# Philosophy of AI: A Structured Overview

*Vincent C. Müller*

FAU Erlangen-Nürnberg
www.sophia.de

Final Draft, 24th July, 2023

Forthcoming in
Smuha, Nathalie (ed.) (2024):
*Cambridge Handbook on the Law,*
*Ethics and Policy of Artificial Intelligence,*
Cambridge: Cambridge University Press.

*Abstract:* This paper presents the main topics, arguments, and positions in the philosophy of AI at present (excluding ethics). Apart from the basic concepts of *intelligence* and *computation*, the main topics of artificial cognition are *perception, action, meaning, rational choice, free will*, *consciousness,* and *normativity*. Through a better understanding of these topics, the philosophy of AI contributes to our understanding of the nature, prospects, and value of AI. Furthermore, these topics can be understood more deeply through the discussion of AI; so we suggest that "AI Philosophy" provides a new method for philosophy.

*Keywords:* action, AI philosophy, artificial intelligence, cognition, computation, consciousness, free will, goals, intelligence, meaning, philosophy of AI, rational choice

## 1. Topic and Method

### 1.1. Artificial Intelligence

The term *Artificial Intelligence* became popular after the 1956 "Dartmouth Summer Research Project on Artificial Intelligence" which stated its aims as follows:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.[1]

This is the ambitious research program that human intelligence or cognition can be understood or modeled as rule-based computation over symbolic representation, so these models can be tested by running them on different (artificial) computational hardware. If successful, the computers running those models would display artificial intelligence. AI and cognitive science are two sides of the same coin. This program is usually called *Classical AI*:[2]

a) AI is a research program to create computer-based agents that have intelligence.

The terms *Strong AI* and *Weak AI* as introduced by John Searle stand in the same tradition. *Strong* AI refers to the idea that: "the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states." *Weak* AI is means that AI merely simulates mental states. In this weak sense "the principal value of the computer in the study of the mind is that it gives us a very powerful tool."[3]

On the other hand, the term "AI" is often used in computer science in a sense that I would like to call *Technical AI*:

b) AI is a set of computer-science methods for perception, modelling, planning, and action (search, logic programming, probabilistic reasoning, expert systems, optimization, control engineering, neuromorphic engineering, machine learning, etc.).[4]

There is also a minority in AI that calls for the discipline to focus on the ambitions of a), while maintaining current methodology under b), usually under the name of *Artificial General Intelligence* (AGI).[5]

This existence of the two traditions (classical & technical) occasionally leads to suggestions that we should not use the term "AI", because it implies strong claims

---

[1] (McCarthy et al. 1955: 1).
[2] As a sample: (Dietrich 2002). The classic historical survey is (Margaret A. Boden 2006).
[3] (Searle 1980: 353).
[4] (Görz et al. 2020; Pearl and Mackenzie 2018; Russell 2019; Russell and Norvig 2020).
[5] AGI conferences have been organized since 2008.

that stem from the research program a) but have very little to do with the actual work under b). Perhaps we should rather talk about "machine learning" or "decision-support machines", or just "automation" (as the 1973 Lighthill Report suggested).[6] In the following we will clarify the notion of "intelligence" and it will emerge that there is a reasonably coherent research program of AI that unifies the two traditions: The *creation of intelligent behavior through computing machines*.

These two traditions now require a footnote: Both were largely developed under the notion of *classical AI*, so what has changed with the move to *machine learning* (ML)? ML is a traditional computational (connectivist) method in neural networks that does not use representations.[7] Since ca. 2015, with the advent of massive computing power and massive data for deep neural networks, the performance of ML systems in areas like translation, text production, speech recognition, games, visual recognition and autonomous driving has improved dramatically, so that it is superior to humans in some cases. ML is now the standard method in AI. What does this change mean for the future of the discipline? The honest answer is: We do not know yet. Just like any method, ML has its limits, but these limits are less restrictive than was thought for many years because the systems exhibit a non-linear improvement – with more data they may suddenly improve significantly. Its weaknesses (e.g. overfitting, causal reasoning, reliability, relevance, black box) may be quite close to those of human rational choice, especially if "predictive processing" is the correct theory of the human mind (sections 4.1, 6).

### 1.2. Philosophy of AI & Philosophy

One way to understand the philosophy of AI is that it mainly deals with three Kantian questions: What is AI? What can AI do? What should AI be? One major part of the philosophy of AI is the *ethics* of AI but we will not discuss this field here, because there is a separate entry on "Ethics of AI" in the present CUP handbook.[8]

Traditionally, the philosophy of AI deals with a few selected points where philosophers have found something to say about AI, e.g. about the thesis that

---

[6] (Lighthill 1973).
[7] (Rosenblatt 1957); (Buckner forthcoming; Garson and Buckner 2019; LeCun et al. 2015).
[8] See Chapter 4, written by Stefan Buijsman, Michael Klenk and Jeroen van den Hoven. See also: (Müller 2020, forthcoming).

cognition is computation, or that computers can have meaningful symbols.[9] Reviewing these points and the relevant authors (Turing, Wiener, Dreyfus, Dennett, Searle, …) would result in a fragmented discussion that never achieves a picture of the overall project. It would be like writing an old-style human history through a few 'heroes'. Also, in this perspective, the philosophy of AI is separated from its cousin, the philosophy of cognitive science, which in turn is closely connected to the philosophy of mind.[10]

In this paper we use a different approach: We look at *components of an intelligent system,* as they present themselves in philosophy, cognitive science and AI. One way to consider such components is that there are relatively simple animals that can do relatively simple things, and then we can move 'up' to more complicated animals that can do those simple things, and more. As a schematic example, a *fly* will continue to bump into the glass many times to get to the light; a *cobra* will understand that there is an obstacle here and try to avoid it; a *cat* might remember that there was an obstacle there the last time and take another path right away; a *chimpanzee* might realize that the glass can be broken with a stone; a *human* might find the key and unlock the glass door … or else take the window to get out. To engage in the philosophy of AI properly, we will thus need a wide range of philosophy: philosophy of mind, epistemology, language, value, culture, society, …

Furthermore, in our approach, the philosophy of AI is not just "applied philosophy": It is not that we have a solution ready in the philosopher's toolbox and "apply" it to solve issues in AI. The philosophical understanding itself *changes* when looking at the case of AI: It becomes less anthropocentric, less focused on our own human case. A deeper look at concepts must be normatively guided by the *function* these concepts serve, and that function can be understood better when we consider both the natural cases *and* the case of actual and possible AI. This paper is thus also a "proof of concept" for doing philosophy through the conceptual analysis of AI: I call this *AI philosophy*.[11]

I thus propose to turn the question from its head onto its feet, as Marx would have said: If we want to understand AI, we have to understand ourselves; and if we want to understand ourselves, we have to understand AI!

