Trust in Autonomy: Cyber Human-Learning Loops

Asun Lera St.Clair, Senior Principal Scientist

Digital Assurance Programme, Group Technology and Research

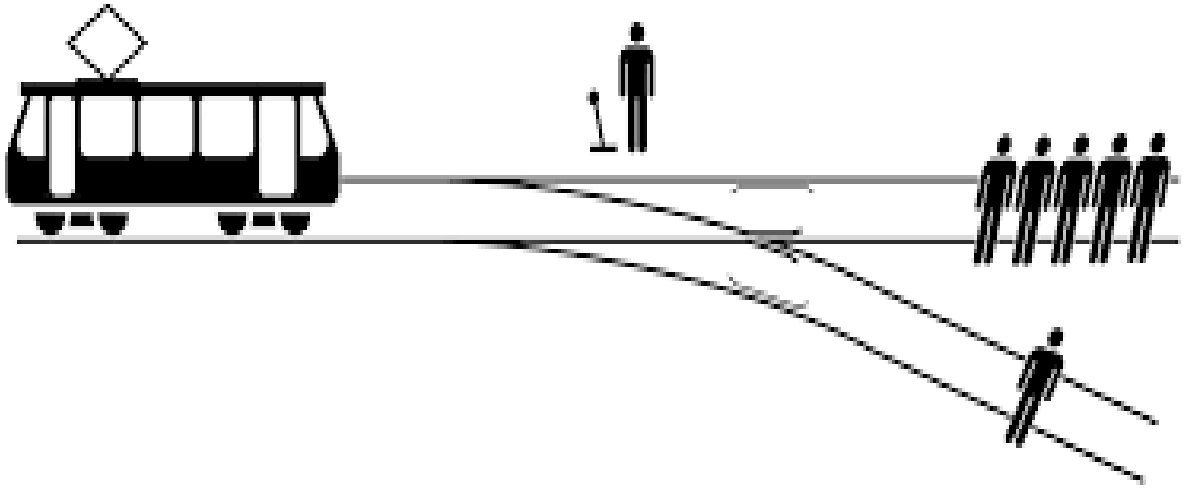
IWASS – Trondheim

11 March 2019

Abstract

- This presentation addresses the ethical and societal implication of autonomous systems. It argues that these are far more complex that the mere aspiration to embed ethical reasoning into algorithms. The presentation argues that morality is a characteristic of human beings and cannot be transported into machines. It is important to distinguish between explainability of autonomous systems versus trustworthiness. Trust is underpinned by shared ethical and societal values, and the conditions for trusting technologies are similar to those of trusting other people or institutions. In the case of autonomy this means both, an assessment of the goals and purpose of the technology as well as assessment of the technical robustness of the system. The core ethical and societal issues associated with autonomous systems emerges from the complex interactions between software, hardware and human beings, alongside the context in which the system operates and the consequences it may have-- directly or indirectly, on people and the environment. Even if autonomous, human beings are part of their design, construction, deployment, operation, maintenance, evaluation and verification of these systems. A potentially normative approach to aim towards is the generation of cyber (physical)-human (social) learning loops, requiring true interdisciplinarity, in particular with the social sciences and the humanities.

The ”Trolley Problem” is insufficient to guide us



- ***Created as a thought exercise in metaethics to illustrate differences in ethical theories***

- ***Framing the ethical and social challenges of autonomy primarily like this is not useful,***

or worse, it blinds us to many key ethical and societal issues needed to have a social lisence to operate

Moral machines?

- Do you know exactly what would be your "moral" choice in case of an accident before you get on your car?

- Would you be a utilitarian or a human rights defender?

Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

Hopeless choices

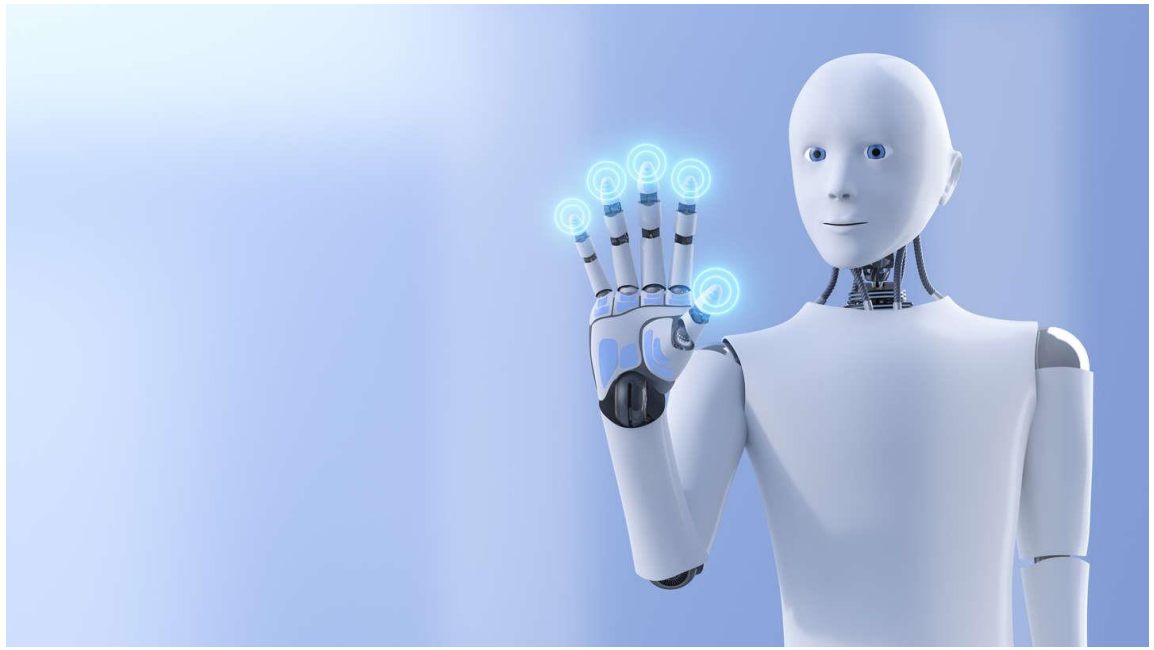
- Is the choice on who to harm in case of unavoidable accidents a necessary question?

- How can a technology that presents us with hopeless choices be a technology that has an ethical purpose?

Awareness of anthropomorphising autonomy

Morality is an attribute of human beings

At the same time, we are asking machines to do things we can't do!



What other questions do we need to ask in order to envision ethical and societal challenges emerging from autonomous systems?

Several options…..

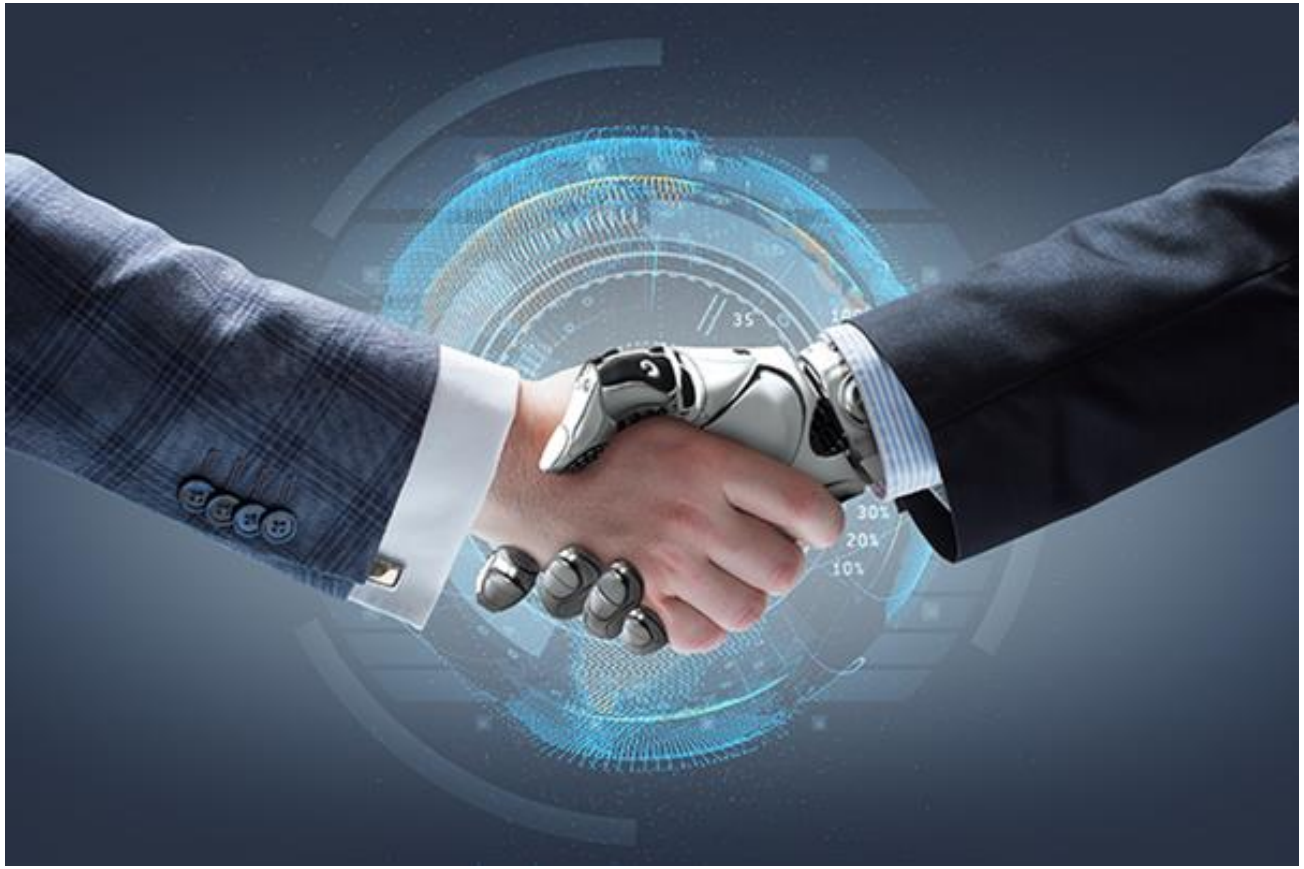
Uber won't be charged with fatal self-driving crash, says prosecutor

