Chapter 2
Perspectives on Artificial Intelligence

Abstract A discussion of the ethics of artificial intelligence hinges on the definition of the term. In this chapter I propose three interrelated but distinct concepts of AI, which raise different types of ethical issues. The first concept of AI is that of machine learning, which is often seen as an example of "narrow" AI. The second concept is that of artificial general intelligence standing for the attempt to replicate human capabilities. Finally, I suggest that the term AI is often used to denote converging socio-technical systems. Each of these three concepts of AI has different properties and characteristics that give rise to different types of ethical concerns.

Keywords Artificial intelligence · Definitions of AI · Machine learning · Artificial general intelligence · Socio-technical systems

A good starting point for an introduction to the term "AI" is the 1956 Dartmouth summer research project on artificial intelligence, where the term was coined by McCarthy and collaborators (McCarthy et al. 2006). In their proposal for this project McCarthy et al. suggest that machines can be made to simulate "every aspect of learning or any other feature of intelligence". As features of intelligence, McCarthy et al. cite the use of language, the formation of abstractions and concepts, solving problems now reserved for humans and self-improvement.

This points to the first problem in understanding AI, namely its aim to replicate or emulate intelligence. Intelligence is itself a contested concept and it is not clear which or whose intelligence AI would have to replicate, in order to be worthy of being called AI. Biological organisms, including humans, seem to work on different principles from digital technologies (Korienek and Uzgalis 2002). Humans have access to "mental abilities, perceptions, intuition, emotions, and even spirituality" (Brooks 2002: 165). Should AI emulate all of those?

This, in turn, points to the second problem in understanding AI. Are there barriers that AI, as a digital technology, cannot overcome, aspects of intelligence that cannot be digitally replicated? This is an interesting question that has been debated for a long time (Collins 1990, Dreyfus 1992). It is ethically interesting because it has a bearing on whether AI could ever be considered an ethical subject, i.e. whether it could have

B. C. Stahl, *Artificial Intelligence for a Better Future*,
SpringerBriefs in Research and Innovation Governance,
https://doi.org/10.1007/978-3-030-69978-9_2

moral obligations in itself. This is similar to the question whether computers can think, a question that Alan Turing found "too meaningless to deserve discussion" (Turing 1950: 442) and that prompted him to propose the imitation game, also known as the Turing Test.[1]

Both problems of understanding AI – namely, what is human intelligence and which part of it might be replicable by AI – make it difficult to define AI. The conceptual subtleties of AI have led to a situation where there are many competing definitions covering various aspects (Kaplan and Haenlein 2019). The OECD (2019: 7) suggests that

> [a]n AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

A similarly policy-oriented definition comes from the European Commission (2020a: 2):

> AI is a collection of technologies that combine data, algorithms and computing power.

One of the most cited academic definitions is from Li and Du (2007: 1) and notes that AI combines

> a variety of intelligent behaviors and various kinds of mental labor, known as mental activities, … [to] include perception, memory, emotion, judgement, reasoning, proving, identification, understanding, communication, designing, thinking and learning, etc.

Virginia Dignum, an AI researcher who has worked extensively on ethical aspects of AI, highlights the fact that AI refers not just to artefacts, but also to an academic community. She considers

> AI to be the discipline that studies and develops computational artefacts that exhibit some facet(s) of intelligent behaviour.
>
> Such artefacts are often referred to as (artificial) agents. Intelligent agents are those that are capable of flexible action in order to meet their design objectives, where flexibility includes the following properties …

- Reactivity: the ability to perceive their environment, respond to changes that occur in it, and possibly learn how best to adapt to those changes;
- Pro-activeness: the ability to take the initiative in order to fulfil their own goals;
- Sociability: the ability to interact with other agents or humans.

As this book is about the ethics of AI, I propose a view of the term that is geared towards eludicating ethical concerns. Both the terms "AI" and "ethics" stand for multi-level concepts that hold a variety of overlapping but non-identical meanings. For this reason, I distinguish three aspects of the term AI, all of which have different ethical challenges associated with them.

[1] In the Turing Test a human participant is placed in front of a machine, not knowing whether it is operated by another human or by a computer. Can the computer's responses to the human made through the machine imitate human responses sufficiently to pass as human responses? That is what the Turing Test tries to establish.

