# Exploring Minds

## Modes of Modelling and Simulation in Artificial Intelligence

Hajo Greif [1,2]

[1] International Center for Formal Ontology (ICFO)
Warsaw University of Technology
Plac Politechniki 1
00-661 Warsaw
Poland

[2] Munich Center for Technology in Society (MCTS)
Technical University of Munich
Arcisstraße 21
80333 Munich
Germany

E: mail@hajo-greif.net W: https://hajo-greif.net

Manuscript accepted for publication in *Perspectives on Science*, Special Issue on Exploratory Models and Exploratory Modelling in Science, Guest Editors: Axel Gelfert, Grant Fisher, Friedrich Steinle

Version of December 31, 2019 (final draft post-refereeing)

# Exploring Minds: Modes of Modelling and Simulation in Artificial Intelligence

**Abstract:** The aim of this paper is to grasp the relevant distinctions between various ways in which models and simulations in Artificial Intelligence (AI) relate to cognitive phenomena. In order to get a systematic picture, a taxonomy is developed that is based on the coordinates of formal versus material analogies and theory-guided versus pre-theoretic models in science. These distinctions have parallels in the computational versus mimetic aspects and in analytic versus exploratory types of computer simulation. This taxonomy cuts across the traditional dichotomies between symbolic / embodied AI, general intelligence / cognitive simulation and human / non-human-like AI.

According to the taxonomy proposed here, one can distinguish between four distinct general approaches that figured prominently in early and classical AI, and that have partly developed into distinct research programmes: first, phenomenal simulations (e.g., Turing's "imitation game"); second, simulations that explore general-level formal isomorphisms in pursuit of a general theory of intelligence (e.g., logic-based AI); third, simulations as exploratory material models that serve to develop theoretical accounts of cognitive processes (e.g., Marr's stages of visual processing and classical connectionism); and fourth, simulations as strictly formal models of a theory of computation that postulates cognitive processes to be isomorphic with computational processes (strong symbolic AI).

In continuation of pragmaticist views of the modes of modelling and simulating world affairs (Humphreys, Winsberg), this taxonomy of approaches to modelling in AI helps to elucidate how available computational concepts and simulational resources contribute to the modes of representation and theory development in AI research – and what made that research programme uniquely dependent on them.

## 1 Introduction

Artificial Intelligence (AI) is a diverse research programme that includes many, partly competing, approaches and that comes with many, often diverging, aspirations. Despite its disunity and its perpetually pre-paradigmatic status in Kuhnian terms, AI is currently experiencing a renaissance. In doing so, it also undergoes numerous transformations that make it ever more difficult to determine its purpose and content.

This paper seeks to systematically account for the diversity of AI and to identify its common themes on the grounds of a critical reconstruction of the various modes of modelling and simulation that are used in the field. These distinctions are instructive for the modelling and simulation debates in the philosophy of science, but also for the philosophy and theory of AI: how do models and simulations relate to world affairs? How and to what purpose are they made to do so? What bearing do answers to these questions have on various approaches to AI?

In order to at least begin to answer these questions, a taxonomy is developed that builds, first, on the distinctions between formal and material analogies in scientific modelling, and between theory-guided and pre-theoretic models (Section 2). These two aspects of models in science have relevant analogies in the realm of computer simulations: on the one hand, simulations comprise computational and phenomenal elements, typically but not necessarily in conjunction. On the other hand, one can distinguish between analytic, mathematical and synthetic, exploratory types of computer simulations (Section 3). Using these conceptual distinctions as co-ordinates, four approaches to modelling and simulation in AI will be outlined (Section 4). These distinctions cut across a variety of well-established dichotomies that have been used to conceptually sort the field of AI. This taxonomy will help to elucidate how available computational concepts and simulational resources contribute to the modes of representation and theory development in AI research – and what made that research programme uniquely dependent on them (Section 5).

## 2 A Variety of Models

In the various analyses of the role of models in science that emerged towards the end of the 19th century, there are two recurring questions: first, how do models represent an object or “target system”? The two basic types of representation are formal and material, but they occur in various forms and arrangements. Second, what is the relationship between models and theories? Models can either be derived from an existing theory, or they can enable the development of a theory. In principle, however, they can also be completely independent of any theory.

The authors who first introduced the concept of models into philosophical reflections about science were quite inclusive, if not vague, in their accounts of what models are and how they represent. Ludwig Boltzmann (1902) refers to models as "tangible representations" in the first sentence of his "Model" *Encyclopaedia Britannica* entry, but then allows models to be "constructed [...] in thought" and "mentally conceived" at the end of the very same sentence (1902, 788).

Heinrich Hertz (1899, 1-2) commences his inquiry into models from a notion of "our conceptions of things" as mental images and develops a picture-theoretic account of models that defines them mathematically as point-to-point mappings between states and transformations in image and original while taking them to directly concern perceptual experience and experimental measurements (Hertz 1899, 30). He maintains that modelling relations are identical to the relations between mental images and the "things themselves" in general (Hertz 1899, 177). All science, then, is model-based, and the models it employs involve both mathematical mapping, which may or may not happen 'in the head', and more material, observation-based relations.

Hertz' and Boltzmann's broad and inclusive views of models have given way to more differentiated accounts of how models relate to world affairs in classical mid-20th century philosophy of science: Max Black (1962, Ch. XIII) distinguishes between analogue models, theoretical models and "archetypes" (plus physical scale models and mathematical models, which are deemed less pertinent to scientific inquiry).