But the backup driver might still be held responsible

- Was there a technical problem?
  - What type of problem?
- What was the right thing to do? (by whom)
- What type of legal conventions and regulations do we have to handle these situations?
- ***How can we trust that those with the ability to prevent or avoid the accident did their best?***

Ethical and societal issues are a component of trust in autonomy

- Trust is underpinned by **shared ethical and societal values**
- The conditions for trusting technologies are **similar to** those of trusting other people or institutions
- **How can we trust** autonomous systems?
  - The answer is an **integration** of technical and non technical issues in a way that creates **ongoing learning**



We delegate responsibility to those people, institutions or technologies that we trust

Different levels of autonomy / different roles of the autonomous system (decision aid, co-worker, manager = different forms of delegation

Explainable versus trustworthiness

- **World Economic Forum:**
  - Bias
  - Transparency
  - Accountability
  - Privacy
- **Microsoft:**
  - Fairness
  - Reliability and safety
  - Privacy and security
  - Inclusiveness
- **IBM:**
  - Fairness
  - Robustness
  - Explainability
  - Lineage

Trustworthy AI has **two components:** (1) it should respect fundamental rights, applicable regulation and core principles and values, ensuring an "**ethical purpose**" and (2) it should be **technically robust** and reliable since, even with good intentions, a lack of technological mastery can cause unintentional harm.

