Bringing ideas from cybernetics to current challenges in AI

Marc STEEN[a], Jurriaan VAN DIGGELEN[a], Tjerk TIMAN[a]
[a] *TNO, The Netherlands Organisation for applied scientific research*

Abstract. We used two ideas from cybernetics—the *sociotechnical system* and the *feedback loop*—to explore ways to deal with some current challenges in the design and application of AI systems, e.g., in Human-Machine Teaming and Algorithmic Decision Making. As thought experiments, we envisioned two 'cybernetic' systems. Our goal is to explore the usefulness of ideas from cybernetics to the field of AI.

Keywords. Cybernetics; Feedback; Sociotechnical system; Human-Machine Teaming; Algorithmic Decision Making.

1. Introduction

Suppose your task is to design a mechanism to read the signals from a CD-disk: the 'zeros and ones'. You have a laser beam and a sensor that need to move *very precisely* to a specific track in order to read the 'zeros and ones'. Do you make a stiff, metal arm? Or do you make the arm of cheap, flexible plastic? Engineers chose to make a flexible, plastic mechanism, *with a feedback loop*. The sensor reads the signals and, *as an integral part of reading*, generates error-signals, which are used to steer the arm in real-time; see Figure 1. *Flexible and precise*. More effective than a metal arm without feedback loop.

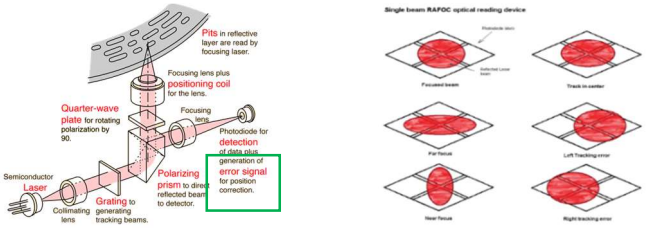

Figure 1. The sensor reads the signal (left) and also produces error signals (right), which are used to steer the laser (left). Images: http://hyperphysics.phy-astr.gsu.edu/hbase/Audio/cdplay.html (left) and https://en.wikipedia.org/wiki/CD_player (right).

Now, let us move to the domain of Artificial Intelligence (AI). There are many images, both in public and in academic discourse, in which AI systems are presented as 'white, shiny, humanoid robots' or as 'stand-alone, blue, floating brains'; see Figure 2.

We believe that these images are misguided. Moreover, they do not help us to address current key challenges in the design and application of AI systems; challenges related to, e.g., control, fairness, and transparency, or questions like: How can AI systems support and enhance human capabilities, rather than replace them or corrode their

capabilities? Such challenges and questions are at play in Human-Machine Teaming (HMT), Algorithmic Decision Making (ADM), and Meaningful Human Control (MHC) [1, 2]. We feel that shiny robots and floating brains are not particularly helpful.

Figure 2. Common imagery of AI: as a 'white, shiny, humanoid robot' (left) or as a 'stand-alone, blue, floating brain' (right). Images: https://robots.ieee.org/robots/charli/ (left) and https://vcatechnology.com/resources/ai-definition-and-app/ (right).

Instead, we propose to use ideas from the field of cybernetics in the 1960s and 1970s, notably the ideas of the *feedback loop* and the *sociotechnical system*, and to apply these ideas to current challenges. A *sociotechnical system* expresses an understanding of people interacting with the world through a complex web of relationships and interactions, involving both social and technical components [3]. We speculate that these concepts can help to create more realistic images in AI projects, in the minds of AI developers—and that these can help to create AI systems that are more aligned with needs and values in society, and more realistic and effective.

2. Ideas from cybernetics

The term 'AI' was coined at the *Dartmouth workshop*, in 1956. The *Macy Conferences* on cybernetics, however, had started already one decade earlier. Ten of these conferences happened between 1946 and 1953; they were famous for their interdisciplinary approach. For about a decade, the two fields (Cybernetics and AI) coexisted, but by the mid-1960s, proponents of symbolic AI became more successful in gaining research funding. Consequently, cybernetics lost traction. 'This effectively liquidated the subfields of self-organizing systems, neural networks and adaptive machines, evolutionary programming, biological computation, and bionics for several decades' [4: p. 89].

