Automation, responsibility and meaningful human control

Giulio Mecacci

Simeon Calvert, Daniël Heikoop, Marjan Hagenzieker,
Filippo Santoni de Sio, Bart van Arem

TU Delft

NWO Netherlands Organisation for Scientific Research

Meaningful Human Control over automated driving systems

Who am I?

TU Delft

Meaningful Human Control
over automated driving systems

Ethics and Philosophy of Technology

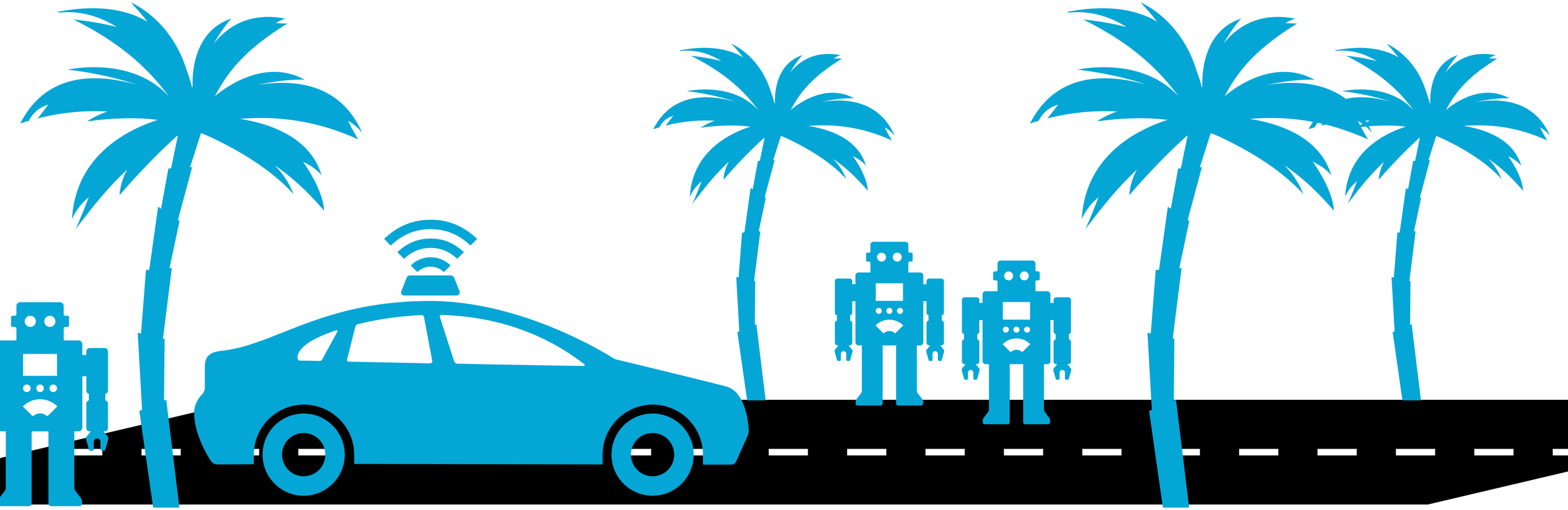
Ethics and philosophy of AI and neurotechnology