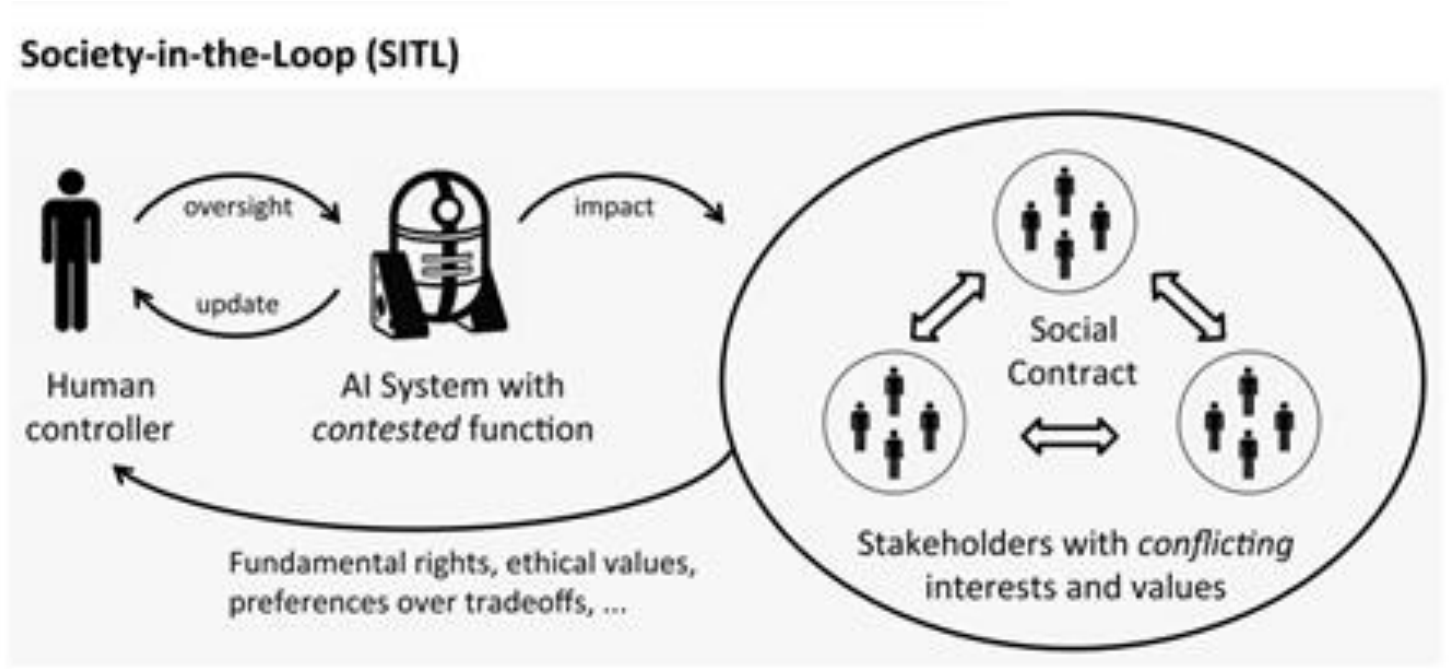
The overarching principle of human-centred technology

The goal of a specific technology needs to be an ethical purpose!

Revert to humans (beyond the “human operator”)

- There are always humans around any autonomous system, even though responsibilities are distributed among elements of the complex system.

From human to society in the loop



Rahwan, I. (2017). Society-in-the-Loop: Programming the algorithmic social contract, Ethics of Information Technology.

Cyber-physical-human systems

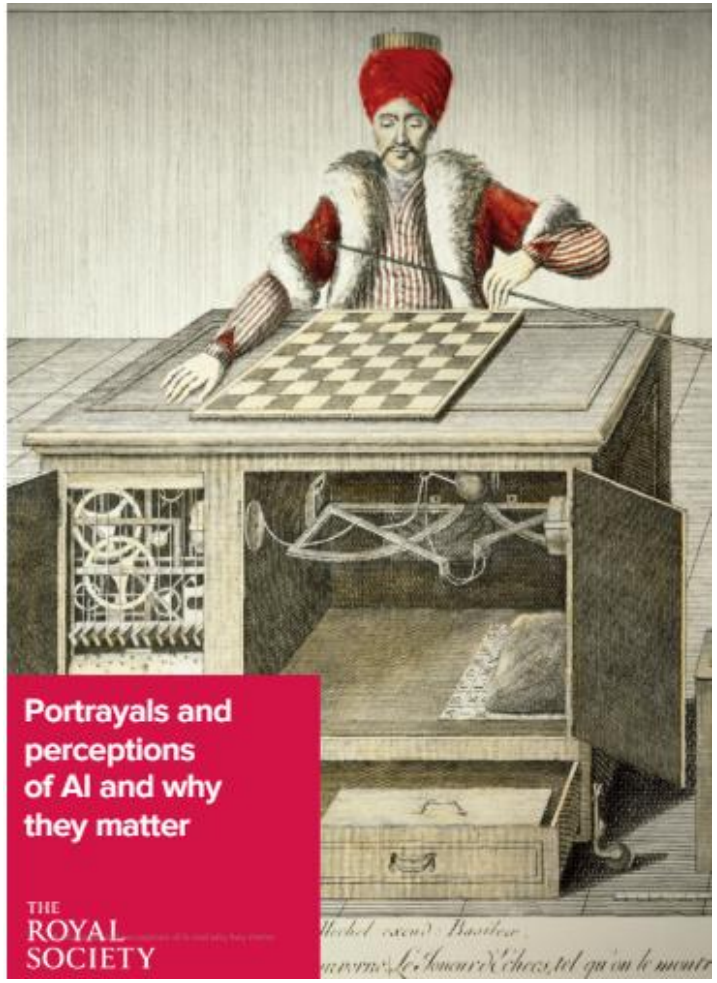
- The ethical and social agenda overlaps with the challenges presented in the White Paper :

- "The challenges regarding SRS lie in identifying failures that may arise from the complex interactions of software, hardware, and human operators, as well as from the propagation of those throughout the system's components and subsystems (IWASS White paper p. 9).