---

[9] There are very few surveys and no recent ones. See (Carter 2007; Copeland 1993; Dietrich 2002; Floridi 2003, 2011; Müller 2016). Some of what philosophers had to say can be seen as undermining the project of AI, cf. (Dietrich et al. 2021).

[10] (Margolis et al. 2012).

[11] A cousin is "technophilosophy" through the analysis of virtual worlds (Chalmers 2022).

## 2. Intelligence

### 2.1. The Turing Test

"I propose to consider the question 'Can Machines Think?'" Alan Turing wrote at the outset of his paper in the leading philosophical journal *Mind*.[12] This was 1950, Turing was one of the founding fathers of computers, and many readers of the paper would not even have heard of such machines, since there were only half a dozen universal computers in the world (Z3, Z4, ENIAC, SSEM, Harvard Mark III, Manchester Mark I)[13]. Turing moves swiftly to declare that searching for a definition of "thinking" would be futile and proposes to replace his initial question by the question whether a machine could successfully play an "imitation game". This game has come to be known as the "Turing Test": A human interrogator is connected to another human and a machine via "teleprinting", and if the interrogator cannot tell the machine from the human by holding a conversation, then we shall say the machine is "thinking". At the end of the paper he returns to the issue of whether machines can think and says: "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted".[14] So, Turing proposes to replace our everyday term of "thinking" by an operationally defined term, a term for which we can test with some procedure that has a measurable outcome.

Turing's proposal to replace the definition of thinking by an operational definition that relies exclusively on behavior fits with the intellectual climate of the time where behaviorism became a dominant force: In psychology, behaviorism is a *methodological* proposal that psychology should become a proper scientific discipline by relying on testable observation and experiment, rather than on subjective introspection. Given that the mind of others is a "black box", psychology should become the science of stimulus and behavioral response, of an input-output relation. Early analytic philosophy led to *reductionist behaviorism*; so, if the meaning of a term is its "verification conditions", then a mental term like "pain" just *means* the person is disposed to behaving a certain way.

Is the Turing test via observable behavior a useful definition of intelligence? Can it "replace" our talk of intelligence? It is clear that there will be intelligent beings

---

[12] (Turing 1950)
[13] (Anonymous 1950)
[14] (Turing 1950: 442).

that will not pass this test, for example humans or animals that cannot type. So, I think it is fair to say that Turing very likely only intended the passing of the test as being sufficient for having intelligence and not as necessary. So, if a system passes that test, does it have to be intelligent? This depends on whether you think intelligence is just intelligent behavior, or whether you think for the attribution of intelligence we also need to look at internal structure.

### 2.2. What Is Intelligence?

Intuitively, intelligence is an ability that underlies intelligent action. Which action is intelligent depends on the goals that are pursued, and on the success in achieving them – think of the animal cases mentioned above. Success will not only depend on the agent, but also on the conditions in which it operates, so a system with fewer options how to achieve a goal (e.g. find food) is less intelligent. In this vein, a classical definition is: "Intelligence measures an agent's ability to achieve goals in a wide range of environments."[15] Here intelligence is the *ability to flexibly pursue goals,* where flexibility is explained with the help of different environments. This notion of intelligence from AI is an *instrumental* and normative notion of intelligence, in the tradition of classical decision theory, which says that a rational agent should always try to maximize expected utility (see section 6).[16]

If AI philosophy understands intelligence as relative to an environment, then to achieve more intelligence, one can change the agent or change the environment. Humans have done both on a huge scale through what is known as "culture": Not only have we generated a sophisticated learning system for humans (to change the agent), we have also physically shaped the world such that we can pursue our goals in it; e.g. to travel, we have generated roads, cars with steering wheels, maps, road signs, digital route planning, and AI systems. We now do the same for AI systems; both the learning system, and the change of the environment (cars with computer interfaces, GPS, etc.). By changing the environment, we will also change our cognition and our lives – perhaps in ways that turn out to be to our detriment.

In sections 4-9, we will look at the main components of an intelligent system; but before that we discuss the mechanism used in AI: computation.

[15] (Legg and Hutter 2007: 402).

[16] (e.g. Simon 1955; Thoma 2019), see also the neo-behaviorist proposal in (Coelho Mollo 2022).

# 3. Computation

## 3.1. The Notion of Computation

The machines on which AI systems run are "computers", so it will be important for our task to find out what a computer is and what it can do, in principle. A related question is whether human intelligence is wholly or partially due to computation – if it is wholly due to computation, as classical AI had assumed, then it appears possible to re-create this computation on an artificial computing device.

In order to understand what a computer is, it is useful to remind ourselves of the history of computing machines – I say "machines" because before ca. 1945 the word "computer" was a term for a human who has a certain profession, for someone who does computations. These computations, e.g. the multiplication of two large numbers, are done through a mechanical step-by-step procedure that will lead to a result once carried out completely. Such procedures are called "algorithms". In 1936, in response to Gödel's challenge of the "Entscheidungsproblem", Alan Turing suggested that the notion of "computing something" could be explained by "what a certain type of machine can do" (just like he proposed to operationalize the notion of intelligence in the "Turing Test"). Turing sketched what such a machine would look like, with an infinitely long tape for memory, a head that can read from and write symbols to that tape. These states on the tape are always specific discrete states, such that each state is of a type from a finite list (symbols, numbers, …), so for example it either is the letter "V" or the letter "C", not a bit of each. In other words, the machine is "digital" (not analog).[17] Then there is one crucial addition: In the "universal" version of the machine, one can *change* what the computer does through further input. In other words, the machine is *programable* to perform a certain algorithm, and it stores that program in its memory.[18] Such a computer is a universal computer, i.e. it can compute any algorithm. It should be mentioned that wider notions of computation have been suggested, e.g. analog computing and hypercomputing.[19]

There is also the question whether computation is a real property of physical systems, or whether it is rather a useful way of describing these. Searle has said: "The electrical state transitions are intrinsic to the machine, but the computation

[17] (Negroponte 1995), see also (Haugeland 1985: 57); (Müller 2013).
[18] (Gödel 1931; Turing 1936). The original programme outlined in (Hilbert 1900). See e.g. (Copeland et al. 2013).
[19] (Piccinini 2021; Shagrir 2022; Siegelmann 1995, 1997).

is in the eye of the beholder."[20] If we take an anti-realist account of computation, then the situation changes radically.