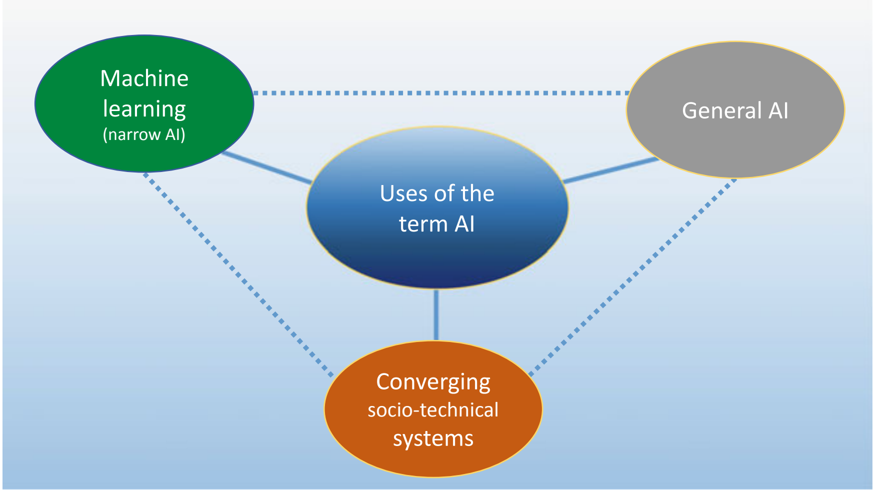
Machine learning (narrow AI)
General AI
Uses of the term AI
Converging socio-technical systems

Fig. 2.1 Uses of the term "AI"

1. machine learning as the key example of a narrow understanding of AI, i.e. as a technique that successfully replicates very specific cognitive processes
2. general AI
3. AI as a synonym for converging socio-technical systems which include but go far beyond narrow AI technologies.

Figure 2.1 gives an overview of use of the term AI that I discuss in this chapter.

2.1 Machine Learning and Narrow AI

A recent review of the AI literature by the academic publisher Elsevier (2018) suggests that there are a number of key concepts and research fields that constitute the academic discipline of AI. Based on a sample of 600 000 AI-related documents, analysed against 800 keywords, the report classified AI publications in seven clusters:

1. search and optimisation
2. fuzzy systems
3. planning and decision making
4. natural language processing and knowledge representation
5. computer vision
6. machine learning
7. probabilistic reasoning and neural networks.

This underlines that AI is not one technology but can better be understood as a set of techniques and sub-disciplines (Gasser and Almeida 2017).

While all these clusters are recognised components of the AI field, the emphasis in current AI ethics is on machine learning and neural networks, clusters 6 and 7. Neither of these is truly novel. Machine learning has been an established part of AI research (Bishop 2006) since its inception, but recent advances in computing power and the availability of data have led to an upsurge in its application across a broad range of domains. Machine learning covers a wide array of techniques and approaches including supervised learning, Bayesian decision theory, various parametric and nonparametric methods, clustering and many others (Alpaydin 2020).

Neural networks are technologies that try to replicate the way in which natural brains are constructed. They represent a bottom-up approach to AI, i.e. a view that intelligence arises from the structure of the brain. Neural networks are not a new idea, but they have only recently achieved success thanks to the availability of large data sets, novel algorithms and increased computing power. Neural networks are an important factor behind the recent success of machine learning, which is the main driver of the current AI wave.

One particular technique of high importance is deep learning (LeCun et al. 2015), which uses different types of neural networks and has contributed to recent successes in areas such as speech recognition, visual object recognition and object detection, as well as other domains such as drug discovery and genomics (Horvitz 2017).

Machine learning, despite its impressive successes, can be characterised as an example of narrow AI. As noted earlier, this is a technique that successfully replicates very specific cognitive processes. It is not able to transfer insights easily from one domain to another. A machine learning system that has learned to distinguish cats from dogs, for example, does not automatically have the ability to recognise natural language or categorise pathology images to identify cancer. The underlying system may well be able to cover other applications but will need to be trained anew for new purposes.