DONDERS INSTITUTE

Artificial Intelligence
Radboud University

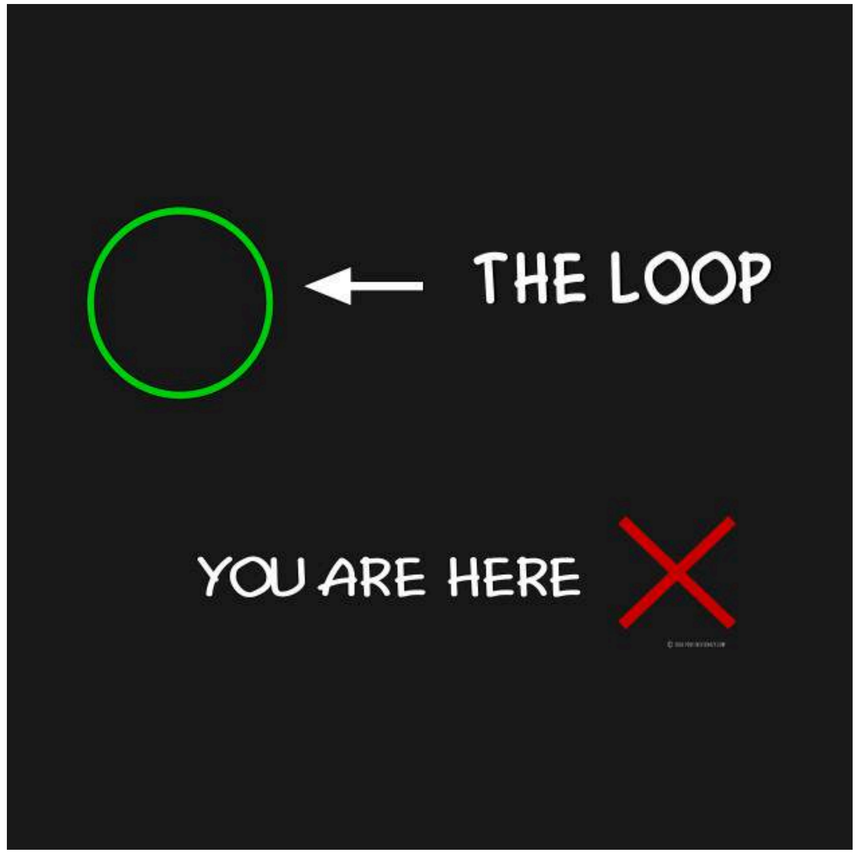
<AI>RADBOUD UNIVERSITY</AI>
Artificial Intelligence Program

The path to enlightenment begins with control and responsibility



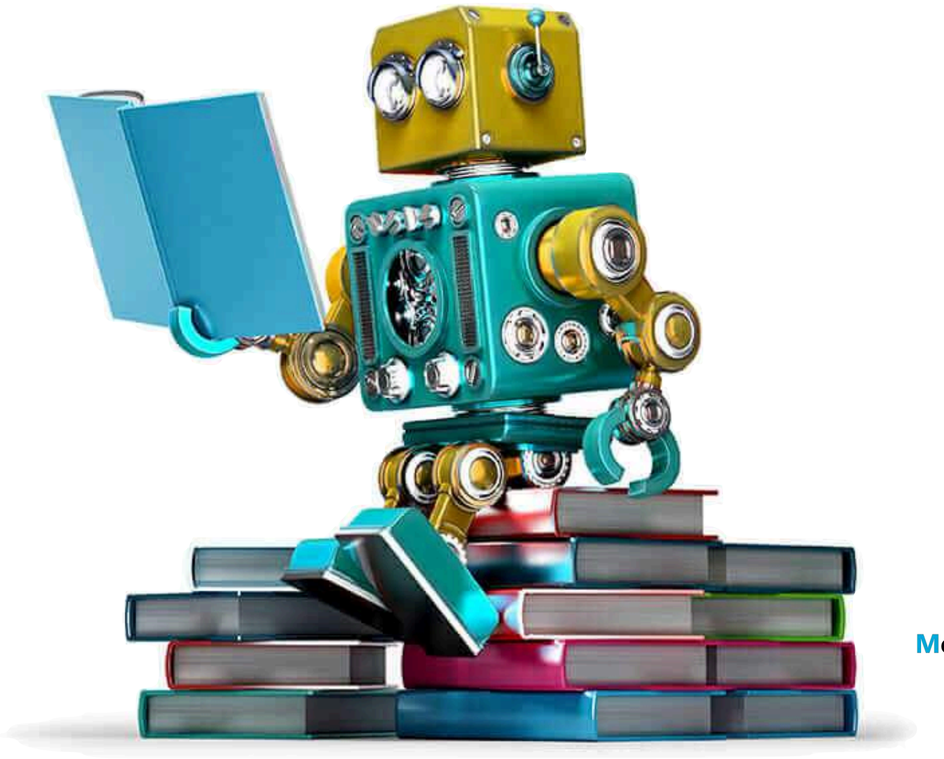
Control is important for responsibility

- Less control can lead to less responsibility, and create the so-called "responsibility gaps"