The exact same computation can be performed on different physical computers, and it can have a different semantics. There are thus three levels of description that are particularly relevant for a given computer: (a) The *physical level* of the actual "realization" of the computer, (b) the *syntactic level* of the algorithm computed and (c) the *symbolic level* of content, of what is computed.

Physically, a computing machine can be built out of anything and use any kind of property of the physical world (cogs and wheels, relays, DNA, quantum states, etc.). This can be seen as using a physical system to encode a formal system.[21] Actually, all universal computers have been made with large sets of switches. A switch has two states (open/closed), so the resulting computing machines work on two states (on/off, 0/1), they are *binary* – this is a design decision. Binary switches can easily be combined to form "logic gates" that operate on input in the form of the logical connectives in Boolean logic (which is also two-valued): NOT, AND, OR, etc. If such switches are in a state that can be *syntactically* understood as 1010110, then *semantically,* this could (on current ASCII/ANSI conventions) represent the letter "V", the number "86", a shade of light grey, a shade of green, etc. etc.

### 3.2. Computationalism

As we have seen, the notion that computation is the cause of intelligence in natural systems, e.g. humans, and can be used to model and reproduce this intelligence is a basic assumption of classical AI. This view is often coupled with (and motivated by) the view that human mental states are functional states and that these functional states are that of a computer: "machine functionalism". This thesis is often assumed as a matter of course in the cognitive sciences and neuroscience, but it is also the subject of significant criticism in recent decades.[22] The main sources for this view are an enthusiasm for the universal technology of digital computation, and early neuroscientific evidence indicated that human neurons (in the brain and body) are also somewhat binary, i.e. they either send a signal to other neurons, they "fire", or they don't. Some authors defend the *Physical Symbol*

[20] (Dodig-Crnkovic and Müller 2011; Searle 2004: 64).
[21] (Horsman et al. 2014).
[22] (Edelman 2008; Miłkowski 2018), for the discussion (Harnad 1990; Scheutz 2002; Shagrir 1997; Varela et al. 1991)

*System Hypothesis,* which is computationalism, plus the contention that only computers can be intelligent.[23]

## 4. Perception & Action

### 4.1. Passive Perception

You may be surprised to find that the heading of this chapter combines perception and action in one. We can learn from AI and cognitive science that the main function of perception is to allow action; indeed that perception *is* a kind of action. The traditional understanding of perception in philosophy is *passive* perception, watching ourselves watching the world in what Dan Dennett has called the *Cartesian Theatre*: It is as though I had a little human sitting inside my head, listening to the outside world through our ears, and watching the outside world through our eyes.[24] That notion is absurd, particularly because it would require there to be yet another little human sitting in the head of that little human. And yet, a good deal of the discussion of human perception in the philosophical literature really does treat perception as though it were something that happens to me when inside.

For example, there is the 2D-3D problem in vision, the problem of how I can generate the visual experience of a 3-dimensional world through a 2-dimensional sensing system (the retina, a 2-dimensional sheet that covers our eyeballs from the inside). There must be a way of processing the visual information in the retina, the optical nerve and the optical processing centers of the brain that generates this three-dimensional experience. Not really.[25]

### 4.1. Active Perception

The 3-dimensional impression is generated by an interaction between me and the world (in the case of vision it involves movement of my eyes and my body). It is better to think of perception along with the lines of the sense of touch: Touching is something that I *do,* so that I can find out the softness of an object, the texture of its surface, its temperature, its weight, its flexibility, etc. I do this by acting and then perceiving the change of sensory input. This is called a perception-action-loop: I do something, that changes the world, and that changes the perception that I have.

---

[23] (cf. Margaret A. Boden 2006: 1419ff; Newell and Simon 1976: 116).
[24] (Dennett 1991: 107).
[25] For an introduction to vision, see (O'Regan 2011, ch. 1-5).

It will be useful to stress that this occurs with perception of my own body as well. I only know that I have a hand because my visual sensation of the hand, the proprioception, and the sense of touch are in agreement. When that is not the case it is fairly easy to make me feel that a rubber hand is my own hand - this is known as the "rubber hand illusion". Also, if a prosthetic hand is suitably connected to the nervous system of a human, then the perception-action-loop can be closed again, and the human will feel this as their own hand.

### 4.1. Predictive Processing and Embodiment

This view of perception has recently led to a theory of the "predictive brain": What the brain does is not to passively wait for input, but it is *always on* to actively participate in the action-perception-loop. It generates *predictions* what the sensory input will be, given my actions, and then it matches the predictions with the actual sensory input. The difference between the two is something that we try to minimize, which is called the "free energy principle".[26]

In this tradition, the perception of a natural agent or AI system is something that is intimately connected to the physical interaction of the body of the agent with the environment; perception is thus a component of embodied cognition. A useful slogan in this context is "4E cognition", which says that cognition is *embodied*; it is *embedded* in an environment with other agents; it is *enactive* rather than passive; and it is *extended,* i.e. not just inside the head.[27] One aspect that is closely connected to 4E cognition is the question whether cognition in humans is fundamentally representational, and whether cognition in AI has to be representational (see section 5).

Embodied cognition is sometimes presented as an empirical thesis about actual cognition (especially in humans) or as a thesis on the suitable design of AI systems, and sometimes as an analysis of what cognition is and has to be. In the latter understanding, non-embodied AI would necessarily miss certain features of cognition.[28]

---

[26] (Clark 2013, 2016, 2023; Friston 2010).
[27] (Clark and Chalmers 1998; Clark 2003; Newen et al. 2018).
[28] (Dreyfus 1972; Pfeifer and Bongard 2007).

# 5. Meaning & Representation

## 5.1. The Chinese Room Argument

As we saw above, classical AI was founded on the assumption that the appropriately programmed computer really *is* a mind – this is what John Searle called *strong AI*. In his famous paper “Minds, Brains and Programs”, Searle presented a thought experiment of the “Chinese Room”.[29] The Chinese Room is a computer, constructed as follows: There is a closed room in which John Searle sits and has a large book which provides him with a computer program, with algorithms, on how to process the input and provide output. Unknown to him, the input that he gets is Chinese writing, and the output that he provides are sensible answers or comments about that linguistic input. This output, so the assumption, is indistinguishable from the output of a competent Chinese speaker. And yet Searle in the room understands no Chinese and will learn no Chinese from the input that he gets. Therefore, Searle concludes, *computation is not sufficient for understanding*. There can be no strong AI.