The purpose of analogue models is to "reproduce as faithfully as possible in some new medium the *structure* or web of relationships in the original", where that reproduction establishes relations of isomorphism, formally defined in similar fashion to Hertz' view as "point-to-point correspondence" between relations in the model and relations in the original, which may but need not depend on a pre-existing theory of the target domain (Black 1962, 222, emphasis in original). In the mathematical parlance introduced by Black, isomorphism is an identity relation between structures in terms of a bijective function that individually pairs every element in one structure with exactly one element in the other. As examples of changes of medium, Black cites "hydraulic models of economic systems, or the use of electrical circuits in computers" (1962, 222), under the premise that the formally expressed isomorphism relations hold independent of the choice of medium.

Black's theoretical models are more conceptual and pragmatic and less formal in nature, involving the transfer of theoretical concepts from a well-explored domain of science to a less explored one, in order to facilitate the building and testing of hypotheses in the latter. They do not presuppose a theory of the target domain but are employed to develop it in the first place. Unlike analogue models though, the performance of their epistemic tasks depends on the choice of medium. Given that, according to Black's example example, Clerk Maxwell modelled the electrical field "in terms of the properties of an imaginary incompressible fluid" (Black 1962, 226), the properties of that incompressible fluid will have a direct bearing on the set of hypotheses concerning the electrical field that are derived from the model.

Lastly, Black's archetypes are, often implicit and always informal, guiding metaphors that organise an entire field of inquiry, and as such bear some resemblance to Kuhnian paradigms. On Black's view, these three types of models stand in a relation of decreasing formality and strictness and increasing generality and epistemic import.

A first attempt at distinguishing models along the kinds of likeness between model and target system on the one hand and between its kinds of relationship to theories on the other was introduced by Ernest Nagel (1961, Ch. 6). Both "formal" and "substantial" analogies, he suggests, can be used to construct a theory or to extend its range of application, or to apply a pre-existing theory. Consequently, two types of analogy are co-ordinated with two types of uses, where these two axes are considered independent.

In her classic account of models and analogies, Mary Hesse (1966) chooses not to detach these two aspects. Instead, she relies on a binary distinction between material, pre-theoretic and formal, theory-guided analogies, which she characterises as follows:

> there is one-to-one correspondence between different interpretations of the same formal theory, which we may call *formal analogy*, and there are pretheoretic analogies between observables [. . . ] which enable predictions to be made from a model. Let us call this second sense *material analogy*. [. . . ] It should be noticed that if material analogies between models and explicanda are to do the predictive job required of them, they must be

> *observable* similarities between corresponding terms and must not depend on a theory of the explicandum. (Hesse 1966, 68, 69, emphasis in original)

Definitionally, Hesse's requirements for models can be pinned down as follows:

**M-1** material models are structures that bear observable and pre-theoretic similarities to, and thereby enable predictions of, their target systems.

**M-2** formal models are structures that express interpretations of a formal theory, without observable similarities being required.

Definition M-2 echoes the "syntactic" view of theories, under which models are formally described structures of which all axioms of a theory are true, and which provide concrete values to its variables (for a paradigmatic formulation of this view, see Tarski 1953 and its critical discussion in Suppes 1960). Under the syntactic view, models may remain independent of any observable phenomena and their explanation, although in practice they typically serve the mapping of a theory onto a given set of phenomena. The formal analogies involved are 'vertically' determined by the theory, in that the same equations are used in otherwise disjunct domains of phenomena that do not display similarities in observables between them from which predictions could be generated. Hesse cites the Mathieu's equation as an example, as it equally applies to the behaviour of elliptic membranes and to the movements of a balancing artist among other things. Despite being described by the same set of theoretical axioms, these phenomena do not display observable analogies that otherwise might allow for prediction or explanation of one domain in light of the properties of the other.

Hence, the primary mark of distinction between material and formal models is not the way in which they are expressed but, first, their mode of reference to observables. Second, they are distinguished by being determinants of theory or determined by theory respectively.

With respect to the role of material models in theory-building, a key role accrues to, in Hesse's words, "neutral analogies". They are relations assumed to hold between model and target system that, at the time of the model's introduction, cannot be proven to hold or not to hold. Their specific value lies in their capacity to explore the properties of the target system by generating predictions of its behaviour in light of that analogy, which are to be confirmed or disconfirmed at a later stage of inquiry.

Neutral analogies are thus identified as the "growing points" of a theory (Hesse 1966, 8-10). In this specific and important sense, the function of models is exploratory, as guides for theory construction that are neither defined nor constrained by pre-existing theory. In order to elucidate the exploratory role of models, Hesse chooses the example of the "ether" in the development of the wave theory of light. When the wave theory of sound was transferred to explanations of the behaviour of light, ether was chosen as the neutral analogue of "air", besides a number of established positive and negative analogies between sound and light. There was well-founded knowledge concerning the behaviour of air as a medium for the transmission of sound waves, from which predictions were derived concerning the behaviour of the medium for the transmission of light waves – which were ultimately not confirmed.

Hesse's dual requirement for material models, according to M-1, to be pre-theoretic *and* display observable similarities to the target system is quite descriptive of how models are in fact often used in science. Conceptually and logically, the parts of the conjunct are independent though, either in the way described by Nagel (1961) or in the more fundamental sense of models being genuinely autonomous from theory (Morrison 1999). A model may bear observable similarities to the target system *and* be based on a pre-existing theory, for example when functioning as an illustration or exemplification of the theory's empirical content. Hence, a third definition can be introduced that aims at a class of models whose role in advancing science may be comparatively modest, but whose role in elucidating it should not be neglected:

**M-3** models are structures that materially exemplify the observable consequences of an established theory, demonstrating its bearing on phenomena.