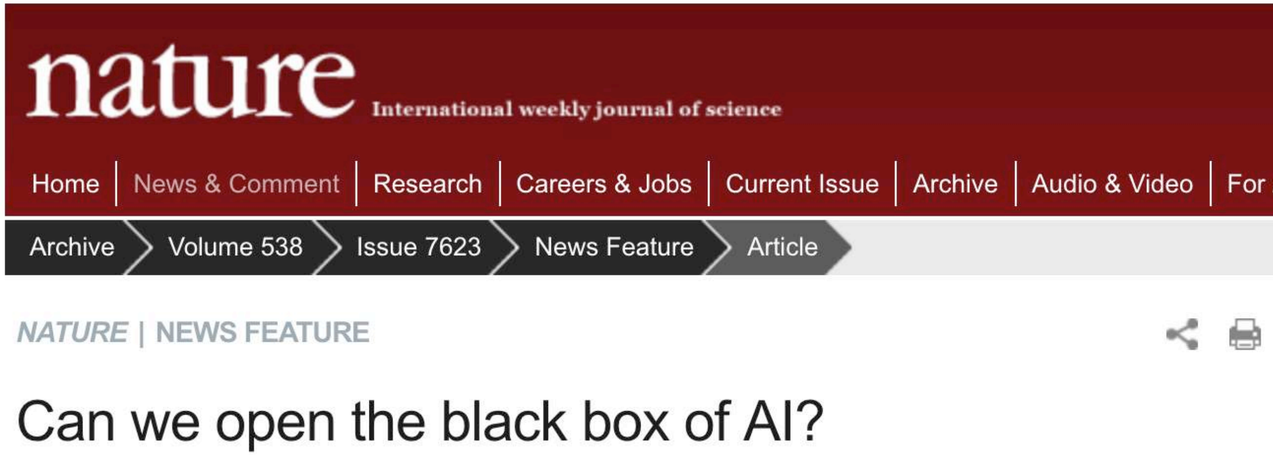
The problems with responsibility: learning automata

- Some machines might be designed to learn new things by themselves, and might become unpredictable
- If nobody can predict what they'll do, who's going to be responsible for their actions?



The problems with responsibility: opacity

- Very competent artificial intelligence means sometimes losing track of how it actually does what it does



TU Delft

Meaningful Human Control
over automated driving systems

The problems with responsibility: us

- Humans suffer from several cognitive limitations in their interaction with intelligent machines

TU Delft

The problems with responsibility: us

- We are naturally lazy, and tend to accept suggestions without debating

TU Delft

The problems with responsibility: us

- We don’t know what’s going on

TU Delft

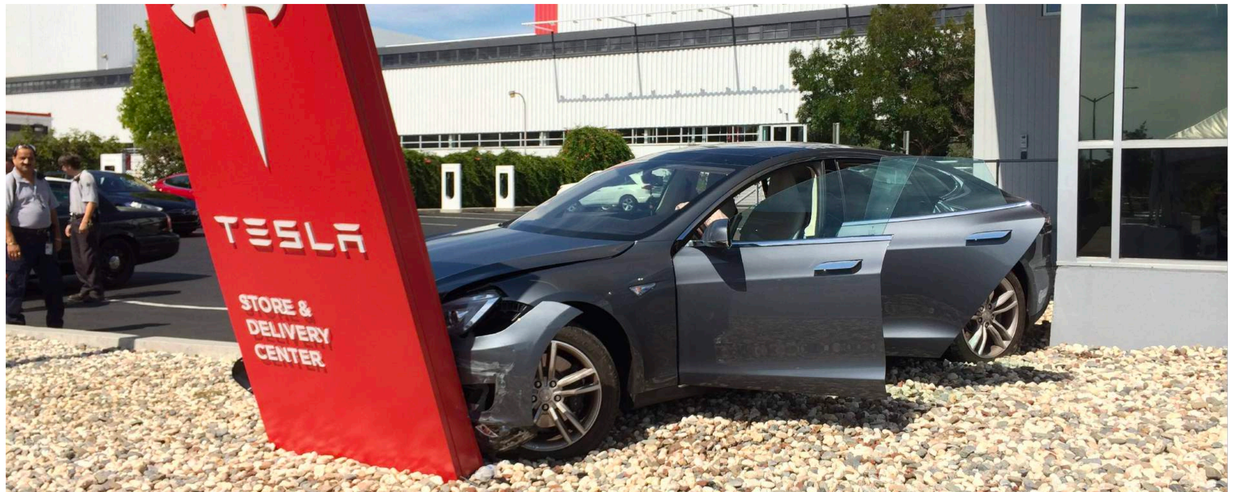
The problems with responsibility: us

- We are slow and easy to distract

The case in vehicle automation

- **Partial automation** might make unclear whether and to what extent users, vehicle manufacturers or even programmers, are involved, and potential **morally responsible,** in case of accidents