In the course of his discussion of the Chinese room argument, Searle looks at several replies: The *systems reply* accepts that Searle has shown that no amount of simple manipulation of the person in the room will enable that person to understand Chinese, but objects that perhaps symbol manipulation will enable *the wider system,* of which the person is a component, to understand Chinese. So perhaps there is a part-whole fallacy here? This reply raises the question, why one might think that the whole system has properties which the algorithmic processor does not have.

One way to answer this challenge, and change the system, is the *robot reply*, which grants that the whole system, as described, will not understand Chinese because it is missing something that Chinese speakers have, namely a causal connection between the words and the world. So, we would need to add sensors and actuators to this computer, that would take care of the necessary causal connection. Searle responds to this suggestion by saying input from sensors would be “just more Chinese” to Searle in the room; It would not provide any further understanding, in fact Searle would have no idea that the input is from a sensor.[30]

---

[29] (Searle 1980).

[30] (Cole 2020; Preston and Bishop 2002).

### 5.2. Reconstruction

I think it is best to view the core of the Chinese room argument as an extension of Searle's remark:

> No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle.[31]

Accordingly, the argument that remains can be reconstructed as:

1. If a system does only syntactical manipulation, it will not acquire meaning.
2. A computer does only syntactical manipulation.

→

3. A computer will not acquire meaning.

In Searle's terminology, a computer has *only syntax* and *no semantics*; the symbols in a computer lack the intentionality (directedness) that human language use has. He summarizes his position at the end of the paper:

> 'Could a machine think?' The answer is, obviously, yes. We are precisely such machines. […] But could something think, understand, and so on solely in virtue of being a computer with the right sort of program? […] the answer is no.[32]

### 5.3. Computing, Syntax and Causal Powers

Reconstructing the argument in this way, the question is whether the premises are true. Several people have argued that premise 2. is false, because one can only understand what a computer does as responding to the program as meaningful.[33] I happen to think that this is a mistake, the computer does not *follow* these rules, it is just constructed in such a way that it *acts according to* these rules, if its states are suitably interpreted by an observer.[34] Having said that, any actual computer, any physical realization of an abstract algorithm processor, *does* have causal powers, it does more than syntactic manipulation. For example, it may be able to turn the lights on or off.

---

31 (Searle 1980: 424).
32 (Searle 1980: 422).
33 (McCarthy 2007);(Margaret A. Boden 1988: 97; Haugeland 2002: 385).
34 (Wittgenstein 1953: §82-86, 198, 217, 34` etc.).

The Chinese room argument has moved the attention in the philosophy of language away from convention and logic towards the conditions for a speaker to mean what they say (speakers' meaning), or to mean anything at all (have intentionality); in particular, it left us with the discussion on the role of *representation* in cognition, and the role of computation over representations.[35]

## 6. Rational Choice

### 6.1. Normative Decision Theory: MEU

A rational agent will perceive the environment, find out which options for action exist, and then take the best decision. This is what decision theory is about. It is a normative theory on how a rational agent *should* act, given the knowledge they have – not a descriptive theory of how rational agents *will* actually act.

So how should a rational agent decide which is the best possible action? They evaluate the possible outcomes of each choice and then select the one that is best, meaning the one that has the highest subjective utility, i.e. utility as seen by the particular agent. It should be noted that rational choice in this sense is not necessarily egoistic, it could well be that the agent puts a high utility on the happiness of someone else, and thus rationally chooses a course of action that maximizes overall utility through the happiness of that other person. In actual situations, the agent typically does not know what the outcomes of particular choices will be, so they act under uncertainty. To overcome this problem, the rational agent selects the action with *maximum expected utility* (MEU), where the value of a choice equals the utility of the outcome multiplied by the probability of that outcome occurring. This thought can be explained with the expected utility of certain gambles or lotteries. In more complicated decision cases the rationality of a certain choice depends on subsequent choices of *other agents*. These kinds of cases are often described with the help of "games" played with other agents. In such games it is often a successful strategy to cooperate with other agents in order to maximize subjective utility.

In artificial intelligence it is common to perceive of AI agents as rational agents in the sense described. For example, Stuart Russell says: "In short, a rational agent acts so as to maximise expected utility. It's hard to over-state the importance of this conclusion. In many ways, artificial intelligence has been mainly about working out the details of how to build rational machines."[36]

---

[35] (Searle 1984, 2004).
[36] (Russell 2019: 23).

### 6.2. Resources and Rational Agency

It is not the case that a rational agent *will* always choose the perfect option. The main reason is that such an agent must deal with the fact that their resources are limited, in particular data storage and time (most choices are time-critical). The question is thus not only what the best choice is, but how many resources I should spend on optimizing my choice; when should I stop optimizing and start acting? This phenomenon is called *bounded rationality*, *bounded optimality* and in cognitive science it calls for *resource rational* analysis.[37] Furthermore, there is no set of discrete options from which to choose, and a rational agent needs to reflect on the goals to pursue (see section 9).

The point that agents (natural or artificial) will have to deal with limited resources when making choices, has tremendous importance for the understanding of cognition. It is often not fully appreciated in philosophy – even the literature about the limits of rational choice seems to think that there is something 'wrong' with using heuristics that are biased, being 'nudged' by the environment, or using the environment for 'extended' or 'situated' cognition.[38] But it would be irrational to aim for perfect cognitive procedures, not to mention for cognitive procedures that would not be influenced by the environment.

### 6.3. The Frame Problem(s)

The original frame problem for classical AI was how to *update a belief system* after an action, without stating all the things that have *not* changed; this requires a logic where conclusions can change if a premise is added – a non-monotonic logic.[39] Beyond this more technical problem, there is a philosophical problem of updating beliefs after action, popularized by Dennett, which asks how to find out what is relevant, how wide the frame should be cast for *relevance*. As Shanahan says "relevance is holistic, open-ended, and context-sensitive" but logical inference is not.[40]

There is a very general version of the frame problem, expressed by Jerry Fodor, who says, the frame problem really is: "Hamlet's problem: when to stop thinking". He continues by saying that "modular cognitive processing is *ipso facto* irrational

---

[37] (Lieder and Griffiths 2020; Russell 2016: 16ff; Simon 1955: 99; Wheeler 2020)
[38] (Kahneman and Tversky 1979; Kahnemann 2011; Thaler and Sunstein 2008) vs. (Kirsh 2009).
[39] (Shanahan 2016).
[40] (Dennett 1984b; Shanahan 2016).

