AI & SOCIETY (2025) 40:3615–3626
https://doi.org/10.1007/s00146-024-02101-z

OPEN FORUM

We need better images of AI and better conversations about AI

Marc Steen[1] · **Tjerk Timan**[1] · **Jurriaan Van Diggelen**[1] · **Steven Vethman**[1]

Received: 16 January 2023 / Accepted: 4 October 2024 / Published online: 29 October 2024

Abstract
In this article, we critique the ways in which the people involved in the development and application of AI systems often visualize and talk about AI systems. Often, they visualize such systems as shiny humanoid robots or as free-floating electronic brains. Such images convey misleading messages; as if AI works independently of people and can reason in ways superior to people. Instead, we propose to visualize AI systems as parts of larger, sociotechnical systems. Here, we can learn, for example, from cybernetics. Similarly, we propose that the people involved in the design and deployment of an algorithm would need to extend their conversations beyond the four boxes of the *Error Matrix*, for example, to critically discuss *false positives* and *false negatives*. We present two thought experiments, with one practical example in each. We propose to understand, visualize, and talk about AI systems in relation to a larger, complex reality; this is the requirement of *requisite variety*. We also propose to enable people from diverse disciplines to collaborate around *boundary objects*, for example: a drawing of an AI system in its sociotechnical context; or an 'extended' Error Matrix. Such interventions can promote meaningful human control, transparency, and fairness in the design and deployment of AI systems.

Keywords Cybernetics · Sociotechnical system · Error matrix · Autonomy · Justice

1 Introduction

It goes without saying that AI systems offer amazing opportunities to be used as tools to do good, as well as tremendous risks to be used as tools to do evil. Therefore, it is pertinent that we organize the development and deployment of AI systems with great care. Here, 'we' refers to the ambition to involve people with diverse backgrounds and roles in such development and deployment: people from disciplines such as technology, law, ethics, social science, public administration, and economics; experts from the domain in which a particular AI system will be used, e.g., public safety; and also potential users or putative beneficiaries of these systems. For public safety, this could be police officers, who will use a particular system, and citizens, for whom this system is meant to promote safety. We can build on the tradition of Participatory Design (Schuler and Namioka 1993), which advocates active participation of future users in the development and deployment of technologies (Doorn et al. 2013). Furthermore, 'great care' refers to the ambition of Responsible Innovation (Stilgoe et al. 2013; Von Schomberg and Hankins 2019), e.g., to promote human autonomy, transparency, and fairness (Hayes et al. 2020, 2023; Steen et al. 2021b), and enable participants to exercise curiosity, creativity, and practical wisdom (Steen et al. 2021a; Steen 2021). With this paper, we aim to contribute to the design and application of 'trustworthy' (High-Level Expert Group on Artificial Intelligence 2019), 'human-centric' (Bryson and Theodorou 2019) or 'responsible' (Dignum 2019) AI systems.

In our work at TNO, a research and technology organization in the Netherlands with over 3000 people, in multiple projects involving the development and evaluation of AI systems, we have noticed that much of the work is done by people with STEM (science, technology, engineering, and maths) backgrounds, e.g., by data scientists. Sometimes, people with other backgrounds are involved in the larger project; notably, in the project's preparation, *before* development, or in the deployment of the project's results, *after* development. Consequently and critically, it can be rather hard to involve people with backgrounds in, e.g., ethics, law or social science, and to take into account *ethical*, *legal* or *societal* aspects. Such involvement has, however, become pertinent (Van Veenstra et al. 2021).

✉ Marc Steen
marc.steen@tno.nl

[1] TNO, The Hague, Netherlands

Fig. 1 Common images of AI systems: a white, shiny, humanoid robot (left; from https://spectrum.ieee.org/how-will-humans-and-robots-coexist), or a flee-floating, blue, electronic brain (right; from https://neolore.com/how-machine-learning-can-help-your-business/)

In this paper, we make the case for finding better ways to visualize AI systems; and for organizing better conversations about AI systems' practical functioning—in order to enable more realistic and inclusive dialogues in the development and deployment of AI systems. Our paper contains two thought experiments. The first thought experiment deals with the question: *What if we visualize represent AI systems differently?* The second thought experiment deals with the question: *What if we talk about AI systems' functioning differently?* We will propose that we need *better images* and *better conversations*, in order to enable people from diverse disciplines to participate in the design and application of AI systems, so that they can collaborate and jointly promote values such as human autonomy and fairness—which will lead to *better* AI systems.

In our thought experiments, we follow a *sociotechnical systems* approach. In line with literature, we understand a sociotechnical system as a system that consists of *people*, *technology*, and *organization* (Mecacci et al. 2023, note 6) and that is organized in such ways that they can collaborate towards specific goals (Novelli et al. 2023, p. 6).[1] In this *sociotechnical systems* approach, we also explore: the notion of *requisite variety*, i.e., the requirement that any viable system, in order to cope with changes in its environment, needs to have a level of variety that matches the level of variety in the environment; and the role of *boundary objects*, i.e., the ways in, e.g., objects or drawings can facilitate communication and collaboration between people with different backgrounds. We will come back to these topics in Sect. 4. More specifically, we chose to position the examples within our thought experiments in the domain of safety and justice; a domain in which the authors have ample experience (references omitted for the review process).

[1] See also: Brandt 2007, p. 460; Sartori and Bocca 2023, p. 444.

2 We need better images, learning from cybernetics

Our first thought experiment involves the question: *What if we visualize AI systems differently?* The background to this question is the prevalence of images of white, shiny, humanoid robots, and of free-floating, blue, electronic brains—see Fig. 1.

These images can be found in popular media, in commercial messages, and even in scientific articles. Such images convey misleading ideas; as if robots can act autonomously and solve problems independently, and as if AI systems can reason in ways superior to people's ways of reasoning.

Critically, these images inform the mental models of the people involved in the design and application of AI systems (Maas 2023). (See https://www.aimyths.org/ for a diagnosis of misleading representations of AI systems, and betterimagesofai.org for alternative images.) Moreover, these images can be conveyed with only words. If, in a project meeting, somebody says, '*This robot will solve problem X*', then this will typically prove to be unrealistic. Soon enough, they will discover that the practical deployment of the robot requires all sorts of changes in existing processes and in the larger organization. Typically, a robot is not a stand-alone silver bullet solution.

We propose that we can use better images: images that appreciate the complex relationships and interactions between people, machines, and their environment. These images, in turn, can enable more realistic conversations, in which in which people from diverse disciplines can participate and contribute more than they currently typically do (more on that in the second thought experiment).