- This may lead to "responsibility gaps", and stimulate, as solution, **opportunistic, unfair forms of attribution of responsibility** (e.g. blaming the drivers as they are supposedly **"in control"**)



The road to full automation

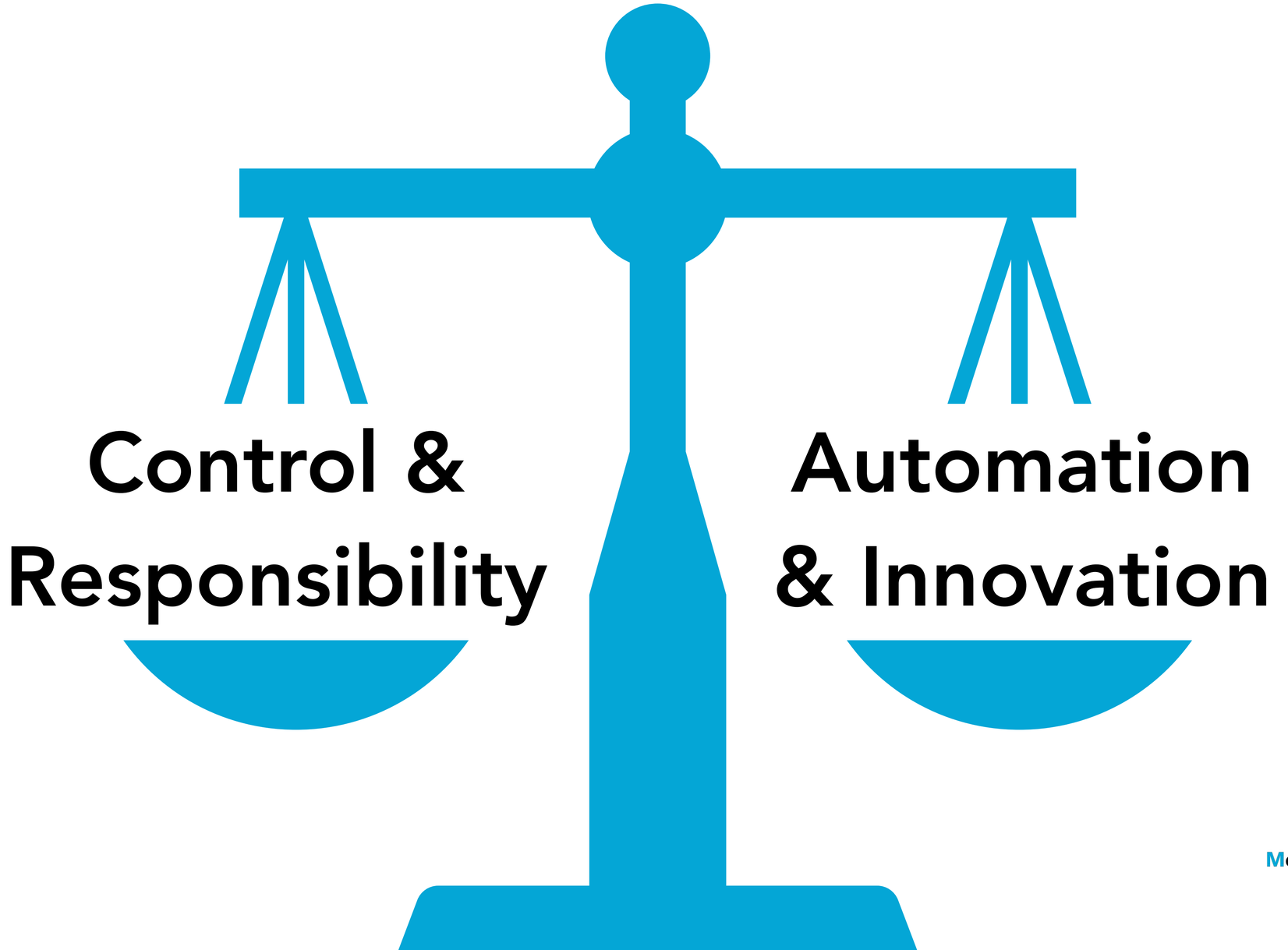
- Full automation seems to **challenge the very possibility of control**, making further difficult to deem somebody responsible. We might be **tempted to resort to legal liability solutions**

The value of moral responsibility

- **Intrinsic** value of moral responsibility: **self-understanding** + **duty to explain** one's behaviour to one another in terms of reasons

- **Instrumental** value of moral responsibility: **promoting safety** via enhancement of **sense of responsibility** and reduction of responsibility shifting

Meaningful Human Control to save both worlds?



