fixed. This depends on how the socio-technical context determines permissions and restrictions.



The zone of discretion

AI in context restricts actions/ judgments that were formerly permitted.

AI in context licenses actions/ judgments that were formerly not permitted.

Professional identity, discretion, and trust

- “Professional service entails risk and vulnerability on the part of clients and the public because of the knowledge and power asymmetries typically present in client-professional relationships. Trust is essential for professional practice precisely because it is the attitude in which one is willing to make oneself vulnerable to the discretionary choices of another person” (Kelly 2018, 21).

Trust in professionals transfers to AI or becomes irrelevant

- Possibility (I): Discretionary authority is transferred to the AI system, and the question of trust follows along, now being applied to the AI system (Nickel 2022)
- Possibility (II): Discretionary authority disappears, and there is no question of trust in relation to the AI system

Arguments for (I):

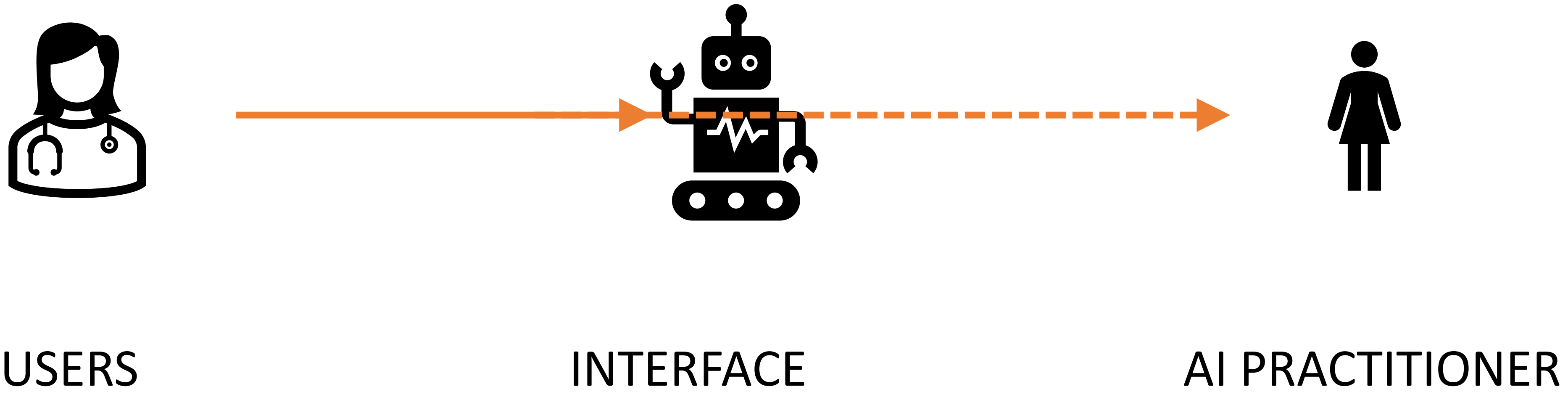
- Attitudinal force of mere reliance is self-directed rather than other-directed
- Explanatory value (style of reliance)
- Linkage arguments: privacy and transparency
- Availability of reductive account for moral implications of trust

Trust in AI?

From Nickel 2022

→ = trusts

⇢ = trusts indirectly



USERS

INTERFACE

AI PRACTITIONER

Transparency doesn’t directly help with trust

- As I understand trust, it … involves *economizing* on monitoring, supervision, and audits, and leaving the trusted to get on with their work with minimal audits and minimal supervision. So increasing these is of course displaying decreasing trust — simply replacing it with audits, supervision, threats, sanctions and coercion. To be worthy of trust, as I understand it, is to be worthy of being left fairly unmonitored and unsupervised, needing to be only minimally checked-up on, given discretionary powers in looking after whatever is entrusted to one. Even if the audits did show good performance, that would be performance when subject to audit, so not yet when really trusted (Baier 2013, 175, italics added).
- Epistemic vs. institutional transparency
- But on the other hand: XAI methods leave plenty of room for vulnerability and discretion

Technical References

Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen, M. Welling. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence, *Computer* 53 (2020) 18–28.

M.R. Endsley, D. Kaber. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics* 42 (1999), 462e492.

B. Shneiderman. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36:6 (2020): 495-504, DOI: 10.1080/10447318.2020.1741118

M. Vagia, A. Aksel, S.A. Transeth, A. Fjerdingen. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics* 53, Part A (2016): 190-202. doi: 10.1016/j.apergo.2015.09.013.

Philosophy References

Akrich, M. 1992. The de-scription of technical objects. In W. Bijker & J. Law (Eds.), *Shaping technology/ building society*. Cambridge, MA: MIT Press.

Baier, A. 2013. What is trust? In Archard, D., Deveaux, M., Manson, N., & Weinstock, D., eds. *Reading Onora O'Neill*. Routledge. Pp. 175–185.

Bankins, S. & Formosa, P. 2023. The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics* 185, 725–740. https://doi.org/10.1007/s10551-023-05339-7

Kelly, T.M. 2018. Professional Ethics: A Trust-Based Approach. Bloomsbury.

Nickel, P.J. 2022. Trust in medical artificial intelligence: a discretionary account. *Ethics and Information Technology* 24. DOI: 10.1007/s10676-022-09630-5Di Nucci, E. 2020*. The Control Paradox: From AI to Populism*. Rowman & Littlefield.

Santoni di Sio, F. & van den Hoven, J. 2018. Meaningful human control over autonomous systems: a philosophical account. *Frontiers in Robotics and AI* 5.