Our approach concurs with Johnson and Verdicchio (2017, p. 575), who remarked that there is a great deal of 'confusion about the notion of "autonomy" that induces

people to attribute to machines something comparable to human autonomy, and a "sociotechnical blindness" that hides the essential role played by humans at every stage of the design and deployment of an AI system'.

2.1 Learning from cybernetics

For more realistic images, we can turn to cybernetics: a field that was popular in the 1950s and 60s and that studies how people and machines *interact* with each other and with their environment.[2] It focuses on how people, machines, and elements in their environment form complex, adaptive systems. The various elements are connected through diverse types of interactions, notably through *feedback loops*, which enable a system to adapt to changes and circumstances, and to have a more or less stable course. Indeed, the term *cybernetics* refers to the steering a ship, on a stable course, through changing winds and waves. Interestingly, the current *relational turn* (Coeckelbergh 2020; Birhane 2021) also foregrounds such interactions between people and technologies.

The added value of cybernetics with regards to images can be illustrated with a drawing by 'cybernetician' Stafford Beer—see Fig. 2. This drawing shows a complex, adaptive system with diverse components, on different levels of abstraction, connected by multiple feedback loops. The *sociotechnical system* and the *feedback loop* have been recurring themes, from Norbert Wiener's *Cybernetics* (1948/1961, pp. 96–97) all the way to Cathy O'Neil's *Weapons of Math Destruction* (2016), in which she argued, that, without a properly functioning feedback loop, 'a statistical engine can continue spinning out faulty and damaging analysis while never learning from its mistakes' (2016, p. 7).

We propose that these ideas—the *sociotechnical system* and the *feedback loop*—are currently under-utilized in the design and application of AI systems, and we will explore how we can use these for the better. We hasten to remark that some people do talk about AI systems as part of sociotechnical systems, and that many systems do have mechanism that enable operators to modify or correct the system's output. Notably, some AI systems also use feedback to 'learn', e.g., in supervised learning or in reinforcement learning; such systems monitor the effects of their actions and take these into account for future actions. Yet, we propose that there is a tendency to view an AI system as an entity separate from the people who operate and use it. Notoriously, in 2022, one Google engineer went as far as believing that the *LaMDA* chatbot had become a sentient entity.[3]

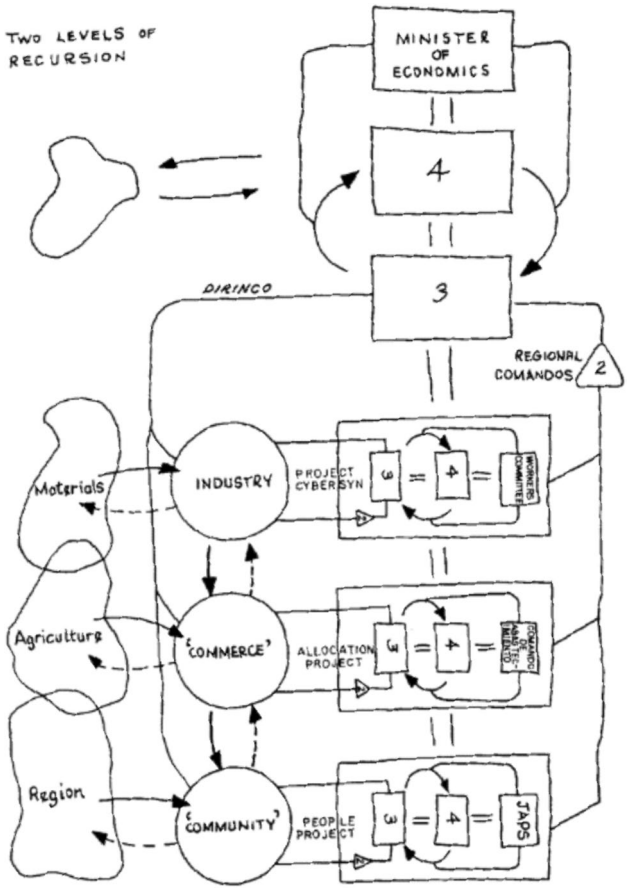

Fig. 2 Drawing of a sociotechnical system, with multiple interactions and feedback loops; by Stafford Beer (1981, p. 325)

2.2 Using better images

As part of our first thought experiment, let us imagine a team with two soldiers and five robots, e.g., the four-legged, dog-like robot. They form a team and have a reconnaissance task: to go into a series of buildings that were recently targeted by adversaries. There may be wounded people in there and it may be dangerous to enter. The robots function as *team members*, with capabilities that are different from people's: on the one hand, they are less skillful in a range of tasks that people are good at, e.g., to interpret a situation in one glance; on the other hand, they can operate in dangerous

[2] The field of cybernetics preceded the field of AI. The term 'Artificial Intelligence' was coined at the 1956 Dartmouth workshop, whereas the Macy Conferences on cybernetics started one decade earlier. Ten of these conferences occurred between 1946 and 1953. Famously, they brought together people from very diverse disciplines. For a decade, the two fields coexisted. In the 1960s, however, proponents of symbolic AI were more successful in gaining research funding—and 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' (Cariani 2010, p. 89).

[3] https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

environments, where we would not want to risk human lives. Practically, the soldiers give orders to the robots, e.g., to move to specific coordinates; the robots then go there and provide images of the situation; the soldiers then use their professional judgement to plan further actions.

Typically, in such a set-up, there are two 'basic' feedback loops—see Fig. 3 (left):

1. The robot uses information from its sensors, e.g., its camera and motion sensors, to modify its movements; this refers to the robot's 'autonomy', e.g., when it climbs a flight of steps;
2. The soldiers use information from the robot, e.g., from its camera, to steer it to new coordinates; they can give commands to the robot, if and when necessary.

Now, if we think of the robots and soldiers as being part of a larger *sociotechnical* system, several more and 'higher' *feedback loops* become available—see Fig. 3 (right):

3. The soldiers can give information to their commanders and ask for further instructions; this would involve additional 'higher ups' in the Human–Machine Teaming—they can make more consequential decisions and can be held accountable for these decisions;
4. The soldiers and their commanders can store information and re-use it for the larger mission; this can enable them to learn across multiple operations or multiple missions—this could also involve 'double loop learning', i.e., learn to become better at learning;
5. There are other feedback loops possible, e.g.: back to the organization that developed the software (so that they can improve the robots); or into the legal system, e.g., into regulations for soldiers' working conditions, or into liability law, for harms that were caused by the robots.