[…] by attending to less than all the evidence that is relevant and available”.[41] Fodor sets the challenge that in order to perform a logical inference, especially an abduction, one needs to have decided what is relevant. However, he seems to underestimate that one cannot attend to *all* that is relevant and available (rationality is bounded). It is currently unclear whether the frame problem can be formulated without dubious assumptions about rationality. Similar concerns apply to the claims that Gödel has shown deep limitations of AI systems.[42] Overall, there may be more to intelligence than instrumental rationality.

### 6.4. Creativity

Choices that involve *creativity* are often invoked as something special, not merely mechanical, and thus inaccessible to a mere machine. The notion of “creation” has significant impact in our societal practice particularly when that creation is protected by intellectual property rights – and AI systems *have* created or co-created music, painting and text. It is not clear that there is a notion of creativity which would provide an argument against machine creativity. Such a notion would have to combine two aspects that seem to be in tension: On the one hand creativity seems to imply causation that includes acquiring knowledge and techniques (think of J.S. Bach composing a new cantata), on the other hand creativity is supposed to be a non-caused, non-predictable, spark of insight. It appears unclear whether such a notion of creativity can, or indeed should, be formulated.[43] Perhaps a plausible account is that creativity involves moving between different spaces of relevance, as in the frame problem.

## 7. Free Will and Creativity

### 7.1. Determinism, Compatibilism

The problem that usually goes under the heading of “free will” is how physical beings like humans or AI systems can have something like free will. The traditional division for possible positions in the space of free will can be put in terms of a decision tree. The first choice is whether *determinism* is true, i.e. the thesis that all events are caused. The second choice is whether *incompatibilism* is true, i.e. the thesis that if determinism is true, then there is no free will.

[41] (Fodor 1987: 140f; Sperber and Wilson 1996).
[42] (Koellner 2018b, 2018a; Lucas 1996).
[43] (Margaret A Boden 2014; Colton and Wiggins 2012; Halina 2021).

The position known as *hard determinism* says that determinism is indeed true, and if determinism is true then there is no such thing as free will – this is the conclusion that most of its opponents try to avoid. The position known as *libertarianism* (not the political view) agrees that incompatibilism is true, but adds that determinism is not, so we are free. The position known as *compatibilism* says that determinism and free will are compatible and thus it may well be that determinism is true *and* humans have free will (and it usually adds that this is actually the case).

This results in a little matrix of positions:

| | **Incompatibilism** | **Compatibilism** |
|---|---|---|
| **Determinism** | Hard Determinism | Optimistic/Pessimistic Compatibilism |
| **Non-Determinism** | Libertarianism | [not a popular option] |

### 7.1. Compatibilism and Responsibility in AI

In a first approximation, when I say I did something freely, it means that it was *up to me* that I was *in control*. That notion of control can be cashed out by saying I could have done otherwise than I did, specifically I could have done otherwise if I had *decided* otherwise. To this we could add that I would have decided otherwise if I had had other *preferences* or *knowledge* (e.g. I would not have eaten those meatballs if I had a preference against eating pork, and if I had known that they contain pork). Such a notion of freedom thus involves an *epistemic condition* and a *control condition*.

So, I act freely if I do as I choose according to the preferences that I have (my subjective utility). But why do I have these preferences? As Aristotle already knew, they are not under my voluntary control, I could not just *decide* to have other preferences and then have them. However, as Harry Frankfurt has pointed out, I can have *second order* preferences or desires, i.e. I can prefer to have other preferences than the ones I actually have (I could want not to have a preference for those meatballs, for example). The notion that I can overrule my preferences with rational thought is what Frankfurt calls the *will,* and it is his condition for being a person. In a first approximation one can thus say, *to act freely is to act as I choose, to choose as I will, and to will as I rationally decide to prefer*.[44]

[44] (Dennett 1984a; Frankfurt 1971).

The upshot of this debate is that the function of a notion of free will for agency in AI or humans is to allow personal *responsibility*, not to determine *causation*. The real question is: What are the conditions such that an agent is *responsible* for their actions and *deserves* being praised or blamed for them. This is independent of the freedom from causal determination; that kind of freedom we do not get, and we do not need.[45]

There is a further debate between "optimists" and "pessimists" whether humans actually do fulfil those conditions (in particular whether they can truly cause their preferences) and can thus properly be said to be responsible for their actions and *deserve* praise or blame – and accordingly whether reward or punishment should have mainly forward-looking reasons.[46] In the AI case, an absence of responsibility has relevance for their status as moral agents, for the existence of "responsibility gaps", and for what kinds of decisions we should leave to systems that cannot be held responsible.[47]

## 8. Consciousness

### 8.1. Awareness & Phenomenal Consciousness

In a first approximation, it is useful to distinguish two types of consciousness: *Awareness* and *phenomenal consciousness*. Awareness is the notion that a system has cognitive states on a base level (e.g. it senses heat) and on a meta level, it has states where it is aware of the states on the object level. This awareness, or access, involves the ability to remember and use the cognitive states on the base level. This is the notion of "conscious" that is opposed to "unconscious" or "subconscious" – and it appears feasible for a multi-layered AI system.

Awareness is often, but not necessarily, connected to a specific way that the cognitive state at the base level *feels* to the subject – this is what philosophers call *phenomenal consciousness,* or how things *seem* to me *(*Greek *phaínetai).* This notion of consciousness is probably best explained with the help of two classical philosophical thought experiments: the bat, and the color scientist.

If you and I go out to have the same ice cream, then I can still not know what the ice cream tastes like to you, and I would not know that even if I knew everything

45 Chapter 6 of this book by Lode Lauwaert and Ann-Katrien Oimann delves further into the subject of AI and responsibility.

46 (Dennett and Caruso 2018; Mele 2006; Pink 2004; Strawson 2004).

47 (Müller 2021; Simpson and Müller 2016; Sparrow 2007).

about the ice cream, you, your brain and your taste buds. Somehow, *what it is like* for you is something epistemically inaccessible to me, I can never know it, even if I know everything about the physical world. In the same way, I can never know what it is like to be a bat.[48]

A similar point about what we cannot know in principle, is made by Frank Jackson in the article "What Mary didn't know".[49] In his thought experiment, Mary is supposed to be a person who has never seen anything with color in her life, and yet she is a perfect color scientist, she knows everything there is to know about color. One day, she gets out of her black and white environment and sees color for the first time. It appears that she learns something new at that point.