TU Delft

Meaningful Human Control
over automated driving systems

The many faces of Meaningful Human Control

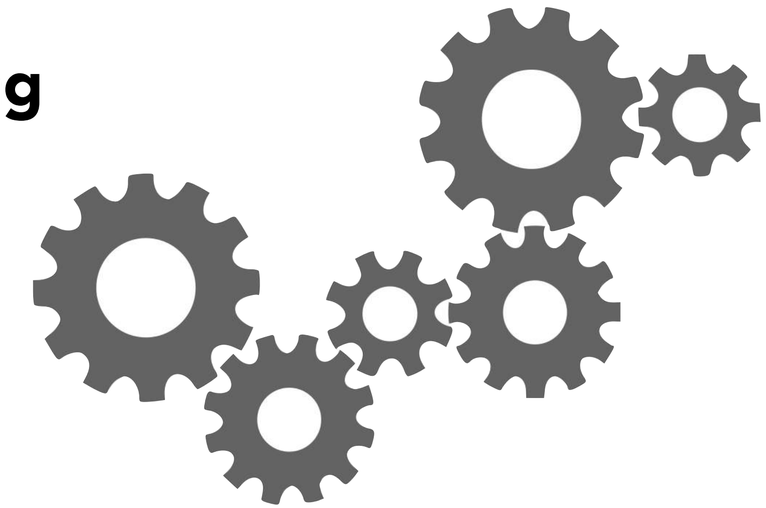
Table 1. Recurring terms, themes, and elements in existing descriptions of human control standards

CNAS	US DoD	Article 36	ICRAC	ICRC
Human operators make **informed, conscious decisions** about the use of force.	The need for operators to make **informed and appropriate decisions** in engaging targets through readily understandable interface.	Reference to timely **human judgment** and action.	There must be active **cognitive participation** in the attack and the ability to **perceive and react** to any change or unanticipated situations.	Reference to **human intervention** in different stages (development, deployment, use).
Human operators have **sufficient information** to ensure the lawfulness of the action they are taking, given **what they know about the target, the weapon, and the context for action.**	Systems will be designed with appropriate human-machine interfaces and controls as well as appropriate safeties, anti-tamper mechanisms and **information assurance**.	**Accurate information** for the user on the **outcome sought, the technology and the context of use**.	Reference to **deliberation on the nature of the target, its significance and likely incidental effects**. Also a reference to the need to have full **contextual and situational awareness of target area**.	**Knowledge and accurate information** about the functioning of the **weapon system and the context** of its intended or expected use.
The weapon is **designed and tested**, and human **operators are properly trained**, to ensure effective control over the use of the weapon.	Need for rigorous **verification and validations, operational testing and evaluation** to ensure the systems function as anticipated.	Reference to need for **predictable, reliable and transparent technology** – that could be linked to **design** features.	Reference to a means for the **rapid suspension or abortion** of the attack-that could be linked to **design** features.	Reference to need for **predictability and reliability** of the weapon - that could be linked to **design** features.
Explicit reference to the need for sufficient information to ensure the **lawfulness** of the action is included in the element's description.	A reference to the need to employ systems **in accordance with the law** is made in the Directive but not as part of the standard itself.	**Accountability** to a certain standard. The requirement to make legal judgments is described in the broader analysis of the concept.	**Necessity and appropriateness** of attack. Meeting the requirements of international law is reflected in broader statement as a driver.	**Accountability** for the functioning of the weapon system following its use. IHL compliance is considered a core driver of the concept.