From this example, we can learn the following: understanding and visualizing AI systems as being part of sociotechnical systems (*not* as a humanoid robot or as an electronic brain), can help us to understand an AI system's functioning as a *team member*, as part of Human–Machine Teaming, or as a *tool* that enhances people's abilities for decision making, *supported* by AI. Furthermore, it offers ways to involve people with diverse backgrounds, to deploy more and diverse feedback loops—which, in turn, offers more and diverse opportunities for learning. Moreover, viewing AI systems through a cybernetics lens can help to draw attention to changes that happen over time: to adapting to circumstances and to co-learning (Van Zoelen et al. 2021). The soldiers and the robots can engage in co-learning (Schoonderwoerd et al. 2022); they can mutually learn to adapt their behaviors. Likewise, the police officers' usage of the ADM would change over time. One specific officer can make the tool into a *personal* tool; like how craftspeople can make their tools more personal. These examples raise a series of questions, notably regarding reliability and safety, like: *Can such learning robots and modified tools remain reliable and safe over the course of time?*

2.3 Benefits of better images

Using such alternative images of AI can help to avoid unrealistic expectations, such as: *'This robot will solve this problem'* or *'this algorithm can predict fraud.'* There are ample examples of how such expectations have proven to be unrealistic, e.g., in self-driving cars (Marx 2022). Understanding and visualizing a particular AI system as part of a larger, sociotechnical system is more realistic. In addition , it is more complex. But the world is complex; our problems, and the solutions they require, are complex. In addition, zooming-out to see a larger, sociotechnical system with all sorts

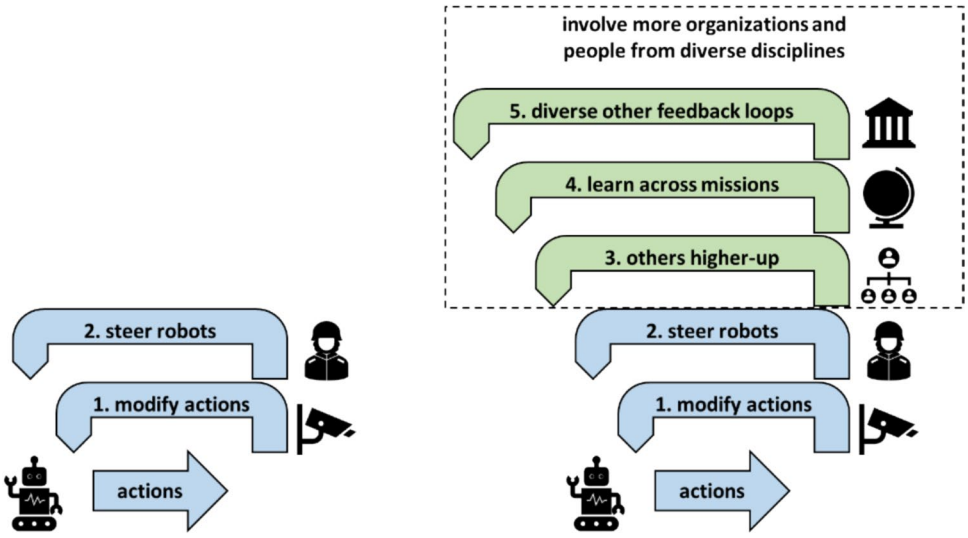

Fig. 3 Soldiers using robots (left); a cybernetics view creates opportunities for additional feedback loops and for involving more organizations and people from diverse disciplines (right)

of feedback loops, can encourage people from diverse disciplines to participate and contribute—which is needed, both to better understand the problem and to explore and develop solutions (Steen 2013). Notably, it can enable people with expertise on ethical, legal, and societal aspects to participate. Moreover, it can help to involve various types of 'users' of the design and application of an AI system, e.g., operators, soldiers, police officers, tax inspectors, and citizens.

More specifically, we propose that cybernetics—notably the notion of feedback loops—can help to better organize *Human–Machine Teaming* (Van Diggelen et al. 2018); notably this can help to clarify the roles and responsibilities of various actors on different levels of aggregation (Fig. 3, right), and can thereby promote Meaningful Human Control (Santoni de Sio and Van den Hoven 2018; Santoni de Sio et al. 2022; Umbrello 2021; Verdiesen et al. 2021).

Critically, we would need to take care that the feedback reaches the right people, at the right moment, in the right form, so that they can act upon it effectively. Zooming-out to the level of society, we can imagine feedback loops that feed into policy cycles (Rahwan 2018; Santoni de Sio and Mecacci 2021). Moreover, zooming-out can help to steer clear from the pitfall of too much focus on technology. A conversation about a robot's reliability (first example) or about an algorithm's fairness (second example) will need to zoom-out and look at the processes and organization in which the system is deployed: *How is reliability promoted in practices of professionals who use this robot? How is fairness promoted in the organization in which this algorithm is deployed?* The people involved will need to carefully organize such higher order feedback loops (Steen et al. 2021b.

3 We need better conversations, extending the Error Matrix

In our second thought experiment, we will explore ways to organize and facilitate better conversations about AI systems and their practical functioning. Sadly, we have become familiar with errors of AI systems that have led to all sorts of ethical, societal, legal, and economic harms. Many data scientists, and also people from the general public, have read books such as *Weapons of Math Destruction* (O'Neil 2016), *Automating Inequality* (Eubanks 2017) or *Algorithms of Oppression* (Noble 2018), and know about grave errors in the domain of justice and security. Well-known examples of suspect algorithms include *COMPAS* (Correctional Offender Management Profiling for Alternative Sanctions), which judges in the US used to assesses risks of recidivism and which had racial biases (Binns 2018; Lagioia et al. 2023), and *SyRI* (System for Risk Indication), which Dutch government agencies used to detect fraud in social benefits, allowances, and taxes, and which was found in breach of human rights, notably Article 8 (Right to respect for private and family life) of the European Convention on Human Rights (Rechtspraak 2020; Wieringa 2023).

Especially notorious are the *false-positive* errors: people whom the system flags as fraudulent, while in reality they are not. Explicitly or implicitly, people often use an *Error Matrix* to discuss a system's intended outcomes: *true positives* (correctly flag cases of fraud) and *true negatives* (correctly non-flag cases of non-fraud); and its unintended errors: *false positives* (incorrectly flag as fraud; was non-fraud) and *false negatives* (incorrectly non-flag non-fraud; was fraud)—see Fig. 4.

In our work, we have noticed that these four categories do not cover the complex reality that our partners or clients deal with. We often hear utterances like: 'This algorithm is 98% accurate'. Critically, many assumptions and conditions need to be in place in order to call something '98% accurate'. In addition, there is the reality of the need for data collection and cleaning, which often requires lots of human labor, behind the scenes, out of sight (Crawford 2021), and diverse modifications in the working processes and in the larger organization that seeks to deploy such an AI system.