In cybernetics, people and their goals are understood as embedded in their environment and surrounded by machines; the combination of people, environment and machines is understood as a *complex, adaptive system*. The various components are connected by inputs and outputs, and, crucially, through *feedback loops*, which enable the system to have a stable course or status, independent of changing circumstances (*cybernetics* refers to steering a ship, on a stable course, through winds and waves).

These two key ideas can be illustrated with a (part of a) drawing by 'cybernetician' Stafford Beer, from his *Brain of the Firm* (1972, 1981) [5]; see Figure 3: a complex, adaptive system with diverse components, connected by arrows and feedback loops.

The sociotechnical system and the feedback loop have been recurring themes, from Norbert Wiener's *Cybernetics* [6: e.g., pp. 96-97], to, e.g., Cathy O'Neil, who argued, that, without feedback loop, 'a statistical engine can continue spinning out faulty and damaging analysis while never learning from its mistakes' [7: p. 7].

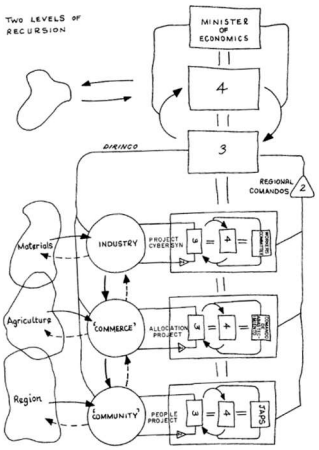

Figure 3. Drawing of a system with diverse components, connected by arrows and feedback loops, by Stafford Beer. Image: *Brain of the Firm*, p. 325.

We speculate that the ideas of *sociotechnical system* and *feedback loop* are under-utilized in the design and application of AI systems, and propose to explore how these might be used. Of course, there is the proverbial 'red button', to halt an AI system. Or, more subtly, operators can use their discretion and correct the system's output. Moreover, some systems use feedback to 'learn', e.g., *supervised learning* or in *reinforcement learning*.

Our proposal is to use these ideas in how we envision and organize *sociotechnical systems*, with AI, and *feedback* mechanisms; not in a narrow sense (e.g., 'red button', correct output, 'learn'), but with feedback mechanisms that involve larger processes or the organization that deploys the AI system. This would enable the people involved in design and application to promote values like fairness and transparency, and better anticipate and deal with undesirable consequences [8, 9]. We can illustrate our proposal with three images; see Figure 4. They portray, in a caricatural manner, the following:

- A stand-alone, AI system, where the user hopes that 'it will work'; like the metal arm in the CD-player; like a white, shiny robot' or a 'blue floating brain' (left);
- A sociotechnical system in which people and AI components collaborate like *team members*; below, we will give an example of HMT (middle);
- People use AI like tools, to enhance their capabilities, like a musical instrument or a craft tool; below, we will give an example of ADM (right).

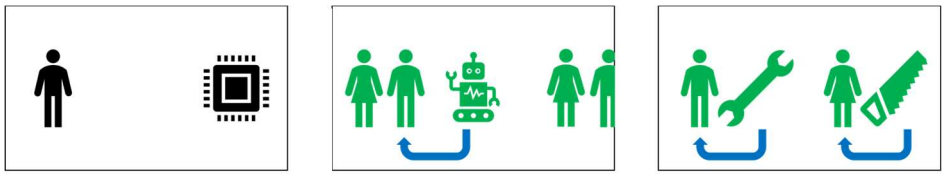

Figure 4. Compare and contrast: a conventional stand-alone AI system (left); people and machines in a sociotechnical system (green), with *AI as a team member* and feedback (blue) (middle); and *AI as a tool* in a sociotechnical system (green) and feedback (blue) that supports and enhances human capabilities (right).

In the next two sections, we envision two *cybernetic* AI systems. Both aim to *support* and *enhance* human capabilities [10]; rather than aim to replace or corrode these. They function as *thought experiments*: *What if we start with professionals' practices and capabilities, and envision a larger sociotechnical system, with feedback mechanisms, around them, with AI components that support them?*

- The first system is concerned with HMT in a military context; there, the AI components functions as *team members*, possibly similar to how we have interacted with animals, e.g., dogs, horses or pigs [11], but also very differently because people and animals share biology—and machines do *not*;
- The second system is concerned with ADM in a police context; there, the AI component functions *as a tool*, similar to how people use a musical instrument or a craft tool; with these, people can develop their craft [12], extend human capabilities [13], and cultivate technomoral virtues [14].