Conversely, a model might be pre-theoretic in a similar fashion to that envisioned by Hesse for material models while referring to observables in a more abstract and indirect way. This possibility has been explored by Bas van Fraassen (1980) in particular:

**M-4** models are structures that bear a formally defined relation of isomorphism to their target system, without either observable similarities or a pre-existing theory being presupposed.

Definition M-4 matches the "semantic" view of theories, under which theories are not sets of propositions articulated in a specified formal language. Instead, they are presented “in the first instance by identifying a class of structures as its models”, where “the same class of structures could well be described in radically different ways” (van Fraassen 1980, 44). The classes of structures involved are identified by reference to relations of isomorphism. Theories, in turn, are descriptions of a set of related isomorphic structures or “a family of models” that share a specific set of properties (van Fraassen 1980, 65).

Accordingly, models are epistemically more central to science than theories. By the same token, theories are genealogically not prior to models, but defined by them. According to van Fraassen’s own example, Isaac Newton attempted to map the apparent, observable movements of bodies both onto one another and onto the true movements of bodies that he assumed to occur in an absolute space. If these images are understood as isomorphic, the observed movements can be represented as differences between true movements in absolute space. Absolute space is not observable itself but described by the family of models of apparent motion.

An important implication of the last of the above approaches to modelling (M-4), but partly also of the first (M-1), is that it substitutes the condition of truth for the propositions of a theory with the condition of empirical adequacy (van Fraassen 1980, 12-13). If models are images, as Hertz (1899) suggested, or even if models are any closer in nature to images than to propositions, they cannot be strictly speaking true or false but only more or less faithful or adequate in a number of respects. A model may vindicate but does not verify any theory that it might support. If one abstains from ascribing truth values to pictures or other non-propositional structures, and hence also to models (as Hertz 1899 does), and especially if one allows theories to be non-propositional structures, too (as van Fraassen 1980 does), empirical adequacy will be the strongest possible normative judgement on the value of a theory or the models on which it may rest.

However, the empirical adequacy of a model is a function not only of empirical fit but also of the purposes and resources of inquiry. This pragmatic leitmotif of modelling in science, and in fact of science altogether, is explicit already in Hertz (1899). It also figures in van Fraassen (1980) and other anti-realist approaches as well as in what is called the “practice turn” in the philosophy of science (inaugurated by Hacking 1983 and recently summarised in Soler et al. 2014).

The degrees of freedom in how models may relate to world affairs, pre-theoretic versus theory-based and materially versus formally, and the pragmatic considerations that go into designing these relations, become particularly manifest in the most recent major addition to repertoire of scientific methodology: computer simulations.

## 3 Simulations and Models

A paradigmatic albeit extremely broad definition of simulations that puts them in the context of scientific modelling is given by Stephan Hartmann: "*a simulation imitates one process by another process*," where "the term 'process' refers solely to some object or system whose state changes in time" (Hartmann 1996, 83, emphasis in original). According to this definition, simulations, in the most straightforward cases, can plainly *be* dynamic models that seek to trace the outward behaviours or the internal dynamics of a system. Consequently, a computer simulation will be a computer implementation of such a dynamic model. In scientific practice, simulations are typically computer simulations, so that the terms "simulation" and "computer simulation" come to be used synonymously. (I will henceforth refer to computer simulations simply as "simulations".)

Paul Humphreys introduces a paradigmatic definition of computer simulations that goes beyond this de facto constraint on simulations and moves from material to formal criteria: "A computer simulation is any computer-implemented method for exploring the properties of mathematical models where analytic methods are unavailable" (Humphreys 2004, 107-8). This definition is too narrow, as the author himself admits, as computer simulations often provide solutions to models that are analytically tractable. It is also too broad, he concedes, as it also covers areas of computational science otherwise unrelated to simulations.

Between these two poles of maximal generality and specificity, one will find accounts of computer simulations that highlight both their dynamic, material and their mathematical, model-solving character. Typically, the connection between a simulation and its target system is conceived of as a two-step relation, as critically discussed by Eric Winsberg (2010, 9-11, 19-25): simulations are, in a first step, computer implementations of formal models. They are designed to algorithmically solve, that is, to provide concrete values for, the variables of those models. The simulation's output counts as the model's solution. In a second step, the solution helps to decide on the model's empirical adequacy. In the parlance of simulation-based science, if the algorithms in question match the formal model, the simulation is *verified.* If the model underlying the simulation is found to correctly represent its target, the simulation is *validated.*

According to this two-step image, a simulation represents its target system by first realising the underlying model. However, verification and validation may not be so clearly distinguishable in practice. A model and its implementation are often mutually adjusted in pragmatic fashion in order to make them both solvable and empirically adequate. Such pragmatic considerations motivate Winsberg (2010, 19) to characterise simulational methods as “motley”. Simulations might be informed by established theories, but also resort to theoretically unprincipled assumptions, intuition, tricks and tinkering, so as to bring model, simulation and, possibly, observation into accordance. The implications, however, might not be as relativistic as it may seem at first sight, as a greater amount of background knowledge, relative to traditional observational and experimental methods, will compensate for ad-hoc practices in simulation modelling (Winsberg 2010, 70-71).

Most authors consider simulations are sufficiently defined by their formal model-solving properties, as far as “core simulations” are concerned, in contrast to complete “computer models” (Humphreys 2004, 110) or what Winsberg (2010, 16-17) calls “models of the phenomena” (see also Hartmann 1996, 84). In boundary cases, simulations may have no empirical referent at all.