Our second thought experiment involves the question: *What if we talk about AI systems' functioning differently?* We propose to organize better conversations about AI systems by enabling people to extend the Error Matrix's four categories; to think outside these boxes. Crucially, we expect that such extending can enable people from diverse disciplines to participate and contribute.

3.1 Extending the Error Matrix

Below, we will focus on systems for Algorithmic Decision Making (ADM), e.g., those used by police officers to detect crimes or by tax inspectors to detect fraud. The risks of such systems—especially in domains such as public administration, justice, and security—have been known for

True positives (correctly flag as fraud)	**False positives** (incorrectly flag as fraud; ***was non-fraud***)
False negatives (incorrectly non-flag non-fraud; ***was fraud***)	**True negatives** (correctly non-flag non-fraud)

Fig. 4 Error Matrix for a system that aims to detect and flag fraud; with true positives, false positives, true negatives and false negatives

years and include, e.g., stigmatization and discrimination (Mittelstadt et al. 2016; O'Neil 2016; Spielkamp 2019).

In order to contribute to the design and deployment of more fair (or less unfair) systems we need to involve experts from diverse disciplines; experts on technology and on ethical, legal, and societal aspects. It is, however, often rather challenging for people from different disciplines to collaborate. Even finding a shared understanding regarding basic concepts like *fairness* or *bias* can be challenging. We have observed misunderstandings about such concepts in numerous projects. A data scientist, an AI expert, a lawyer, a moral philosopher, a civil servant ('user'), a citizen ('data subject') or a project leader (or other decision maker) can have very different understandings of *fairness* or *bias*.

Typically, the people involved in the design and application of such an algorithm will (implicitly or explicitly) use an Error Matrix to discuss the system's performance, to optimize its functionality and to minimize its errors. Critically, however, its four categories are, very often, too rigid and too much of a simplification. The requirement of *requisite variety* would ask for more flexibility and more complexity. In addition, the categories contain assumptions that can limit the types of discussions people can have. We, therefore, propose to extend the Error Matrix's four categories, in order to promote collaboration between people from different disciplines.

Often and implicitly, data scientists and software developers use a range of assumptions, which they refer to as the 'ground truth', when they work on an algorithm (Rommes 2013). In the metaphor that proposes that '*the map is not the terrain*', they construct a map out of data, in order to describe the terrain, which they cannot access directly. They need to 'ground' their map on the data they have, and assume that these data convey 'truth' about the terrain. Looking at the terrain, however, there are many cases that do not neatly fit into the map and its categories. From a sociotechnical system perspective, we would need to make and use maps that not only includes the *technology*, but also the *people* who will be affected by the system's output, and the *organization* that will use the system.

3.2 Organizing better conversations

We can imagine extending the Error Matrix's four boxes in order to enable people with different backgrounds and roles to participate more actively, with an example of an algorithm to detect and flag cases of fraud. Let us look at the original four boxes in turn and imagine how we can extend these boxes—see Fig. 5; the letters A-F are discussed as follows:

True positives (*correctly flag as fraud*); ideally, this box contains people who commit fraud.

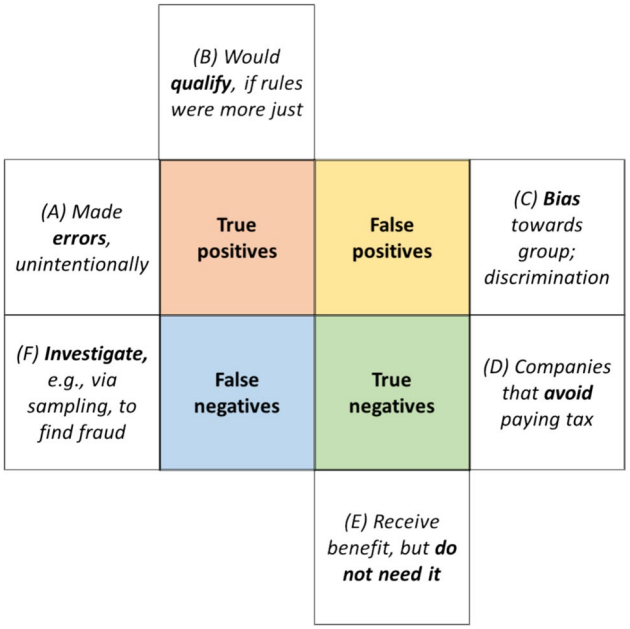

Fig. 5 Example of an Extended Error Matrix

- It is, however, possible that a person made an error, unintentionally, in filling-out some form (A). Some forms *are* notoriously complex. Or they forgot to mention one detail. The algorithm will flag these cases or people as 'fraud'. But is this fair? This can be a starting point for a dialogue between a data scientist, an expert in administrative law, and an expert in human-centered design and usability.
- Moreover, this box can include people who would qualify for a specific benefit, but due to an over-stringent implementation of rules, they commit 'fraud' (B). They may, e.g., occasionally receive groceries from a family member and not report this. This counts as fraud. But is that fair? This can be a starting point for a dialogue between people in administrative law, human rights law, and applied ethics.

False positives (*incorrectly flag as fraud*—was non-fraud); we can extend this box as follows:

The system can be biased towards some *type* of false-positive errors, and thus stigmatize or discriminate against specific groups (C). This is what happened in the Dutch childcare benefits scandal (Amnesty International 2021). Bothering people—sometimes repeatedly—with incorrect data and unjustified accusations, and an exasperating process of investigation and correction, on top of that, is not a fair way to deal with innocent people.

True negatives *(correctly non-flag non-fraud)*; typically, for algorithms that aim to detect fraud or crime, a majority

of cases goes into this box—we can, however, look at this box as follows:

- We can look at companies that spend millions on lawyers to evade taxes and save hundreds of millions (D). Technically, this is not fraud. Increasingly, however, such companies are critiqued. We can zoom-out and see that this fraud detection system looks for small fish (citizens who depend on allowances, some of whom commit fraud for hundreds or thousands of euros), and ignores big fish (corporations that routinely avoid paying hundreds of millions of euros in tax).
- We can also look at the legal rules that make citizens eligible for some benefit that they do *not really need* (E). E.g., a reduction on VAT for energy for *all* citizens, regardless of their income or assets. It sounds laudable to treat citizens equally. But what about equity? It would be fairer to give priority to people who actually need this benefit or allowance. This can be a starting point for a dialogue between people with backgrounds in law, public administration, and ethics.