(Merel Ekelhof, 2018)

Two conditions for a system to be under *meaningful human control*

Tracking

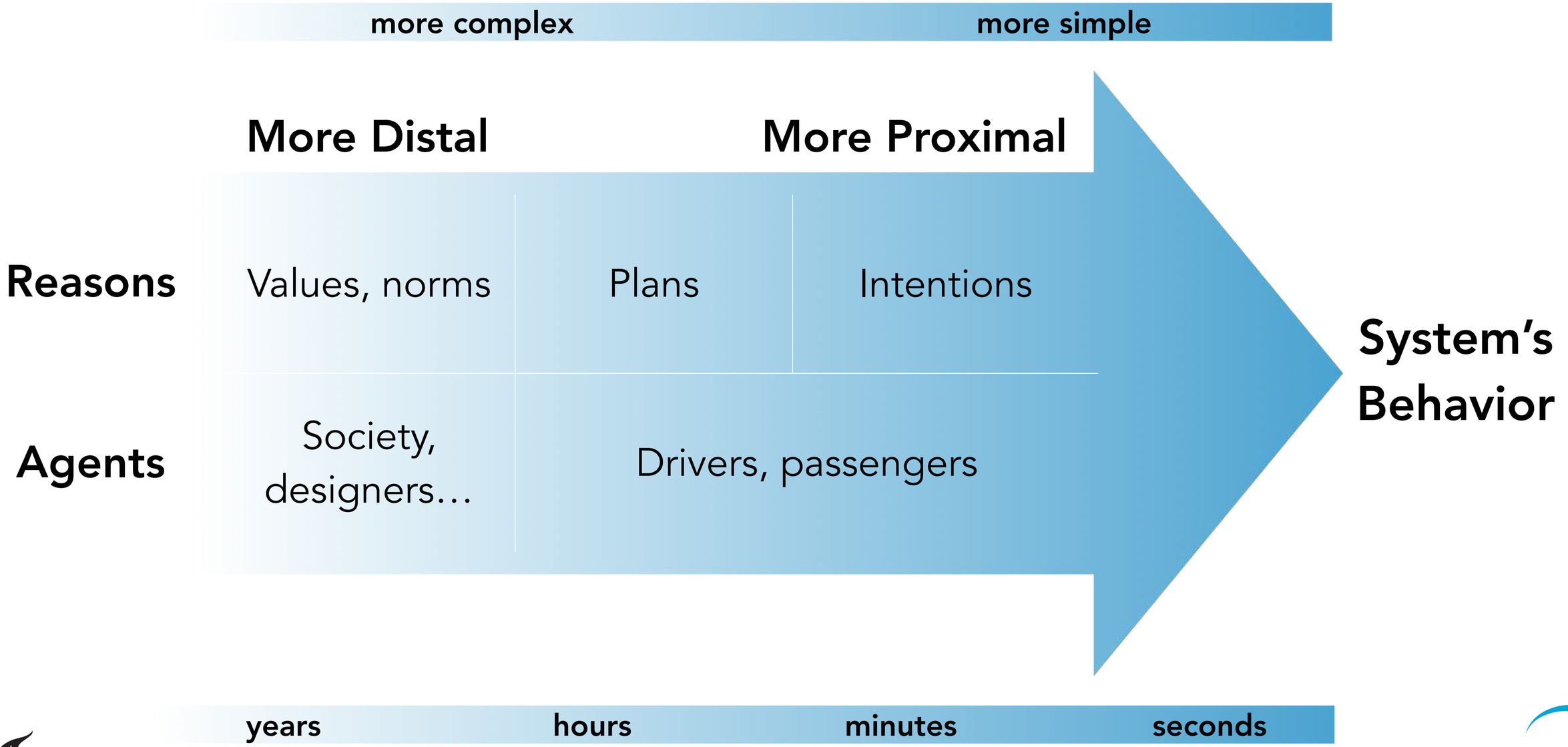


*The system (human operators, operated devices, infrastructures…) should be able to **co-vary** its behavior with the **relevant reasons** of the **relevant human agent(s)** for carrying out X or omitting X*

Tracing

*There is at least **one human agent** in the system design history or use context who can **appreciate the capabilities of the system** and their own role as target of **potential moral consequences** for the system's behaviour*

Distributing control and responsibility to the actors with our scale



Mecacci & Santoni de Sio (2019), *Ethics and Information Technology*

Concluding remarks

- Automated systems are hard to control
- Meaningful human control can provide the kind of control that can help keeping human agents responsible
- It also offer suggestions on how to assess control and how to design systems to maximise it

TU Delft