The argument that is suggested here seems to favor an argument for a mental-physical *dualism* of *substances* or at least *properties*: I can know all the physics, and I cannot know all the phenomenal experience, therefore phenomenal experience is not part of physics. If dualism is true, then it may appear that we cannot hope to generate phenomenal consciousness with the right physical technology, such as AI. In the form of *substance dualism*, as Descartes and much of religious thought had assumed, dualism is now unpopular since most philosophers assume physicalism, that "everything is physical".

Various arguments against the reduction of mental to physical *properties* have been brought out, so it is probably fair to say that *property dualism* has a substantial following. This is often combined with substance monism in some version of "supervenience of the mental on the physical", i.e. the thesis that two entities with the same physical properties must have the same mental properties. Some philosophers have challenged this relation between property dualism and the possibility of artificial consciousness. David Chalmers has argued that "the physical structure of the world—the exact distribution of particles, fields, and forces in spacetime—is logically consistent with the absence of consciousness, so the presence of consciousness is a further fact about our world". Despite this remark, he supports computationalism: "… strong artificial intelligence is true: there is a class of programs such that any implementation of a program in that class is conscious."[50]

---

[48] (Nagel 1974; 1987, ch. 3)
[49] (Jackson 1986).
[50] (Chalmers and Searle 1997; Chalmers 1999: 436; Davidson 1970). Donald Davidson, 'Mental Events' in L. Foster and J. Swanson (eds), *Experience and Theory* (University of Massachusetts Press, Amherst, Mass. 1970)

What matters for the function of consciousness in AI or natural agents is not the discussion about dualisms, but rather why phenomenal consciousness in humans is the way it is, how one could tell whether a system is conscious, and whether there could be a human who is physically just like me, but without consciousness (a "philosophical zombie").[51]

### 8.2. The Self

Personal identity in humans is mainly relevant because it is a condition for allocating responsibility (see section 6.4): In order to allocate blame or praise there has to be a sense in which I am *the same person* as the one who performed the action in question. We have a sense that there is a life in the past that is mine, and only mine – how this is possible is known as the "persistence question". The standard criteria for me being the same person as that little boy in the photograph are my *memory* of being that boy, and the *continuity of my body* over time. Humans tend to think that *memory* or *conscious experience,* or *mental content* are the criteria for personal identity, which is why we think we can imagine surviving our death, or living in a different body.[52]

So, what is a "part" of that persistent self? Philosophical phantasies and neurological rarities[53] aside, in the human case there is now no doubt what is "part of me" and what is not – I continuously work on maintaining that personal identity by checking that the various senses are in agreement, e.g. I try to reach for the door handle, I see my hand touching the handle, I can feel it … and then I can see the door opening and feel my hand going forward. This is very different from a computer: The components of the standard Von Neumann architecture (input-system, storage, random-access memory, processor, output-system) can be in the same box or miles apart, they can even be split into more components (e.g. some off-board processing of intensive tasks) or stored in spaces like the 'cloud' that are not defined through physical location. And that is only the hardware, the software faces similar issues, so a persistent and delineated self is not an easy task for an AI system. It is not clear that there is a function for a self in AI, which would have repercussions for attributing moral agency and even patiency.

---

[51] (O'Regan 2011).

[52] (Metzinger 2009; Olsen 2019).

[53] E.g. "The Man Who Fell out of Bed" in (Sacks 1985) or the view of humans as superorganisms, based on the human microbiome.

## 9. Normativity

Let us return briefly to the issues of rational choice and responsibility. Stuart Russell said that "AI has adopted the standard model: we build optimising machines, we feed objectives into them, and off they go."[54] On that understanding, AI is a tool, and we need to provide the objectives or goals for it. AI has only *instrumental intelligence* on how to reach given goals. However, *general intelligence* also involves a metacognitive reflection on which goals are relevant to my action now (food or shelter?) and a reflection on which goals one should pursue.[55] One of the open questions is whether a non-living system can have "real goals" in the sense required for choice and responsibility, e.g. of goals that have subjective value to the system, and that the system recognizes as important after reflection. Without such reflection on goals, AI systems would not be moral agents and there could be no "machine ethics" that deserves the name. Similar considerations apply to other forms of normative reflection, e.g. in aesthetics and politics. This discussion in AI philosophy seems to show that there is a function for normative reflection in humans or AI as an elementary part of the cognitive system.

## 10. References


Anonymous (1950), 'Digital Computing Newsletter', *Office of Naval Research, Mathematical Sciences Division, Washington DC,* 2 (1), 1-4.

Boden, Margaret A (2014), 'Creativity and artificial intelligence: A contradiction in terms?', in Elliot Samuel Paul and Scott Barry Kaufman (eds.), *The philosophy of creativity: New essays* (Oxford University Press).

Boden, Margaret A. (1988), *Computer models of the mind: Computational approaches in theoretical psychology* (Cambridge: Cambridge University Press).

— (2006), *Mind as machine: A history of cognitive science*, 2 vols. (Oxford: Oxford University Press).

Buckner, Cameron (forthcoming), *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence* (New York: Oxford University Press).

Carter, Matt (2007), *Minds and computers: An introduction to the philosophy of artificial intelligence* (Edinburgh: Edinburgh University Press).

Chalmers, David J (1999), 'Précis of "The Conscious Mind"', *Philosophy and Phenomenological Research,* LIX (2), 435-38.

— (2022), *Reality+: Virtual worlds and the problems of philosophy* (New York: W. W. Norton).

Chalmers, David J and Searle, John (1997), 'Consciousness and the philosophers: An exchange', *New York Review of Books,* May 15.


[54] (Russell 2019: 172).

[55] (Müller and Cannon 2022).

<https://www.nybooks.com/articles/1997/05/15/consciousness-and-the-philosophers-an-exchange/>.

Clark, Andy (2003), *Natural born cyborgs: Minds, technologies, and the future of human intelligence* (Oxford: Oxford University Press).

— (2013), 'Whatever next? Predictive brains, situated agents, and the future of cognitive science.', *Behavioral and Brain Sciences,* 36 (6), 181-204.

— (2016), *Surfing uncertainty: Prediction, action, and the embodied mind* (Oxford University Press).

— (2023), *The experience machine: How our minds predict and shape reality* (New York: Pantheon Books).

Clark, Andy and Chalmers, David J. (1998), 'The extended mind', *Analysis,* 58 (1), 7-19.

Coelho Mollo, Dimitri (2022), 'Intelligent Behaviour', *Erkenntnis,* online first.

Cole, David (2020), 'The Chinese room argument', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 edn.; Stanford, Cal.: CSLI).