False negatives (*incorrectly non-flag non-fraud*—was fraud); these case are not flagged for fraud but actually were fraudulent—these are cases of fraud that the algorithm was unable to find. Crucially, this box has a systemic problem. Without further investigation, e.g., through a sample of all 'negatives', it is impossible to distinguish between 'false negatives' and 'true negatives'. On the surface, they look exactly the same: 'nothing to see here', 'no reason to further investigate'.

By definition, this box contains cases that would, after investigation, prove to be fraud. They are, however, *not* flagged and are therefore *not* further investigated. Such investigations, however, are very rarely organized and conducted. From a fairness perspective, it would be better, however, if not only 'usual suspects' (e.g., C) are scrutinized, but also, e.g., a random sample of *not-usual-suspects*—persons whom the system is currently biased towards non-flagging.

To further illustrate how the Error Matrix can be extended, we provide one example. It deals with an algorithm (*Smart Check*) that assesses the likelihood that a citizen is eligible for a specific benefit.[4] The aim of using this algorithm is to *avoid* paying a benefit to a person who is *not eligible*—and thus to avoid having to reclaim these benefits later, which can cause a lot of harm (Amnesty International 2021). Predictions of non-eligible cases are flagged for further investigation. An interdisciplinary discussion of using this algorithm in a rigid manner reveals two systemic problems: the problem of *false negatives*; and the seemingly deterministic line between *true positives* and *false positives*.

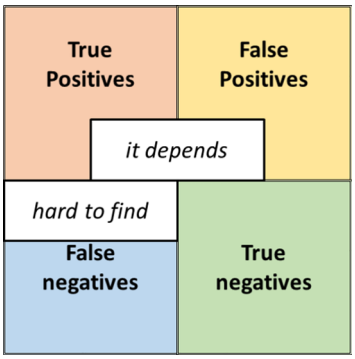

Fig. 6 Error Matrix, with false negatives 'hard to find', because negatives are rarely investigated; and for true and false positives, 'it depends' on people and processes)

First, recall that *false negatives* are not flagged and therefore not investigated. However, if they *had been* investigated, they *would have been* found to be *non-eligible*. This is a problem in at least two ways: these people's fraud may be found out later, and then they will have to pay back the wrongly received benefits; and, handing out benefits to non-eligible persons is a matter of wrongly allocating scarce resources. This is a systemic problem: one focuses further investigation only on cases that are commonly 'wrong' or the 'target', in this case, the people who are commonly not eligible. Less stereotypical cases of people that were not eligible will be harder to find by an algorithm as they do not form a pattern in the data used to train the model. As a result, *negatives* are not or rarely investigated, so that *false negatives* are hard to find—see Fig. 6 (*'hard to find'*).

Second, conversations led by the *regular* Error Matrix divide *positive* cases between those worth the investigator's time as they were indeed not eligible (*true positives)* and those that were eligible despite the algorithm's prediction (*false positives*). This fine line between the true and false positives is largely dependent on the investigator's ability to detect eligibility or ineligibility of the citizen, and on the citizen's ability to provide evidence that they are eligible—or hide evidence that they are *not*. Regardless of protocols, these are human processes and human judgement and various questions: *How should an investigator act if they have received hints about a citizen hiding cars at their friends' houses under different names, but without solid evidence? Should this count as a* false positive*, so that similar cases are not predicted, and thus not further investigated in the future? Or should it be counted as a* true positive *and thus be used to further train the algorithm, based on speculation rather than evidence?* Extending the Error Matrix can

[4] The algorithm is dubbed *Smart Check* (*Slimme Check*, in Dutch) and is described in the Algorithm Register of Amsterdam; https://algoritmeregister.amsterdam.nl/ai-system/onderzoekswaardigheid-slimme-check-levensonderhoud/1086/

facilitate conversations about the dependence on human judgement—see the comment 'it depends' in Fig. 6.[5]

3.3 Benefits of better conversations

Extending the Error Matrix can help to foreground and discuss fundamental or strategic questions about the goals of using an AI system: *What were its goals again? And which criteria or measures can we use to assess whether it can help to achieve those goals?* Ideally, such questions are part of participatory and iterative process, and can help to engage in both problem-setting (rather than uncritically take the initial brief and follow that without questions) and solution-finding (with some going back and forth, in order to explore and evaluate solutions in an iterative fashion) (Steen 2023a, b). Such a process can enable interlocutors to have better conversations about values, such as fairness and transparency (Hayes et al. 2020; Steen et al. 2021b).[6] Moreover, promoting transparency can help citizens, or other people at the receiving end of the algorithm's outcomes, to question and critique the system (Hayes et al. 2023).

Notably, we would expect that bringing together people from different disciplines can help to design better systems. Initially, however, things can become rather confusing. Imagine six people sitting around a table, discussing a prototype ADM system. One talks about fraud that exists (out there, 'the terrain'). Another about fraud that they can detect or can prove (an operational lens, 'the map'). Yet another about fraud that is done intentionally or unintentionally (a legal or moral lens). A fourth person wants to focus on fraud that is significant—large enough to spend money on systems to detect it (an economic or political lens). Facilitating a dialogue between them can be challenging. It can, however, also lead to a better and shared understanding of the project's problem and objective.

4 Discussion

In both thought experiments, we have steered away from unrealistic simplifications and unwarranted enthusiasm about AI systems. Instead, we have tried to usher in a larger, complex reality, into the imaginations and into the conversations of the people involved in the design and application of AI systems. Moreover, we expect that better images and better conversations can help to invite people from different disciplines and with different roles into a project and around to the table.

What happens if we do *not* use such better images or organize such better conversations? Somewhat hyperbolically, we can expect that sales people, higher management, clients, and the general public will believe in the overblown promises of shiny robots and electronic brains that solve all of our problems—a process that currently see, for Large Language Models, chatbots, and Generative AI, for example. We are witnessing a logic in which data scientists and software developers lead AI projects, and people with expertise in ethical, legal or societal aspects are seen as less relevant, and not worthwhile to be involved, and being fired, dismissed or ignored. Big Tech does not like to be bothered too much with social consequences or ethical deliberation.[7] A popular strategy is to spread images and stories of oversimplified, clean, and frictionless AI systems. Where some may react with a plea to the innocence of such images and stories (*'I just quickly needed a visual for our social media post')* the effects of such oversimplification has far-reaching consequences both of the general societal understanding of what AI is an can (not) do, as well as for scientific and policy agendas. Simplified images and stories of AI can lead to societal harms such as discrimination, manipulation, and centralization of power. They also lead to scientific harms, such as ignoring or neglecting studies of social consequences or ethical deliberation, in order to be viewed as innovative, and too much hope is placed in AI as a means to achieve scientific progress.[8] We also see harms to the functioning of policy and government processes, as many areas of police and government are rushing to play a part in the so-called 'AI race', leading to a distorted allocation of resources and attention towards techno-solutionism and quick, and often misguided, regulatory fixes. The latter is happily fed by private AI companies and investors that point to so-called 'existential' and reputational risks for

[5] This problem can be addressed by a procedure suggested by Vethman et al. (2024). This procedure starts with a random sample, to decide who is further investigated; this provides information for a certain number of citizens on whether they are eligible or not. Then, ex-post, the algorithm makes suggestions for those same citizens (that were already investigated) on whether it suggests whether they should be further investigated (*positives)* or not (*negatives). This* procedure enables the detection of *false negative* cases without yet impacting any citizens to algorithm in development. It also provides a safe space to start a conversation on the implication of using an algorithm, here in particular on the reliability of the current investigation process of detecting someone's eligibility for social welfare.