Conversely, the design and logic of computer simulations are subject to material constraints imposed by the available mathematical and computational resources. Humphreys (2004, 56) argues that “*[m]ost scientific models are specifically tailored to fit, and hence be constrained by, the available mathematics*” (emphasis in original), and that “*It is the invention and deployment of tractable mathematics that drives much progress in the physical sciences*” (Humphreys 2004, 55, emphasis in original). Johannes Lenhard (2015) adds the complementary observation that mathematical models have to be tailored to a given set of computational resources. At any given time, an empirical problem may be treated by simulational methods only to the extent that a specific set of computational and mathematical tools is available.

Within the limits of those pragmatic enablers and constraints, a simulation will typically also comprise a material, observable rendering of their output, mostly in the form of visualisations or other representations of its output with respect to the variables identified in the underlying formal model (Humphreys 2004, 111; Winsberg 2010, 31-34). That rendering has its own criteria of empirical adequacy, in terms of observable similarities. The kind of presentation is chosen in accordance with the visualisation methods that are available, the characteristics of the target system that are considered particularly relevant, and the ways in which these are best communicated. Conversely, it is historically plausible that the lack of suitable means of visualisation and communication has impeded the introduction of advanced simulation-based models in the early days of computing, as Rainer Hegselmann (2017) argues with reference to the Sakoda model as an equally sophisticated and unsuccessful early foray into computer-based social science.

Despite de facto usually travelling together, the formal and material aspects of modelling are partly separate affairs in systematic respects: a straightforward mathematical solution to a model is amenable only in the case of what Peter Asaro (2011, 93) calls "analytic" simulations, where it is possible to (locally) apply a pre-existing (general) theory. In this paradigmatic set of simulation-based investigations, a formal theory of the target system is involved, which will inform a set of formal models to be implemented in a computer. On the basis of input data from the target system, the computer will then generate an output that can be demonstrated to be isomorphic to what the theoretical propositions would predict. In boundary cases, no empirical referent might be involved at all, if and when the simulation solves equations using fictional or altogether non-referring data.

Conversely, "synthetic" simulations are designed to produce analogues of the phenomena that are to be investigated in the absence of theory. Hessean material analogies, in terms of observable similarities, take centre stage. From the observation of these analogies, a set of theoretical hypotheses may be generated and then be subject to further testing. For example, a sequence of variant or even contradicting models can be tested and compared or the effects of fictitious values of key variables surveyed. In this type of simulation, the use of computers as universal machines actually *exploits* the lack of material constraints. In a complementary boundary case to that of non-referring models, it might happen that a genuine theoretical understanding of the target system is not even sought, or is considered out of reach. In such cases, behavioural similarities between model and target may be considered sufficient for the purposes at hand. To some authors, the latter seems the primary use of simulations under 'technoscientific' conditions (Nordmann 2011; Galison 2017).

The two types of boundary case aside, the distinction between analytic and synthetic simulations bears an analogy to the distinctions between theory-guided and exploratory experimentation (Burian 1997; Steinle 1997; Ribe and Steinle 2002; Waters 2007) and, relatedly, between theory-guided and exploratory modelling (Fisher 2006; Gelfert 2016). Where theory-guided experimentation and modelling amount to the testing of theoretical hypotheses by empirical means that are to a large degree determined by the respective theories, exploratory experimentation and modelling lack both the guidance and the constraints provided by a pre-existing theory. Instead, they have been described as practices of "getting a feel" for the phenomenon or model (Gelfert 2016, 96). This seeming lack of conceptual and empirical focus is not to be considered a deficit though, but serves to ground an alternative approach that has respectable historical credentials, from Goethe to Faraday (Steinle 1997; Ribe and Steinle 2002).

With respect to exploratory experimentation, "Its defining characteristic is the systematic and extensive variation of experimental conditions to discover which of them influence or are necessary to the phenomena under study" (Ribe and Steinle 2002, 46). Its aim is "to open up the full variety and complexity of a field, and simultaneously to develop new concepts and categories that allow a basic ordering of that multiplicity" (ibid.). Variation of conditions, as the previous quotes suggest, is anything but random. Instead, it follows preliminary conceptions of how the phenomena might be affected by such variation, where these conceptions are judged to have a promise of furnishing explanations. Exploration paradigmatically – but not always, as will be demonstrated – serves the development of a theory, and embodies the key characteristics of the material, pre-theoretic type of models described by Hesse (1966) in more poignant fashion perhaps than she envisioned herself.

The material character of exploratory modelling in particular stands in a peculiar relation to the properties of computer simulations outlined above: Given that the mathematical structures involved in many models are particularly keen to exploratory manipulation, variation of parameters may come "too cheaply" as compared to experimental practice and its material constraints (cf. Gelfert 2016, 79, 82). Simulations, if anything, facilitate the manipulation of mathematical structures across the entire range of available possibilities. Given that computers are universal machines in Turing's sense, for being capable in principle of accomplishing any logico-mathematical task that is amenable to a solution at all (Turing 1936), the role of an elaborate theory to constrain the range of simulational possibilities, and give direction to an inquiry, has been argued to be more important than in experimental practice or direct, artefact-based modelling (Asaro 2011; Guala 2002). Such a theory is not always available though, nor does it have to be, as long as the range of possibilities is meaningfully and methodically constrained.

Simulations with exploratory functions have assumed particular importance in one scientific field that has received comparatively little attention in the philosophy of modelling and simulation to date but that displays an unrivalled variety of approaches to modelling and simulation: Artificial Intelligence. In its origins as an extremely open and theoretically under-defined field, it offered itself for a multitude of approaches to modelling and simulation, some of which are still of relevance today.