Colton, Simon and Wiggins, Geraint A (2012), 'Computational creativity: The final frontier?', *ECAI* (12: Montpelier), 21-26.

Copeland, Jack B. (1993), *Artificial intelligence: A philosophical introduction* (Oxford: Blackwell).

Copeland, Jack B.; Posy, Carl J and Shagrir, Oron (2013) *Computability: Turing, Gödel, Church, and Beyond* [online text], MIT Press

Davidson, Donald (1970), 'Mental Events', in L. Foster and J. Swanson (eds.), *Experience and Theory* (University of Massachusetts Press, Amherst, Mass.).

Dennett, Daniel C (1984a), *Elbow room: The varieties of free will worth wanting* (Cambridge, Mass. edn.: MIT Press).

— (1984b), 'Cognitive wheels: The frame problem of AI', in Christopher Hookway (ed.), *Minds, machines, and evolution: Philosophical studies* (Cambridge: Cambridge University Press), 129-52.

— (1991), *Consciousness explained* (New York: Little, Brown & Co.).

Dennett, Daniel C and Caruso, Gregg D (2018), 'Just deserts', *Aeon,* (04.10.2018).

Dietrich, Eric (2002), 'Philosophy of artificial intelligence', *The Encyclopedia of Cognitive Science*, 203-08.

Dietrich, Eric; Fields, Chris; Sullins, John P. ; Van Heuveln, Bram and Zebrowski, Robin (2021), *Great philosophical objections to artificial intelligence: The history and legacy of the AI wars* (London: Bloomsbury Academic).

Dodig-Crnkovic, Gordana and Müller, Vincent C. (2011), 'A dialogue concerning two world systems: Info-computational vs. mechanistic', in Gordana Dodig-Crnkovic and Mark Burgin (eds.), *Information and computation: Essays on scientific and philosophical understanding of foundations of information and computation* (Boston: World Scientific), 149-84.

Dreyfus, Hubert L. (1972), *What computers still can't do: A critique of artificial reason* (2 edn.; Cambridge, Mass.: MIT Press 1992).

Edelman, Shimon (2008), *Computing the mind: How the mind really works* (Oxford University Press).

Floridi, Luciano (ed.), (2003), *The Blackwell guide to the philosophy of computing and information* (Oxford: Blackwell)

— (2011), *The philosophy of information* (Oxford: Oxford University Press).

Fodor, Jerry A. (1987), 'Modules, frames, fridgeons, sleeping dogs, and the music of the spheres', in J. L. Garfield (ed.), *Modularity in knowledge representation and natural-language understanding.* (Cambridge, MA, US: The MIT Press), 25-36.

Frankfurt, Harry (1971), 'Freedom of the will and the concept of a person', *The Journal of Philosophy,* LXVIII (1), 5-20.

Friston, Karl J. (2010), 'The free-energy principle: A unified brain theory?', *Nature Reviews Neuroscience* 11, 127–38.
Garson, James and Buckner, Cameron (2019), 'Connectionism', in Edward N Zalta (ed.), *Stanford Encyclopedia of Philosophy* (CSLI, Stanford).
Gödel, Kurt (1931), 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I', *Monatshefte für Mathematik und Physik,* 38, 173-98.
Görz, Günther; Schmid, Ute and Braun, Tanya (2020), *Handbuch der künstlichen Intelligenz* (5th edn.; Berlin: De Gruyter).
Halina, Martha (2021), 'Insightful artificial intelligence', *Mind and Language,* 36 (2), 315-29.
Harnad, Stevan (1990), 'The symbol grounding problem', *Physica D,* 42, 335-46.
Haugeland, John (1985), *Artificial intelligence: The very idea* (Cambridge, Mass.: MIT Press).
— (2002), 'Syntax, semantics, physics', in John Preston and Mark Bishop (eds.), *Views into the Chinese room: New essays on Searle and artificial intelligence* (Oxford: Oxford University Press), 379-92.
Hilbert, David (1900), 'Mathematische Probleme', *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Math.-Phys. Klasse*, 253-97.
Horsman, Clare; Stepney, Susan; Wagner, Rob C. and Kendon, Viv (2014), 'When does a physical system compute?', *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science,* 470 (2169), 1-25.
Jackson, Frank (1986), 'What Mary didn't know', *Journal of Philosophy,* 83, 291-95.
Kahneman, Daniel and Tversky, Amos (1979), 'Prospect theory: An analysis of decision under risk', *Econometrica,* 47, 263–91.
Kahnemann, Daniel (2011), *Thinking fast and slow* (London: Macmillan).
Kirsh, David (2009), 'Problem solving and situated cognition', in P Robbins and M Aydede (eds.), *The Cambridge handbook of situated cognition* (Cambridge: Cambridge University Press), 264–306.
Koellner, Peter (2018a), 'On the question of whether the mind can be mechanized, II: Penrose's new argument', *Journal of Philosophy,* 115 (9), 453-84.
— (2018b), 'On the Question of Whether the Mind Can Be Mechanized, I: From Gödel to Penrose', *Journal of Philosophy,* 115 (7), 337-60.
LeCun, Yan; Bengio, Yoshua and Hinton, Jeffrey (2015), 'Deep learning', *Nature,* 521 (7553), 436-44.
Legg, Shane and Hutter, Marcus (2007), 'Universal intelligence: A Definition of machine intelligence', *Minds and Machines,* 17 (4), 391-444.
Lieder, Falk and Griffiths, Thomas L. (2020), 'Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources', *Behavioral and Brain Sciences,* 43, e1.
Lighthill, James (1973), 'Artificial intelligence: A general survey', *Artificial intelligence: A paper symposion* (London: Science Research Council).
Lucas, J. R. (1996), 'Minds, machines and Gödel: A retrospect', in Peter J. R. Millican and Andy Clark (eds.), *Machines and Thought* (Oxford: Oxford University Press).
Margolis, Eric; Samuels, Richard and Stich, Stephen (eds.) (2012), *The Oxford handbook of philosophy of cognitive science* (Oxford: Oxford University Press)
McCarthy, John (2007), 'John Searle's Chinese room argument', (updated 10.06.2007) <http://www-formal.stanford.edu/jmc/chinese.html>, accessed 06.10.2007.
McCarthy, John; Minsky, Marvin; Rochester, Nathaniel and Shannon, Claude E. (1955), 'A proposal for the Dartmouth summer research project on artificial intelligence', <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>, accessed October 2006.
Mele, Alfred R. (2006), *Free will and luck* (Oxford: Oxford University Press).