[6] Also, writing this manuscript has been an example of transdisciplinary collaboration—and reflexivity (Steen 2021); it involved discussions between the authors, working from different disciplinary angles: as an expert in ethical and societal aspects, as an expert in administrative law, as an AI expert, and as a bias researcher with a data science background.

[7] https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/

[8] For example: https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-science_a8d820bd-en or https://adr-association.eu/wp-content/uploads/2023/06/ADRA-roadmap-May23_v2-3.pdf.

countries not partaking in this race (Bareis and Katzenbach 2022). Much more can be said on the ways in Big Tech shapes innovation, its societal and environmental hazards, and the limited impact of regulation (Sharon and Gellert 2023). In the two examples in this paper, we tried to show how insights from cybernetics can provide inroads to offer alternative ways to look at, and talk about, AI.

In our first thought experiment, we critiqued popular images of shiny, humanoid robots and of free-floating, artificial brains. We turned to the field of cybernetics in order to understand AI systems as elements in larger, sociotechnical systems, in which people and machines interact with each other and with the environment. This can help the people involved to visualize more realistically the *sociotechnical systems* in which AI systems are used, and to focus on how people can use and control AI systems, and on the diverse organizations and processes that play critical roles. Using more realistic images can help to design and apply systems in ways that support and enhance people's abilities (High-Level Expert Group on Artificial Intelligence 2019), rather than replace people or corrode their abilities. Such an approach foregrounds people's abilities and provides ways to enable people, e.g., professionals, such as soldiers, police officers or tax inspectors, to cultivate and express sociotechnical virtues, such as self-control or justice (Vallor 2016).

Furthermore and regarding questions about human autonomy and 'autonomous' systems, it may be relevant to follow Shneiderman's (2020) critique of using a one-dimensional scale, as is very often done, in which a machine's 'autonomy' grows, e.g., from level 1 to level 5, at the expense of human autonomy. Instead, Shneiderman proposes two perpendicular axes: of one for *human control* and one for *computer automation*, and to combine high levels of *human control* with high levels of *computer automation*, in order to create reliable, safe, and trustworthy AI systems.

In our second thought experiment, we looked at the ways in which people discuss an AI system's performance. Typically, the people involved in the design and application of, e.g., an algorithm for ADM, use an Error Matrix to discuss the system's intended results, e.g., to flag cases of fraud and non-flag cases of non-fraud; and two types of errors: *false positives* and *false negatives*. We argued that these categories are too much of a simplification (the 'map') of the complex reality that they refer to (the 'territory'). Sticking too rigidly to these categories can cause harms, like treating a person who made a mistake in filling-out a form as fraudulent, or applying legal rules so stringently, that very mild cases of fraud lead to harsh punishments. We proposed to extend the Error Matrix in order to have better conversations. This can help the people involved to more realistically talk about the larger sociotechnical system—and not restrict themselves to strictly technology-oriented evaluations and measures (Weerts et al. 2024). Moreover, we experienced that questioning these categories can facilitate conversations between people from different disciplines, e.g., from data science, AI, administrative law, and human-centered design, and can facilitate more informed discussions in public arenas. It remains to be seen, however, how such an approach would work in practice, in day-to-day innovation projects in companies or organizations. It will be interesting to see which person can take the initiatives and at what level of responsibility (Georgieva et al. 2022), and whether there is sufficient room and agency for project team members to question and critique, e.g., the Error Matrix. Is the culture in the organization or project sufficiently safe for that?

Here, we turn to cybernetics, and other fields that look at the co-shaping of technology and society, for inspiration. Regarding the division of tasks between people and machines, we can learn from Weizenbaum's (1976) proposal to distinguish between making *decisions*, i.e., what computers can do, through calculation; and making *choices*, i.e., what people can do, through judgement. His proposal is to allocate tasks wisely: tasks that require number crunching and calculation can be delegated to machines, whereas tasks that require, e.g., moral judgement, need to be done by people. These distinctions are not always clear cut. Some have envisioned, e.g., ways in which AI systems can support people in moral tasks (Haselager and Mecacci 2020).

There are two more topics that we mentioned in the introduction and that we would like to discuss briefly, as inspiration to combat oversimplification of AI and its impacts: the need for requisite variety; and the role of boundary objects.

4.1 Requisite variety

A key concept in cybernetics is the idea of *requisite variety* (Ashby 1958); it refers to the requirement that any viable system, in order to cope with changes in its environment, needs to have a level of variety that matches the level of variety in the environment. A system that operates in a simple environment needs to be accordingly simple; e.g., a system that performs a series of fixed tasks in an assembly line. A system that operates in a complex environment, however, needs to be accordingly complex; e.g., a system that detects all sorts of fraud by all sorts of people. Making a complex system for a simple task makes little sense. Similarly, a simple system for a complex task would be unwise.

Now, the ideas conveyed by images of humanoid robots or of electronic brains are of simplicity. *Just buy the robot, it will solve your problem. Just run the algorithm, it will reduce operational costs*.

The situations in which these systems will be deployed are, however, not simple. In the real world, there are not only fraudsters or non-fraudsters; there are also people who make mistakes, people who are victims of overly stringent regulation, and people who can afford to pay lawyers to evade

taxes. Fortunately, there are, of course, systems that do more than give a flag or not give flag. There have been advances in the field of Explainable AI (XAI), that aim to provide information to people who use the system, so that they can better understand how the system calculated its predictions. An algorithm can, e.g., display a certainty percentage next to the flag, or add several keywords, taken from the data that went into the calculation. Human operators can then use their professional discretion and take this additional information into account when choosing follow-up actions.