For this book it is important to understand which of the characteristics that machine learning possesses are of ethical relevance. Key among them are the following:

1. *Opacity*: Machine learning algorithms and neural networks are complex to the point that their internal workings are not straightforward to understand, even for subject experts. While they remain purely technical and determined systems, it is impossible (partly because they are learning systems and therefore change) to fully understand their internal working.
2. *Unpredictability*: As a consequence of point 1, the prediction of outputs of the systems based on an understanding of the input is difficult, if not impossible.
3. "*Big data*" *requirement*s: Machine learning systems in their current form require large training datasets and significant computer capacity to create models.

The reference to machine learning as an example of "narrow AI" suggests that there are other types of AI which are not narrow. These are typically referred to as general AI and are discussed in the next section. Before we come to these, it is important to point out that machine learning, with its use of neural networks, is not the only type of narrow AI. Other examples are decision support systems based on decision trees and fuzzy logic systems. I focus on machine learning in this book because it is the most prominent example of narrow AI right now, mostly owing to its recent successes. This is not to say that other examples of narrow AI may not gain similar prominence in future or raise other types of ethical concerns.

2.2 General AI

General AI, sometimes also referred to as "strong AI", goes back to the early days of AI research and stands for the idea that it is possible to build systems that display true human (or other higher mammal) levels of intelligence. It is also known as "good old-fashioned AI" (GOFAI). The original tenet of GOFAI was that the world could be represented through symbols and that the manipulation of these symbols would lead to intelligent behaviour (Moor and Bynum 2002). In this view the human brain was seen as a computer that performs logical operations, and the same or at least functionally equivalent ones could be performed in a digital computer (Floridi 1999).

Maybe the most interesting observation about the GOFAI project is that it has not succeeded in the 65 years since its inception. At this point there is no general AI (Babuta et al. 2020). This indicates that either its assumptions are wrong or they cannot be implemented in the type of digital computer we currently have at our disposal. There are many suggestions about why exactly GOFAI has not (yet) achieved its objectives. One suggestion is that the core of the problem is ontological, i.e. that the world simply cannot be represented comprehensively through symbols that are defined in a top-down manner (Smith 2019). This is the suggestion of phenomenology as expressed in an early critique of AI by Dreyfus (1972).

Another interesting question is whether the current failure of GOFAI is temporary, which would mean that we will be able to build general AI systems at some point, or whether it is fundamental, which would mean that there is some component of true intelligence that is incapable of being captured and reproduced by machines, or at least by the types of digital computers we are using today.

General AI has a strange status in this 2020 AI ethics book. On one hand it seems clear that general AI does not exist. It can therefore arguably not cause ethical concerns and can happily be ignored. On the other hand, general AI is probably the most prominent subject of discussions related to AI and ethics in science fiction, where a large number of characters represent general AI for good or ill. *2001: A Space Odyssey*, featuring the sentient computer HAL, *Blade Runner*, the Terminator movies, *I, Robot*, *WALL-E*, *Westworld* and a host of other stories are about general AI. These narratives cannot be ignored, partly because science fiction is hugely influential in guiding technical design choices, and partly because the public discussion is guided by them. High-profile interventions by celebrities and well-recognised scientists like Elon Musk and Stephen Hawking lend credence to the idea that general AI may create significant ethical risks.

In addition, general AI is of interest because many of the questions it raises are of relevance to ethics. I am agnostic about the possibility of ever creating general AI, partly because I am not sure we understand what constitutes natural intelligence and hence am not convinced that we could recognise general AI even if it appeared. The history of AI has been one of shifting goalposts, and we are now in a world where many of the early dreams of AI have been realised. Examples of successful AI implementation include the ubiquitous voice recognition that is now standard in most smart phones and the ease of organising vast amounts of data that any internet user encounters when using a search engine. Despite these successes few would say that we are anywhere near general AI. For instance, GPS systems integrated into our cars can remember our usual routes to work and suggest the most efficient one depending on current traffic conditions. They also talk to us. At the same time, we are still waiting for Lieutenant Commander Data, the android from *Star Trek: Picard*. General AI is nevertheless an important ingredient in the AI ethics debate because it brings to the fore some fundamental questions about what makes us human, and about what, if anything, the difference is between humans, other animals and artificial beings. Some of the aspects that have led to the failure of general AI so far – namely, the neglect of human nature, and of the phenomenological and existential aspects of being in the world (Heidegger 1993, Moran 1999, Beavers 2002) – are crucial for ethics and I will return to them in the next chapter.