## 4 Varieties of Models and Simulations in AI

AI co-originated with computer science and played a formative role in the development of the cognitive sciences. Unlike most well-established sciences, and unlike physics in particular, the cognitive sciences were not in a position to rely on an axiomatic theory of their subject matter. The cognitive sciences arose in the mid-20th century from a growing dissatisfaction both with classical introspective psychology and with behaviourism. Neither of these two approaches had a reasonably well-developed, let alone an axiomatic, theory of cognitive processes at their disposal. Introspective psychology lacked the credentials of objective science altogether, whereas behaviourism did not accept cognitive phenomena as amenable to and worthy of scientific consideration, and restricted itself to a systematic inquiry into observable behaviour. Confronted with a number of explanatory problems unsolvable by either behaviourism or introspective psychology, the cognitive sciences were founded in order to develop a novel approach to scientific psychology. Lacking a psychological theory to rely on, the endeavour commenced with models, and with computer models and their implementations in particular, in the hope to be able work its way upwards to a comprehensive theoretical account.

AI played a prominent and particular role in the development of the cognitive sciences in that it, first and foremost, provided them with a novel set of methods. However, AI was in a unique position that went much deeper than methodological innovation. In his seminal contribution to applied mathematics, “On Computable Numbers”, Alan Turing (1936) conceived of theoretical machines, now known as “Turing Machines”, along the lines of a subset of cognitive operations, namely those involved in the accomplishment of basic routines of arithmetic or “computation”. As Turing demonstrated, complex mathematical operations may be broken down into such elementary operations in such a way that they could in principle also be accomplished by those machines. Hence, these machines would be able to solve any logical-mathematical task that is amenable to a solution at all. However, the machines designed by Turing were first and foremost *theoretical* machines, based on his theory of computation. The functions of these theoretical machines were purposefully defined in abstraction from any specific realisation and application, while being modelled on the basis of the performances of human “computers”. In this sense, Turing used the material model of human computers as a neutral analogy (M-1) in the development of his theory of computation. (Another material model that inspired the design of Turing’s theoretical machines on a more concrete level was the mechanical typewriter, as suggested by Andrew Hodges 1983, 96-98.)

Accordingly, there is an element of modelling human cognitive processes that went into the foundations of computer science. In turn, the endeavour of the cognitive sciences was to build on this analogy in order to develop computer-implementable scientific models of higher-level human cognitive abilities. However, there also was a, more or less implicit, suggestion of a direct, behaviour-based analogy between human and machine accomplishments. This analogy was introduced in fairly playful manner in Turing’s later essay “Computing Machinery and Intelligence” (1950), and there is a remarkable ambiguity in Turing’s work between these analogies (lucidly described in Sprevak 2017). However, the latter analogy was transformed into a research programme by later authors, and subsequently shaped public perception of AI, as may be illustrated by the fact that the Loebner Prize competition for a computer that passes the “Turing Test” and related efforts (see also Warwick and Shah 2016).

With respect to the role of human cognition in AI modelling, it is worth highlighting that from its very beginnings, AI comprised two expressly distinct research programmes. Only one of them has human cognition as its *topic* and thereby belongs to the cognitive sciences proper, whereas the other paved the way for the success of AI in many fields of application, but did not pursue a scientific agenda with respect explaining human cognitive abilities. Remarkably, the latter approach corresponds to the original definition of "Artificial Intelligence", in which intelligent, cognitively advanced human behaviour serves as a *resource* for modelling. An early anthology of foundational works in AI (Feigenbaum and Feldman 1963) was divided into two parts accordingly (see also Ringle 1979; Asaro 2011):

S-1 computer programs that solve complex intellectual tasks without prima facie regard to modelling cognitive functions; the aim is "to construct computer programs which exhibit behavior that we call 'intelligent behavior' when we observe it in human beings" (Feigenbaum and Feldman 1963, 3);

S-2 simulations that implement models of the structure or functions of natural cognitive processes, without prima facie regard to creating similarities in observable behaviour; the aim is the "simulation of cognitive processes" (op. cit., 269).

Simulations of the first type (S-1) are also called "behaviour-based simulations" because they make no or no systematic attempt to establish analogies between the computational processes involved in a computer's attempt to solve a task and the cognitive processes involved when a person tries to solve the same task. Conversely, it is the task of S-2 simulations to provide analogies at the level of cognitive processes and functions, which are not necessarily linked to the generation of behavioural similarities at the observational level. Taking such similarities to be indicators of underlying processes and proposing the supposed analogies as definitions of intelligence, or even as evidence for machine intelligence, has been identified as a fateful misunderstanding of AI by several authors (e.g., Copeland 2000; Moor 1976; Whitby 1996).

The distinction between S-1 and S-2 simulations in AI can be further refined and rendered in a new light by reference to the previously introduced coordinates of material versus formal and theory-guided versus explorative modelling (M-1 to M-4). It will thereby become possible to map out more precisely which scientific programs and theories the various AI approaches serve, and how they do this.

**AI-1 – behavioural simulations:** Some AI simulations are of an entirely phenomenal and material kind. They are concerned with the imitation of a subset of the observable behaviours of human beings, without a prima facie regard to the underlying structures and functions they may serve. Such is paradigmatically the case for Turing's "imitation game" (1950) and the so-called "Turing Test"-based approaches to AI that were derived from it. Turing introduces the imitation game as a thought experiment in which he asks the reader to imagine computers being involved in a blinded conversation between a (female) human being, a (male) impersonator and an interrogator. The blinding of the conversation would be accomplished by restricting communication between the players to teletype messages. The question is whether a machine substitute of the human impersonator in such a game could be identified by the human interrogator within a certain time with a certain degree of reliability. Turing himself emphasised that his imitation game neither provides a definition of intelligence nor a proof of machine intelligence, as he states in a 1952 BBC broadcast quoted in Copeland (2004, 494-5). He even calls the question "'Can machines think?' [...] too meaningless to deserve discussion" (Turing 1950, 442), and proposes to replace it with the question whether a digital computer would do well in the imitation game.