Understanding that values are everywhere

- In the provenance of data,
- In the assumptions we make when designing a model,
- In the weighing of options we make in constructing algorithms

We need a better understanding of public perception



Cave, S., and K.Dihal (2019), Hopes and fears for intelligent machines in fiction and reality, Nature

Wider context of the technology

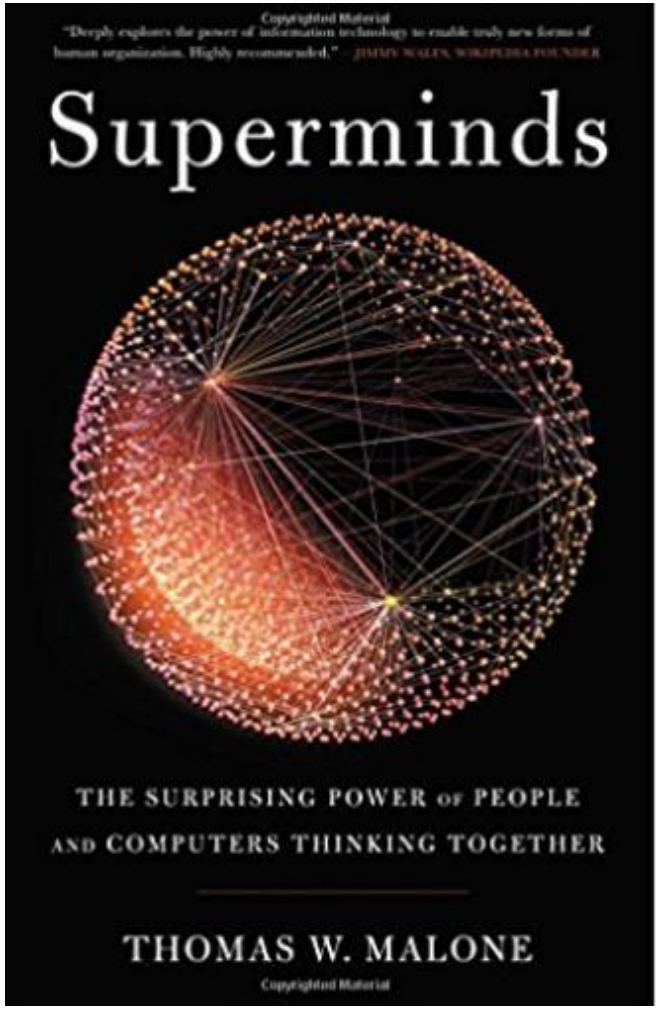
- Job losses for some
- New skills required (not in university study plans)
- Unequal consequences geographically
- Networks of agents and operators
- Changes in transportation systems (moving more toward aviation traffic control systems)
- Need to collaborate with regulators
- Understanding stakeholders needs and expectations

Trust in complex and intelligent systems is not binary (yes or no)

Need for testing, learning, testing, learning

Learning loops across all the technical and non technical elements, and the life cycle of the autonomous system

Cyber (physical) human (social) learning loops



- A Learning Loop is a way to ensure the work we do now informs what we do next. It provides a high-level perspective on how the elements of a complex systems can be broken down into a gradual process of iterative cycles.

- An ethical approach to autonomy can be to aim for cyber (physical) human (social) learning loops.

In conclusion

- Autonomous systems are complex and intelligent systems with humans in the loop & embedded in society.
- Human beings are part of their design, construction, deployment, operation, maintenance, evaluation and verification.
- An ethical and societal analysis of autonomous systems needs to be an integral part of the challenges that arise from the complex interactions of software, hardware, and human operators…
- But should also inquire how human-centric the technology is, and its organisational & societal contexts.
- A potentially normative approach to aim towards is the generation of cyber (physical)-human (social) learning loops, requiring true interdisciplinarity!

Thank you!

Asuncion Lera St.Clair

Asun.lera.st.clair@dnvgl.com

@asunstclair

+ 47 452 619 02

www.dnvgl.com

SAFER, SMARTER, GREENER