The relevant characteristics of general AI are:

1. Nature of intelligence: General AIGeneral AI raises the question of what constitutes intelligence.
2. By implication, general AIGeneral AI points to fundamental questions such as:
    a. Human nature: What does it mean to be human?
    b. Nature of reality: What is reality?
    c. Nature of knowledge: What can we know about reality?

General AI thus points us to some of the most fundamental philosophical questions, many of which may not have an answer or may have many inconsistent answers but are important for humans to ask to make sense of their place in the world.

While narrow AI and general AI are widely recognised concepts in the AI literature, there is another meaning of the term AI that is of high relevance to the AI ethics debate, even though it is not strictly speaking about AI in a technical sense.

2.3 AI as Converging Socio-Technical Systems

There are numerous fields of science and technology that are closely linked to AI and that are often referred to in discussions about AI. Some of these are technologies that produce the data that machine learning requires, such as the internet of things. Others are technologies that can help AI to have an effect on the world, such as robotics (European Commission 2020b). One could also use the term “smart information system” (SIS) to denote this combination of several types of technologies, which typically are based on machine learning and big data analytics (Stahl and Wright

2018). In practice AI rarely appears as a stand-alone technology but is usually linked to and embedded in other technologies.

The distinction between different technologies is increasingly difficult. Fifty years ago, a computer would have been a readily identifiable large machine with clearly defined inputs, outputs and purposes. Since then the increasing miniaturisation of computing devices, the introduction of mobile devices, their linking through networks and their integration into communications technologies have led to a situation where computing is integrated into most technical devices and processes. AI tends to form part of these technical networks.

Some authors have used the abbreviation NBIC (nano, bio, information and cognitive technologies) to denote the apparent convergence of these seemingly different technologies (Khushf 2004, Van Est et al. 2014). AI and brain-related technologies have a central role in this convergence.

Perhaps not surprisingly, there is much work that links AI with neuroscience, the scientific study of the brain and the nervous system. Since the brain is the seat of human intelligence, research on the brain is likely to be relevant to understanding artificial as well as natural intelligence. AI has always drawn from our understanding of the brain, with artificial neural networks being a prominent example of how neuroscientific insights have influenced AI development. At present there is much interest in what neuroscience and machine learning can learn from each other (Marblestone et al. 2016, Glaser et al. 2019) and how neuroscience and AI research, in their further progress, can support each other (Hassabis et al. 2017). One hope is that neuroscientific insights may help us move beyond narrow AI to general AI, to the development of machines that "learn and think like people" (Lake et al. 2017).

The term "AI" in this context is thus used as shorthand for technical systems and developments that have the potential to grow together, to support and strengthen one another. Crucially, these systems are not just technical systems but *socio*-technical systems. While this is true for any technical system (they never come out of nothing and are always used by people) (Mumford 2006), it is particularly pertinent for the converging technologies that include AI. Examples of such socio-technical systems include most of the high-profile examples of AI, such as autonomous vehicles, embedded pattern recognition – for example, for the scrutiny of CVs for employment purposes – and predictive policing. All of these have a narrow AI at their core. What makes them interesting and ethically relevant is not so much the functioning of the AI, but the way in which the overall socio-technical system interacts with other parts of social reality.

This use of the term "AI" to denote socio-technical systems containing AI and other technologies points to some characteristics of these technologies that are ethically relevant. These socio-technical systems appear to be autonomous, i.e. they create outputs that affect people in ways that do not allow responsibility to be ascribed to human beings. This does not imply a strong concept of the autonomy of AI, a concept I will return to in the following chapter, but rather a lack of visible oversight and control. For instance, if embedded pattern recognition is used to scan CVs to identify candidates suitable for interviewing, the system is not an example of strong autonomy (as a human short-lister would be), but the ethical issues in terms of oversight are still obvious.