Our first thought experiment can illustrate that the visualizations that people use would need to be as complex as the situations in which the systems are used. Our second thought experiment can illustrate that the categories ('map') that people talk about would need to reflect the complexity of the world to which these categories refer ('territory').

4.2 Boundary objects

We would like to draw attention to the role of *boundary objects* (Star 2010): a concept from the field of Science and Technology Studies (STS). Social scientists, e.g., use this concept of boundary objects to (retrospectively) describe practices in which people use specific objects (more or less consciously) to communicate and collaborate across disciplinary boundaries (hence the name). Here, however, we propose to use the concept slightly differently. Following recent scholarship (Islind et al. 2019; Beck et al. 2021), we propose to use boundary objects to (prospectively) facilitate communication and collaboration between people from different disciplines. For this purpose, Carlile (2002) distinguished three functions of boundary objects: a *syntactic* function, to *transfer* knowledge, e.g., by using a shared lexicon; a *semantic* function, to *translate* knowledge, which acknowledges that people can interpret the same term differently; and a *pragmatic* function, to *transform* knowledge, e.g., into prototypes, which appreciates also that people can have different or conflicting interests.

Using the three functions of Carlile (2002), the interlocutors can develop a shared lexicon (*syntactic*) (e.g., '*What do you mean with fair?'*); make translations between disciplines (e.g., '*I would call the labeling of an unintentional mistake a false positive')* (semantic); and jointly work on improving the algorithm (e.g., ask: *What* was the overall goal *of the algorithm again?* (pragmatic). Interestingly, Carlile (2002) warns us that knowledge does not necessarily lead to collaboration and innovation: 'knowledge is both a source of and a barrier to innovation. The characteristics of knowledge that drive innovative problem solving within a function actually hinder problem solving and knowledge creation across functions.' This points to the challenges of boundary crossing.

In our first thought experiment, we proposed that the people involved can use more realistic drawings of the sociotechnical system in which the AI system will be used. Such drawings can function as boundary objects and enable people from different disciplines, or with different roles, to communicate and collaborate. They can, e.g., make drawings on a white board in a project meeting, and refer to these in their conversations. In our second thought experiment, we proposed that the people involved in development and deployment can use an Extended Error Matrix as a boundary object. They can, very practically, make a drawing, put it on the table, and discuss around it.

Extending the Error Matrix can, at first, feel like making matters unnecessarily complex: *Do we need to discuss a basic term like fairness? Do you* really *want to discuss the project brief?* Following the requirement for *requisite variety*, however, the people involved would need to make the system, *and the conversations about it*, as complex as the environment in which it will be deployed. Furthermore, the people involved can choose, more consciously than they typically would, to simplify matters. They can keep track of their assumptions, so that they can revisit them, if needed. This can function as a feedback loop in the project. Moreover, a bit of discussion at the start of a project can lead to more clarity and shared understanding in the project, which will help to create a better system.

4.3 In closing

Based on two thought experiments (above), and the insights derived from cybernetics, and other fields that investigate AI as a sociotechnical system, we can articulate the following invitations to the various people who are involved in the design and deployment of AI systems:

- *Please use better images; start with the practices and capabilities of professionals—military personnel, in our example—and envision AI systems as parts of larger sociotechnical systems, with feedback mechanisms, to support these professionals in their work.*
- *Please organize better conversations; enable people from different disciplines to discuss their assumptions and approaches in designing and deployment of AI systems—e.g., how to understand and deal with false positives and false negative errors.*

We can see examples of the opposite around us: what happens when people use images of shiny, humanoid robots and of free-floating, electronic brains; what happens when people use an Error Matrix too rigidly and do not question its assumptions. Too often, we act unrealistically and unwisely, as if robots and algorithms can magically solve complex problems. Surely, we can do better.

Funding Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWA.1306.19.021

Data availability Not applicable.

References

Amnesty International (2021) Xenophobic achines: discrimination through unregulated use of algorithms in the Dutch childcare benefits. Amnesty International, London, UK

Bareis J, Katzenbach C (2022) Talking AI into being: the narratives and imaginaries of national AI strategies and their performative politics. Sci Technol Human Values 47(5):855–888. https://doi.org/10.1177/01622439211030007

Beer S (1981) Brain of the firm, 2nd edn. John Wiley, Chisester

Binns R (2018) Fairness in machine learning: lessons from political philosophy. Proc Mach Learn Res 81:149–159

Birhane A (2021) Algorithmic injustice: a relational ethics approach. Patterns 2(2):100205. https://doi.org/10.1016/j.patter.2021.100205

Brandt D (2007) The global technology laboratory. AI Soc 21(4):453–470. https://doi.org/10.1007/s00146-007-0082-9

Bryson JJ, Theodorou A (2019) How society can maintain human-centric artificial intelligence. In: Toivonen M, Saari E (eds) Human-centered digitalization and services. Springer, Singapore

Cariani P (2010) On the importance of being emergent. Construct Found 5(2):86–91

Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 26(4):2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven and London

de Santoni Sio F, Mecacci G (2021) Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 34(4):1057–1084. https://doi.org/10.1007/s13347-021-00450-x

de Santoni Sio F, Van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robot AI 5:1–15

de Santoni Sio F, Mecacci G, Calvert S, Heikoop D, Hagenzieker M, van Arem B (2022) Realising meaningful human control over automated driving systems: a multidisciplinary approach. Mind Mach. https://doi.org/10.1007/s11023-022-09608-8

Dignum V (2019) Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer Nature, Cham, Switzerland

Doorn N, Schuurbiers D, Van de Poel I, Gorman ME (eds) (2013) Early engagement and new technologies: Opening up the laboratory. Springer Science + Business Media, Dordrecht, The Netherlands

Eubanks V (2017) Automating inequality. St. Martin's Press, New York

Georgieva I, Lazo C, Timan T, Veenstra AFV (2022) From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI Ethics 2(4):697–711. https://doi.org/10.1007/s43681-021-00127-3

Hayes P, Van de Poel I, Steen M (2020) Algorithms and values in justice and security. AI Soc 35:533–555. https://doi.org/10.1007/s00146-019-00932-9

Hayes P, Van de Poel I, Steen M (2023) Moral transparency of and concerning algorithmic tools. AI Ethics 3:585–600. https://doi.org/10.1007/s43681-022-00190-4

High-Level Expert Group on Artificial Intelligence (2019) Ethics Guidelines for trustworthy AI. European Commission, Brussels

Verdiesen I, De Santoni Sio F, Dignum V (2021) Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight. Minds Mach 31(1):137–163. https://doi.org/10.1007/s11023-020-09532-9