Metzinger, Thomas (2009), *The ego tunnel: The science of the mind and the myth of the self* (New York: Basic Books).

Miłkowski, Marcin (2018), 'Objections to computationalism: A survey', *Roczniki Filozoficzne,* LXVI (8), 1-19.

Müller, Vincent C. (2013), 'What is a digital state?', in Mark J. Bishop and Yasemin J. Erden (eds.), *The Scandal of Computation - What is Computation? - AISB Convention 2013* (Hove: AISB), 11-16.

— (2016), 'New developments in the philosophy of AI', in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Synthese Library; Berlin: Springer), 1-6.

— (2020), 'Ethics of artificial intelligence and robotics', in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Summer 2020; Palo Alto: CSLI, Stanford University), 1-70.

— (2021), 'Is it time for robot rights? Moral status in artificial entities', *Ethics & Information Technology,* 23 (3), 579–87.

— (forthcoming), *Can machines think? Fundamental problems of artificial intelligence* (New York: Oxford University Press).

Müller, Vincent C. and Cannon, Michael (2022), 'Existential risk from AI and orthogonality: Can we have it both ways?', *Ratio,* 35 (1), 25-36.

Nagel, Thomas (1974), 'What is it like to be a bat?', *Philosophical Review,* 83 (4), 435-50.

— (1987), *What does it all mean? A very short introduction to philosophy* (Oxford & New York: Oxford University Press).

Negroponte, Nicholas (1995), *Being digital* (New York: Vintage).

Newell, Arthur and Simon, Herbert (1976), 'Computer science as empirical enquiry: Symbols and search', *Communications of the Association of Computing Machinery,* 19 (3), 113-26.

Newen, Albert; Gallagher, Shaun and De Bruin, Leon (2018), '34E Cognition: Historical Roots, Key Concepts, and Central Issues', in Albert Newen; Leon De Bruin and Shaun Gallagher (eds.), *The Oxford Handbook of 4E Cognition* (Oxford University Press), 0.

O'Regan, Kevin J. (2011), *Why red doesn't sound like a bell: Understanding the feel of consciousness* (New York: Oxford University Press).

Olsen, Eric (2019), 'Personal identity', *Stanford Encyclopedia of Philosophy*.

Pearl, Judea and Mackenzie, Dana (2018), *The book of why: The new science of cause and effect* (New York: Basic Books).

Pfeifer, Rolf and Bongard, Josh (2007), *How the body shapes the way we think: A new view of intelligence* (Cambridge, Mass.: MIT Press).

Piccinini, Gualtiero (2021), 'Computation in physical systems', *Stanford Encyclopedia of Philosophy* (Summer 2021 Edition; Palo Alto, Cal.: CSLI-Stanford University).

Pink, Thomas (2004), *Free will: A very short introduction* (Oxford: Oxford University Press).

Preston, John and Bishop, Mark (eds.) (2002), *Views into the Chinese room: New essays on Searle and artificial intelligence* (Oxford: Oxford University Press)

Rosenblatt, Frank (1957), 'The Perceptron: A perceiving and recognizing automaton (Project PARA)', *Cornell Aeronautical Laboratory Report,* 85–460–1, 1-29.

Russell, Stuart (2016), 'Rationality and intelligence: A brief update', in Vincent C. Müller (ed.), *Fundamental issues of artificial intelligence* (Synthese Library; Cham: Springer), 7-28.

— (2019), *Human compatible: Artificial intelligence and the problem of control* (New York: Viking).

Russell, Stuart and Norvig, Peter (2020), *Artificial intelligence: A modern approach* (4 edn.; Upper Saddle River: Prentice Hall).

Sacks, Oliver (1985), *The Man Who Mistook His Wife for a Hat, and Other Clinical Tales* (Summit Books).

Scheutz, Matthias (ed.), (2002), *Computationalism: New directions* (Cambridge: Cambridge University Press)

Searle, John R. (1980), 'Minds, brains and programs', *Behavioral and Brain Sciences,* 3, 417-57.

— (1984), 'Intentionality and its place in nature', *Consciousness and language* (Cambridge: Cambridge University Press 2002), 77-89.

— (2004), *Mind: A brief introduction* (Oxford: Oxford University Press).

Shagrir, Oron (1997), 'Two dogmas of computationalism', *Minds and Machines,* 7, 321-44.

— (2022), *The nature of physical computation* (New York: Oxford University Press).

Shanahan, Murray (2016), 'The frame problem', in Edward N Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2016 edition; Palo Alto: CSLI, Stanford University).

Siegelmann, Hava T. (1995), 'Computation beyond the Turing limit', *Science,* (268), 545-48.

— (1997), *Neural networks and analog computation: Beyond the Turing limit* (Progress in theoretical computer science; Basel: Birkhäuser).

Simon, Herbert (1955), 'A behavioral model of rational choice', *Quarterly Journal of Economics,* 69, 99-118.

Simpson, Thomas W and Müller, Vincent C. (2016), 'Just war and robots' killings', *The Philosophical Quarterly,* 66 (263), 302-22.

Sparrow, Rob (2007), 'Killer robots', *Journal of Applied Philosophy,* 24 (1), 62-77.

Sperber, Dan and Wilson, Deirdre (1996), 'Fodor's Frame Problem and Relevance Theory', *Behavioral and Brain Sciences,* 19 (3), 530-32.

Strawson, Galen (2004), 'Free will', *Routledge Encyclopedia of Philosophy* (updated 2011) <https://www.rep.routledge.com/articles/thematic/free-will/>.

Thaler, Richard H and Sunstein, Cass (2008), *Nudge: Improving decisions about health, wealth and happiness* (New York: Penguin).

Thoma, Johanna (2019), 'Decision Theory', in Richard Pettigrew and Jonathan Weisberg (eds.), *The open handbook of formal epistemology* (PhilPapers), 57-106.

Turing, Alan (1936), 'On Computable Numbers, with an Application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society,* 2 (42), 230–65.

— (1950), 'Computing machinery and intelligence', *Mind,* LIX, 433-60.

Varela, Francisco J; Thompson, Evan and Rosch, Eleanor (1991), *The embodied mind: cognitive science and human experience* (Cambridge, Mass.: MIT Press).

Wheeler, Gregory (2020), 'Bounded Rationality', in Edward N Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition; Stanford: CSLI).

Wittgenstein, Ludwig (1953), 'Philosophische Untersuchungen', *Schriften I* (Frankfurt: Suhrkamp 1980), 279-544.