Another important aspect of these systems is that they structure the space of options that individuals have. Coeckelbergh (2019) uses the metaphor of theatre roles. Drawing on Goffman (1990), Coeckelbergh argues that human actions can be seen as embodied performances. The scope of content of these performances is structured by what is available on the stage. AI-driven socio-technical systems take the role of the theatre, often of the director. Even if they do not directly instruct humans as to what they should do (which is also often the case; think of the Uber driver receiving her instructions from her phone), they determine what can or cannot be done. Where humans are not aware of this, such a structuring of options can be seen as a covert manipulation of human actions. And, given the economic and social reach and importance of these technologies, the social impact of these systems can be significant. For instance, the use of an internet search engine and the algorithms used to determine which findings are displayed structure to a large extent what users of this search engine are aware of with regard to the search. Similarly, the information made available to social media users, typically prioritised by AI, can strongly influence people's perception of their environment and thereby promote or limit the prevalence of conspiracy theories. To summarise, the AI-enabled socio-technical systems have the following characteristics.

1. *Autonomy*: AI socio-technical systems lead to consequences for humans that are not simply results of identifiable actions of human beings.
2. *Manipulation*: AI socio-technical systems structure human options and possible actions, often in ways that humans do not realise.
3. *Social impact*: Consequences for individuals and society of the use of AI socio-technical systems can be significant.

Figure 2.2 provides a graphical representation of the features of the different meanings of AI discussed in this chapter.

This view of AI and its sub-categories helps us better understand and deal with the ethical issues currently discussed in the context of AI. It should be clear, however, that I do not claim that it is the only way of categorising AI, nor would I argue that the three categories are distinctly separate. Machine learning may well hold the key to general AI and it certainly forms part of the converging socio-technical systems. Should general AI ever materialise, it will no doubt form a part of new socio-technical systems. The purpose of the distinction of the three aspects is to show that there are different views of AI that point to different characteristics of the term, which, in turn, raises different ethical issues. It therefore facilitates engagement with ethical issues.

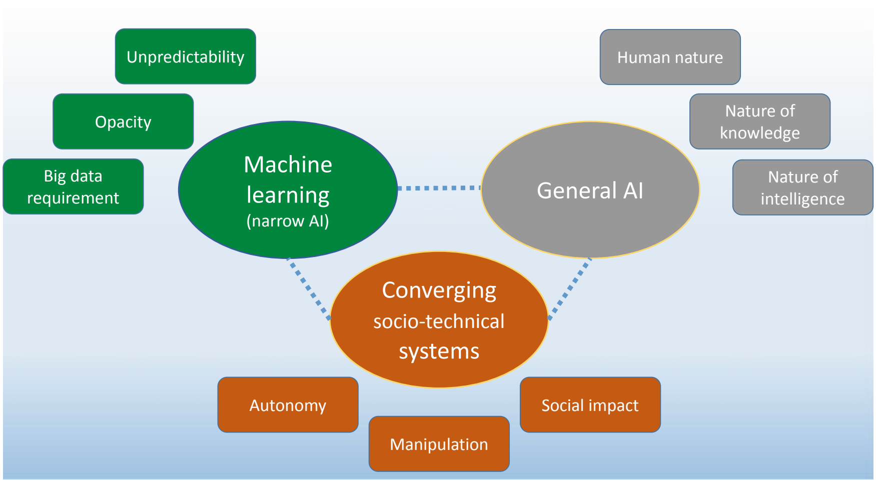

Fig. 2.2 Key characteristics of the different uses of the term "AI"

References

Alpaydin E (2020) Introduction to machine learning. The MIT Press, Cambridge MA

Babuta A, Oswald M, Janjeva A (2020) Artificial intelligence and UK national security: policy considerations. RUSI Occasional Paper. Royal United Services Institute for Defence and Security Studies, London. https://rusi.org/sites/default/files/ai_national_security_final_web_version.pdf. Accessed 21 Sept 2020

Beavers AF (2002) Phenomenology and artificial intelligence. Metaphilosophy 33:70. https://doi.org/10.1111/1467-9973.00217

Bishop CM (2006) Pattern recognition and machine learning. Springer Science+Business Media, New York

Brooks RA (2002) Flesh and machines: how robots will change us. Pantheon Books, New York