3. Human-Machine Teaming (HMT) in the military

Let us imagine two soldiers with five robots, e.g., the familiar dog-shaped ones, in a reconnaissance task. These robots function as *team members*, with capabilities different from people: on the one hand, less skilful; on the other hand, able to operate in dangerous environments. They could function like pigs that help to find truffles; see Figure 5.

In current set-ups, the soldiers are able to provide feedback on two basic levels: 1) by controlling the robots, e.g., giving new coordinates; or 2) by asking others, at a distance, to investigate the robots' camera feeds and to provide advice.

With robots *as team members*, in a *cybernetic system*, we would have additional options and levels of feedback: 3) the soldiers could extend the HMT to include people higher-up in the chain of command; these people can take more consequential decisions, and be accountable for these; and 4) they could share feedback with other and future operations. This could involve 'double loop learning', e.g., sharing feedback data across multiple operations, or missions, even.

The soldiers and 'their' robots can, over time, engage in 'co-learning' [15] and mutually adapt their behaviours. It is an open question whether these patterns can be shared across teams (and what would then happen).

Figure 5. We can envision HMT involving soldiers and robots in a reconnaissance task, similar to HMT involving people and pigs in the task of finding truffles. Image: https://en.wikipedia.org/wiki/Truffle_hog

4. Algorithmic Decision Making (ADM) by the police

Let us imagine a police officer (or other public servant) with the task to detect criminal (or other) activities, working with a system that puts orange flags for the names of people who are associated with higher risks for such activities. With such a system, an operator can provide feedback on two basic levels: 1) to correct the output, based on their experience and discretion; or 2) to conduct some investigation to verify or falsify the output, and make corrections accordingly. Crucially, such systems are often designed in a central government body, and deployed by some decentralized government body. They typically function in one-way; feedback from the periphery does not reach the centre.

With an AI component as tool, in a *cybernetic system*, we would have additional options and levels of feedback: 3) other professionals, from other organizations, in the periphery, could conduct (additional) investigations and produce (additional) findings, e.g., about true positive or false positive, and bring this feedback to the centre, where it can be used to improve the tool; and 4) the larger organization could enable citizens (whom currently are typically viewed as 'data subjects') to ask questions, to contest the risk assessment and the decision, and to require redress. Their responses are currently

viewed as complaints. Instead, it would be possible to find *patterns* in citizens' responses, which can be used to improve the system (feedback that is otherwise not used).

This tool would 'grow' over time, during use, so that it becomes a personal tool, for one specific operator, based on how they approach the task and uses the AI. Again, it is an open question whether this tool can be shared with other operators.

Figure 6. We can envision a ADM to find criminal behaviour, similar to using a head-mounted torch to find specific objects in the dark. Image: https://www.digitalcameraworld.com/buying-guides/best-head-torch

5. Discussion

What is different for cybernetic systems, compared to conventional AI systems, and what is potentially of added value, is that a cybernetic system is understood as a *sociotechnical* system, an assemblage of people and machines in an environment, connected through all sorts of inputs, outputs and *feedback loops*, so that 'it works'; it is 'performative' [16: pp. 17-27]. This is different from a conventional AI system, through which we attempt to collect knowledge, which is then processed and presented to us ('epistemological'). It may be interesting to further explore this difference. Possibly, the 'performative' nature of cybernetic systems makes them **less transparent**, a 'black box' [16: p. 27].

Another interesting topics is co-learning. In both cases (HMT and ADM), we touched on the issue of mutual adaptations, of **co-learning**, over time—including open questions about the (im)possibilities of sharing learning across teams, operations or missions, or professionals. It would be interesting to further explore this topic [15, 17]. What would need to be in place, e.g., to enable teams to spend days or months getting used to working (or socializing) with their robots in a team, or to enable professionals to spend days or months to learn to use their tools skilfully (or artfully)?