As there is no pretence of systematically accounting for the structure or function of human thought in this approach to AI, there is no prima facie theoretical underpinning to these 'black box' simulations. Nor do these simulations presuppose a model that represents relevant properties of its target system in order to develop a theory from it. All that Turing does is to informally suggest some analogies between human and machine capabilities. He explores these analogies more systematically in other works (especially Turing 1948), which are closer to what will be discussed in AI-3. Accordingly, the scientific modelling relations M-1, M-2, and M-4 are not applicable to the imitation game and Turing-Test-based AI. The status of this kind of simulation as scientific has consequently been contested, which has not prevented them from being perceived as the paradigm of AI by the general public and AI critics alike. However, these approaches may serve relevant demonstrative purposes (Ringle 1979; Asaro 2011), and hence fall under the M-3 category of theory-guided material models. They demonstrate the force and scope of Turing's theory of computation by making computers exhibit partly human-like behaviours, which were beyond what, in Turing's time, was deemed within the reach of machines de facto and by definition. The expected effect was that our notions both of machines and of thinking are altered in the course of the spread of digital computers. This is Turing's declared goal in (1950, 442). Conversely, the simulation's observable human-like behaviour can be mobilised to identify some of the behavioural cues by which human beings recognise each other as intelligent beings.

**AI-2 – models of general intelligence:** Some AI simulations are implementations of models that are designed to represent a selection of cognitive accomplishments in such a way that they collectively serve to define a theory of general intelligence. These AI models embody abilities of, for example, theorem-proving, logical problem-solving, chess- and Go-playing or data analysis that one would call intelligent when observed in human beings. General intelligence is what systems like logic-based AI (McCarthy 1960) or the Logic Theory Machine (Newell, Shaw, et al. 1963) were supposed to achieve. The models involved were not designed to furnish or support explanations of how, in particular, human nervous systems or, more generally, human beings accomplish the logic-based tasks in question. Instead, they support explanations of how these tasks are to be solved by any intelligent system.

Different in approach but similar in general outlook is the more application-oriented side of classical connectionism or neural network modelling (Rumelhart and McClelland 1986, Vol. 1). In one of the most successful current AI approaches derived from the latter, Deep Learning algorithms are used for object or image recognition (LeCun et al. 2015; Schmidhuber 2015). Structures are extracted from data sets on numerous levels of abstraction in order to generate representations or classifications that are interpretable for humans. However, the levels of abstraction and the processing stages do not need to correspond to structures and processes in the human nervous system. Similar to AI-1, observable similarities between logic-based or Deep Learning models and human cognitive processes are welcome, but they are not the actual aim of inquiry. However, based on partial formal isomorphisms between the logical operations or neuronal processes in model and human thinking, these models can be argued to form unified classes that define theories of general intelligence in human beings and machines.

To the extent that a theory is a family of models, and to the extent that formal isomorphisms are what unites this family, AI-2 models fall under the M-4 class. A separate and explicit, let alone an axiomatic, definition of what general intelligence is will not be required under this approach, as the models in the class sufficiently define such a theory. While ‘brute force’ computational solutions of logico-mathematical problems cannot provide foundations to a theory of general intelligence, general-level formal analogies of the aforementioned kind will be sufficient for a system to count as Artificial Intelligence.

Remarkably, theorem proving, chess playing and their kin were very explicitly chosen by early and classical AI researchers over other constituents of human mental life, such as emotion or embodied phenomena, not because these were deemed irrelevant features, but for two complementary reasons: On the one hand, logically explicable cognitive tasks were amenable to formalisation and computer simulation by the means available at the time (as recounted, for example, by Boden 2006, 11). On the other hand, the solution of logic-based problems was also considered to be the core or even the exclusive domain of cognition *sensu strictu*, and as such treated apart from other aspects of the human condition. Consequently, this approach to AI remains only tangentially concerned with systematic inquiries into the human mind as a whole. Instead, it informed all kinds of application-oriented AI, where intelligent problem-solving remains the key objective.

**AI-3 – material models of cognitive processes:** Some AI simulations are implementations of models designed to contribute to the development of a theory of the general laws of human cognition. The observable properties of the computational models and the regularities therein serve as templates for those laws. The modelling relations involved here are of the pre-theoretic material kind (M-1): a set of positive, negative and neutral analogies between computational and mental processes is postulated, where a central task of further inquiry is to determine the neutral ones as being either positive or negative.

Prominent examples for this kind of approach include Marr’s account of the stages of visual processing (Marr 1982) and classical cognitively oriented connectionism (Rosenblatt 1958), its precursors (McCulloch and Pitts 1943; Turing 1948) and descendants in the more cognitively oriented side of neural network modelling (Rumelhart and McClelland 1986, Vol. 2, where the two volumes of this collection neatly replicate the distinction in Feigenbaum and Feldman 1963 discussed as S-1 / S-2 above). In distinct ways and with respect to distinct levels of cognitive processes, computational concepts are used to generate theoretical hypotheses concerning the properties and regularities of the respective cognitive processes and structures. The focus moves away from general problem-solving (as in AI-2) to an investigation into the constituents of natural cognition, such as perception and the functions of the nervous system. Marr (1982) conceives of vision as the process of constructing, throughout the stages of perceptual processing, representations of the information contained in the retinal image: from the representation of the two-dimensional retinal image, a "primal sketch" is developed of edges, surfaces, and textures, and then a representation of the orientation and apparent motion of these forms. Finally, a three-dimensional representation of spatially situated objects is generated. All of these stages are described in terms of computational operations, and involve testable predictions as to how visual perception actually works in biological organisms, however without postulating one-to-one correspondences between the algorithms and hardware used and organic structures and neuronal processes. Many of Marr's assumptions have meanwhile been refuted by neurobiological evidence – and the neutral analogies thereby established as negative.