Johnson DG, Verdicchio M (2017) reframing AI discourse. Mind Mach 27(4):575–590. https://doi.org/10.1007/s11023-017-9417-6

Lagioia F, Rovatti R, Sartor G (2023) Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. AI Soc 38:459–478. https://doi.org/10.1007/s00146-022-01441-y

Maas M (2023) AI is like… A literature review of AI metaphors and why they matter for policy. In: Institute for Law & AI Working Paper Series

Marx P (2022) Road to nowhere: what silicon valley gets wrong about the future of transportation. Verso Books, London, UK

Mecacci G, Calvert SC, de Santoni Sio F (2023) Human–machine coordination in mixed traffic as a problem of meaningful human control. AI & Soc 38(3):1151–1166. https://doi.org/10.1007/s00146-022-01605-w

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679

Noble SU (2018) Algorithms of oppression: now search engines reinforce racism. New York University Press, New York

Novelli C, Taddeo M, Floridi L (2023) Accountability in artificial intelligence: what it is and how it works. AI Soc. https://doi.org/10.1007/s00146-023-01635-y

O'Neil C (2016) Weapons of math destruction. Penguin, London

Rahwan I (2018) Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol 20(1):5–14. https://doi.org/10.1007/s10676-017-9430-8

Rechtspraak, De (2020) SyRI legislation in breach of European convention on human rights. https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-legislation-in-breach-of-European-Convention-on-Human-Rights.aspx

Rommes E (2013) Feminist interventions in the design process. In: Waltraud E, Ilona H (eds) Gender in science and technology. Verlag, Bielefeld, pp 41–56

Sartori L, Bocca G (2023) Minding the gap(s): public perceptions of AI and socio-technical imaginaries. AI Soc 38(2):443–458. https://doi.org/10.1007/s00146-022-01422-1

Schoonderwoerd TAJ, van Zoelen EM, van den Bosch K, Neerincx MA (2022) Design patterns for human-AI co-learning: a wizard-of-Oz evaluation in an urban-search-and-rescue task. Int J Hum Comput Stud 164:102831. https://doi.org/10.1016/j.ijhcs.2022.102831

Schuler D, Namioka A (eds) (1993) Participatory design: principles and practices. Lawrence Erlbaum Associates, Hillsdale, New Jersey

Sharon T, Gellert R (2023) Regulating Big Tech expansionism? Sphere transgressions and the limits of Europe's digital regulatory

strategy. Inf Commun Soc. https://doi.org/10.1080/1369118X.2023.2246526
Spielkamp M (2019) Automating society: taking stock of automated decision-making in the EU. AW AlgorithmWatch gGmbH, Berlin, Germany
Steen M (2013) Co-design as a process of joint inquiry and imagination. Des Issues 29(2):16–29
Steen M (2021) Slow Innovation: the need for reflexivity in responsible innovation (RI). J Responsib Innov 8(2):254–260. https://doi.org/10.1080/23299460.2021.1904346
Steen M (2023a) Ethics as a participatory and iterative process. Commun ACM 66(5):27–29
Steen M (2023b) Ethics for people who work in tech. Routledge/CRC Press, Boca Raton, FL
Steen M, Sand M, Van de Poel I (2021a) Virtue ethics for responsible innovation. Bus Prof Ethics J 40(2):243–268
Steen M, Timan T, van de Poel I (2021b) 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 1(4):501–515. https://doi.org/10.1007/s43681-021-00063-2
Vethman S, Schaaphok M, Hoekstra M, Veenman C (2024) Random sample as a pre-pilot evaluation of benefits and risks for ai in public sector. In: Nowaczyk S et al (eds) Artificial intelligence. ECAI 2023 international workshops. ECAI 2023. Springer, Cham
Stilgoe J, Owen R, Macnaghten P (2013) Developing a framework for responsible innovation. Res Policy 42:1568–1580
Umbrello S (2021) Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach. Ethics Inf Technol 23(3):455–464. https://doi.org/10.1007/s10676-021-09588-w
Van Diggelen J, Neerincx M, Peeters M, Schraagen JM (2018) Developing effective and resilient human-agent teamwork using team design patterns. IEEE Intell Syst 34(2):15–24
Van Zoelen EM, Emma M, Van den Bosch K, Neerincx M (2021) Becoming team members: identifying interaction patterns of mutual adaptation for human-robot co-learning. Front Robot AI. https://doi.org/10.3389/frobt.2021.692811
Van Veenstra, Fleur A, Van Zoonen L, Helberger N, (Eds). 2021. ELSA labs for human centric innovation in AI: Netherlands AI Coalition
Von Schomberg R, Hankins J (eds) (2019) International handbook on responsible innovation: a global resource. Edward Elgar, Cheltenham, UK
Weerts H, Xenidis R, Tarissan F, Olsen HP, Pechenizkiy M (2024) The neutrality fallacy: when algorithmic fairness interventions are (Not) positive action. https://doi.org/10.48550/arXiv.2404.12143. In *arXiv*.
Wiener N (1948) Cybernetics: or control and communication in the animal and the machine. MIT Press, Cambridge, MA
Wieringa M (2023) "Hey SyRI, tell me about algorithmic accountability": lessons from a landmark case. Data Policy 5(2). https://doi.org/10.1017/dap.2022.39
Ashby WR (1958) Requisite variety and its implications for the control of complex systems. Cybern 1(2):83–99
Beck A-MT, Rasmussen BM, Nielsen TKH (2021) Action plans as active boundary objects. Res Soc Work Pract 31(4):382–389. https://doi.org/10.1177/10497315211002637
Carlile PR (2002) A Pragmatic view of knowledge and boundaries: boundary objects in new product development. Organ Sci 13(4):442–455. https://doi.org/10.1287/orsc.13.4.442.2953
Haselager P, Mecacci G (2020) Superethics instead of superintelligence: know thyself, and apply science accordingly. AJOB Neurosci 11(2):113–119. https://doi.org/10.1080/21507740.2020.1740353
Islind AS, Lindroth T, Lundin J, Steineck G (2019) Co-designing a digital platform with boundary objects: bringing together heterogeneous users in healthcare. Health Technol 9(4):425–438. https://doi.org/10.1007/s12553-019-00332-5
Shneiderman B (2020) Human-Centered artificial intelligence: reliable, safe & trustworthy. Int JHuman–Comput Interact 36(6):495–504. https://doi.org/10.1080/10447318.2020.1741118
Star SL (2010) This is not a boundary object: reflections on the origin of a concept. Sci Technol Human Values 35(5):601–617
Vallor S (2016) Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press
Weizenbaum J (1976) Computer power and human reason: From judgment to calculation. W.H. Freeman and Company