Coeckelbergh M (2019) Technology, narrative and performance in the social theatre. In: Kreps D (ed) Understanding digital events: Bergson, Whitehead, and the experience of the digital, 1st edn. Routledge, New York, pp 13–27

Collins HM (1990) Artificial experts: social knowledge and intelligent systems. MIT Press, Cambridge MA

Dreyfus HL (1972) What computers can't do: a critique of artificial reason. Harper & Row, New York

Dreyfus HL (1992) What computers still can't do: a critique of artificial reason, revised edn. MIT Press, Cambridge MA

Elsevier (2018) ArtificiaI intelligence: how knowledge is created, transferred, and used. Trends in China, Europe, and the United States. Elsevier, Amsterdam. https://www.elsevier.com/__data/assets/pdf_file/0011/906779/ACAD-RL-AS-RE-ai-report-WEB.pdf. Accessed 22 Sept 2020

European Commission (2020a) White paper on artificial intelligence: a European approach to excellence and trust. European Commission, Brussels. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Accessed 22 Sept 2020

European Commission (2020b) Report on the safety and liability implications of artificial intelligence, the internet of things and robotics. European Commission, Brussels. https://ec.europa.eu/info/files/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics_en. Accessed 22 Sept 2020

Floridi L (1999) Information ethics: on the philosophical foundation of computer ethics. Ethics Inf Technol 1:33–52

Gasser U, Almeida VAF (2017) A layered model for AI governance. IEEE Internet Comput 21:58–62. https://doi.org/10.1109/MIC.2017.4180835

Glaser JI, Benjamin AS, Farhoodi R, Kording KP (2019) The roles of supervised machine learning in systems neuroscience. Prog Neurobiol 175:126–137. https://doi.org/10.1016/j.pneurobio.2019.01.008

Goffman E (1990) The presentation of self in everyday life, New Ed edn. Penguin, London

Hassabis D, Kumaran D, Summerfield C, Botvinick M (2017) Neuroscience-inspired artificial intelligence. Neuron 95:245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Heidegger M (1993) Sein und Zeit, 14th edn. Max Niemeyer Verlag GmbH & Co KG, Tübingen

Horvitz E (2017) AI, people, and society. Science 357:7. https://doi.org/10.1126/science.aao2466

Kaplan A, Haenlein M (2019) Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz 62:15–25

Khushf G (2004) Systems theory and the ethics of human enhancement: a framework for NBIC convergence. In: Roco MC, Montemagno CD (eds) Coevolution of human potential and converging technologies. New York Academy of Sciences, New York, pp 124–149

Korienek G, Uzgalis W (2002) Adaptable robots. Metaphilosophy 33:83–97

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. Behav Brain Sci 40:e253. https://doi.org/10.1017/S0140525X16001837

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444

Li D, Du Y (2007) Artificial intelligence with uncertainty. Chapman and Hall/CRC, Boca Raton FL

Marblestone AH, Wayne G, Kording KP (2016) Toward an integration of deep learning and neuroscience. Front Comput Neurosci 10:94. https://doi.org/10.3389/fncom.2016.00094

McCarthy J, Minsky ML, Rochester N, Shannon CE (2006) A proposal for the Dartmouth summer research project on artificial intelligence. AI Magazine 27:12–14. https://doi.org/10.1609/aimag.v27i4.1904

Moor JH, Bynum TW (2002) Introduction to cyberphilosophy. Metaphilosophy 33:4–10

Moran D (1999) Introduction to phenomenology, 1st edn. Routledge, London

Mumford E (2006) The story of socio-technical design: reflections on its successes, failures and potential. Inf Syst J 16:317–342. https://doi.org/10.1111/j.1365-2575.2006.00221.x

OECD (2019) Recommendation of the council on artificial intelligence. OECD/LEGAL/0449

Smith BC (2019) The promise of artificial intelligence: reckoning and judgment. The MIT Press, Cambridge MA

Stahl BC, Wright D (2018) Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Secur Priv 16:26–33. https://doi.org/10.1109/MSP.2018.2701164

Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460

Van Est R, Stemerding D, Rerimassie V, Schuijff M, Timmer J, Brom F (2014) From bio to NBIC: from medical practice to daily life. Rathenau Instituut, The Hague