Conversely to Marr's approach, connectionism devises models of the basic structure and operations of the nervous system. These models comprise a large number of units, representing neurones organised in several (input, hidden and output) layers, and weights that represent the effects of the synapses that connect them. The computer simulates the connection patterns and activation values of the neurones, where neurones are taken to perform computational operations on the input signals they receive, thereby transform them into a new signal and pass it on to the next level. Instead of modelling certain, rather abstractly conceived, elements of cognitive processes, connectionism targets concrete components of the nervous system that realise these processes. A recent approach that develops connectionist ideas into a Bayesian model of "predictive processing" in animal and human nervous systems is presented in Clark (2013). Another contemporary approach that explores the possible import of advanced Deep Learning methods on cognitive inquiries, in particular perceptual abstraction, is developed in Buckner (2018).

If the cognitive sciences are defined as "*the study of mind as machine*" (Boden 2006, 9, emphasis in original), AI-3 approaches do not merely provide it with some methods of inquiry but also with – diverging – research programmes that say *as which kind* *of machine* the mind shall be studied. If the way in which a model is realised partly depends on the material and conceptual resources available, and if these both enable and constrain the ways in which that model relates to a world affair, Marr's assumption that one can separate a computational theory of cognitive phenomena from the "gory details of algorithms that must be run" (1977, 38) might turn out to be too idealising. Apart from its role in the cognitive sciences, this type of AI also inspired a variety of practical applications, but in a different way than AI-2: certain cognitive features serve as models *for* a technological solution, as, for example, in neural network-based applications that provide analogues of neuronal processes but are dedicated to other, more practical purposes.

**AI-4 – cognition as computation:** Some AI simulations are implementations of models that are supposed to provide direct analogues of cognitive processes, with strict one-to-one correspondence relations between computational operations and cognitive processes that are based on Turing's (1936) theory of computation. In this case, the elements of an axiomatic theory of computation and their interrelations determine the elements of a model of cognitive processes and their interrelations, under the premise that the same theory equally applies to both domains. The computational states and processes proposed in the model are expressions of the propositions of a theory that determines relations of formal analogy between computational and cognitive processes. The theory of cognition in AI-4 is co-extensive with a theory of computation. Hence, the modelling relation in question involved here formal and theory-guided (M-2).

This analysis applies to the paradigmatic statement of strong symbolic AI, the physical symbol system hypothesis (Newell and Simon 1976; Newell 1980). The basic hypothesis is that both computation and human thinking consist in the rule-governed manipulation of meaningful symbols. More precisely, the theory of computation involved postulates that computation is the rule-governed manipulation of symbols, whereas cognitive processes are one class of phenomena involving content-bearing symbols that exemplifies this theory. A physical symbol system is physical because it consists of physical entities, namely symbol tokens. It is a system because those symbol tokens are parts of expressions in which their relations are logically determined, and hence can be computed. And it is symbolic because, first, those expressions refer to objects, processes or other expressions and, second, the system can interpret these expressions. If a computing system exhibits these properties, according to the physical symbol system hypothesis, it can, in and of itself, realise the necessary and sufficient conditions for the presence of cognition. It will thereby not merely represent features of cognitive processes, but embody them. Any mind will then be a physical symbol system, and vice versa, so that cognition *is* computation (Pylyshyn 1980). This sort of argument works in two complementary ways: on the one hand, cognitive models, according to strong symbolic AI, are "computational in a dual sense, in that they not only use computers to do the complex calculations required for modeling, but also postulate that minds are actually performing a kind of computation" (Thagard 2014, 534). This kind of dual computational modelling relation raises the question whether it applies to "biologically realistic neurocognitive models", too (ibid.). On the other hand, some computational processes will by definition be cognitive processes, provided that certain conditions concerning the kind and complexity of the computations involved are met.

Hence, strong symbolic AI regards human thought as computational by nature, provided that the computations are performed on semantically meaningful symbols. This latter premiss is not further explained in this approach. This has exposed strong symbolic AI to critiques along the lines of the "symbol grounding problem" (Harnad 1990): The foundational question remains unanswered as to how the symbols involved in cognitive phenomena come to have their semantic content. Taken by itself, the symbols' computability does not provide the requisite information.

## 5 Discussion

The preceding descriptions should have made clear that the modelling relations and their roles are clearly distinct in each case. Only in AI-3, the simulations assume the exploratory, theory-guiding role of models in science envisioned by Hesse (1966), but the models developed under this type of approach have been developed into competing, and partly mutually exclusive, theoretical accounts of human cognitive abilities. They certainly have not given rise to a unified theory of human cognition.

In AI-2, a more general but still exploratory character of the modelling relations is bought at the cost of being focused not on natural, human cognition but on higher-level isomorphisms between computational and some cognitive processes. This, however, was the original self-understanding of AI as a research programme in the first place. The scientific clout of general intelligence approaches has turned out to be relatively more modest with respect to explaining natural cognitive phenomena, but they proved to be more enduring, although mostly outside cognitive inquiries.

Turing Test-based approaches (AI-1) remain highly visible but are as tenuously related to scientific modelling as ever, while having shaped the public image of AI. In the beginning, they primarily stood in the service of one (namely Turing's) theory of computation, being employed by its inventor to demonstrate the power and scope of that theory. The intended analogies to cognitive phenomena remained informal, but meanwhile have come to be interpreted in a stronger sense.