We also noted a recurring pattern: sociotechnical systems can grow rather complex. This raises questions about **which scale or scope** could be appropriate, e.g., for analysing problems or developing solutions. Imagine an AI component that works nicely on the microlevel. How could we anticipate undesirable effects on, e.g., the level of society? Twitter may work fine for me. But on the level of society, it fuels polarization. Or how could we take into account larger societal injustices when we design one AI component. This question is at play, e.g., when judges use algorithms that 'predict' recidivism [18]; a 'fair' (micro) algorithm is used in a larger system that can be unfair. Is it then 'fair'?

Finally, there are questions that follow from Weizenbaum's [19] proposal to distinguish between two different activities or verbs: *deciding*, making a decision, what computers can do through calculation; and *choosing*, making a choice, what people can do through judgement. His proposal is to choose **wisely**: what we can (not) delegate to computers (calculation), and what we must leave to people (judgement). In line with virtue ethics, we can speculate that people can cultivate virtues like justice, courage, self-

control, by using by using AI components wisely. This could be relevant for *frontline workers, e.g.,* police officers, nurses, teachers, who use AI components in their work [20].

6. Conclusion

We used two ideas from cybernetics—the *sociotechnical system* and the *feedback loop*—to explore ways to deal with some current challenges in the design and application of AI systems. We envisioned two cybernetic' systems, one for Human-Machine Teaming and one for Algorithmic Decision Making. These thought experiments left us with a series of questions and suggestions for further research.

References

1. Siebert, L.C., et al., *Meaningful human control over AI systems: beyond talking the talk.* arXiv preprint, 2021.
2. van der Waa, J., et al., *Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations.* Frontiers in Robotics and AI, 2021. **8**.
3. Latour, B., *Reassembling the Social: An Introduction to Actor-Network-Theory*. 2005, Oxford, UK: Oxford University Press.
4. Cariani, P., *On the importance of being emergent.* Constructivist Foundations, 2010. **5**(2): p. 86-91.
5. Beer, S., *Brain of the firm (2nd ed.).* 1981, Chisester: John Wiley.
6. Wiener, N., *Cybernetics: Or control and communication in the animal and the machine*. 1948/1961, Cambridge, MA: MIT Press.
7. O'Neil, C., *Weapons of Math Destruction*. 2016, London: Penguin.
8. Hayes, P., I. van de Poel, and M. Steen, *Algorithms and values in justice and security.* AI & SOCIETY, 2020. **35**: p. 533-555.
9. Steen, M., T. Timan, and I. van de Poel, *Responsible innovation, anticipation and responsiveness: case studies of algorithms in decision support in justice and security, and an exploration of potential, unintended, undesirable, higher-order effects.* AI and Ethics, 2021. **1**(4): p. 501-515.
10. High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*. 2019, Brussels: European Commission.
11. Darling, K., *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. 2021: Henry Holt and Company.
12. Sennett, R., *The craftsman*. 2008, London: Penguin Books.
13. Robeyns, I., *Wellbeing, Freedom and Social Justice The Capability Approach Re-Examined*. 2017, Cambridge, UK: Open Book Publishers.
14. Vallor, S., *Technology and the virtues: A philosophical guide to a future worth wanting*. 2016, New York, NY: Oxford University Press.
15. Schoonderwoerd, T.A.J., et al., *Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task.* International Journal of Human-Computer Studies, 2022. **164**: p. 102831.
16. Pickering, B., *The cybernetic brain: Sketches of another future*. 2010, Chicago: University of Chicago Press.
17. van Zoelen, E.M., K. van den Bosch, and M. Neerincx, *Becoming Team Members: Identifying Interaction Patterns of Mutual Adaptation for Human-Robot Co-Learning.* Frontiers in Robotics and AI, 2021. **8**.
18. Binns, R., *Fairness in Machine Learning: Lessons from Political Philosophy.* Proceedings of Machine Learning Research, 2018. **81**: p. 149-159.
19. Weizenbaum, J., *Computer Power and Human Reason: From Judgment To Calculation*. 1976, New York and San Francisco: W.H. Freeman and Company.
20. Schwartz, B. and K. Sharpe, *Practical Wisdom: The Right Way to Do the Right Thing*. 2011, New York: Riverhead Books (Penguin Group).