Only in strong symbolic AI (AI-4), the analogies are strictly formal and theory-guided, to the point of making the theory of computation a theory of cognition – and thereby making the properties of model and target system fall into one. It was the key target of philosophical AI critiques but lost much of scientific credentials after the days of "Good Old-Fashioned AI" (GOFAI). It would be worth to further pursue the argument that strong symbolic AI turned out scientifically sterile precisely because it was not in the business of opening up new domains of phenomena to model-based investigation and subsequent theory-building but imposed one theory on one domain of phenomena in top-down fashion.

It should be noted that these approaches are not necessarily strictly disjunct. Sometimes they blend into each other and can only be differentiated according to their respective primary foci. For example, while AI-2 approaches are separated from AI-3 by their explanatory aims, they are distinguished from AI-4 by their ontological presuppositions: does AI seek a computational theory of general intelligence (AI-2), or does it propose a theoretically principled identity relation to hold on a general level between computational and cognitive processes (AI-4)? It should also be noted that in some prominent cases the same AI researchers have pursued different approaches in different projects (Turing, Newell and Simon are the most notable examples here).

The conceptual distinctions developed in this taxonomy cut across a number of established dichotomies that have traditionally been used to sort the field of AI: First, the proposed classification remains *prima facie* indifferent to the dichotomy between symbolic and embodied processes, which largely separates GOFAI and its heirs from modern or “Nouvelle AI” approaches and the most influential AI critiques alike: do symbolic forms and logical operations of the kind involved in abstract, higher-level human thought, provide the necessary and sufficient ingredients for a model of human cognition in its entirety, or must such a model also or even primarily capture the embodied and environmentally embedded nature of cognitive phenomena? According to the most prominent critique of the symbolic AI in Dreyfus (1979), an approach of the latter kind would be required but is unattainable for AI in principle. Second, the matrix presented above is more differentiated and should be more systematic than the rather intuitive programmatic distinction between the simulation of intelligent behaviour and the simulation of cognitive processes present already in early AI. Third, it operates on a different level than the distinction with which the previous one has been combined in the leading AI textbook, namely between human-like and non-human-like AI on the level of observable similarities, from which the following matrix has been generated: “thinking humanly” versus “acting humanly” versus “thinking rationally” versus “acting rationally” (Russell and Norvig 2010, 1-5).

The distinctions between formal and material models and between pre-theoretical and theoretical models can provide further analysis to the similarity criterion at issue here: if a selection of formally reconstructed features of human cognition is used to design systems that provide solutions to intellectual problems on a general level (AI-2), it is unlikely that the aim of this kind of approach will be a material model or even a theory of human cognition in particular – unless an identity relationship is assumed (AI-4). Conversely, if models of human cognitive traits are used to design systems that are supposed to be similar to human thought or behaviour in some relevant respects, the desired modelling relationship will certainly be material, as it either serves to develop a theory of these phenomena (AI-3) or seeks to demonstrate the observable consequences of an existing theory that is based on a presumed fundamental material likeness between human and machine computation (AI-1).

The key difference between the conceptual matrix proposed here and the established classifications is that the focus here is not on elucidating *what* is modelled and simulated, but on *how* and *for what purpose* this is done. In S-1 and S-2, AI comprised two very different types of answers to “what” questions, but that distinction alone offers little systematic insight into how modelling is or shall be accomplished on either side. The symbolic / embodied dichotomy, in turn, is concerned with presumed properties of the target systems but has implications for how models are supposed to look like: they either will be symbolic all the way down, or they will necessarily have to be embodied. This decision is made on the level of ontological presuppositions rather than methodology though.

The distinction between formal and material models in particular offers an indirect route to capturing a point that is missed by the previous dichotomies: much of classical AI had little concern for how cognitive processes are materially realised, but this is plainly the reverse side of the same computational coin, in that AI had equally little prima facie concern for how its models are materially realised. Computational principles, and hence logico-mathematical rules alone, were taken to serve as the enablers and constraints of AI modelling, as if they were implementation-independent concepts. They certainly were so in principle, by virtue of being thus defined in Turing's theory of computation, but they were not so in practice, where modelling decisions had major consequences for the answers that would be given to "how" and "what" questions alike.

To conclude my discussion by explicating this point, I will now argue for a special relation of AI models to their implementation. This argument has one pragmatic and one conceptual aspect. A specific trait that sets AI apart from other sciences in pragmatic terms is its particular entwinement with the realisation of its models, in terms of a very concrete and material reliance on the tools and technologies of modelling. In fact, if there is a science that is so closely tied to the way in which its models are realised that it will be difficult to point to its models in abstraction from their realisation, it will be AI. All approaches took the functional architecture of the computer as such as their starting point, and most relied on the factual availability of computers. Of course, there is the principled possibility of developing and using computational models that are not actually implemented in digital computers – which was the case for those models which were devised before the advent of the first stored-program computers in 1948. The primary examples are Turing Machines (1936) and McCulloch and Pitts' Logical Neurone (1943). These un-implemented models on their own would have been unlikely to establish and sustain a research programme that fully relies on computational principles. Conversely, these un-implemented models contributed, and in Turing's case were foundational, to the design of the digital computer itself, and hence ultimately their own implementation.

There is no such direct dependence on the availability of computers in the classical natural sciences, which adopted computer simulation methods and integrated them into their methodological repertoire once they were available and came to be perceived as useful. Although physics, chemistry, biology or economics are dependent on the availability of these methods when it comes to mastering increasingly complex subjects matter, these disciplines as such do not comprehensively and existentially depend on the availability of digital computers. (However, classical Baconian science might have existentially depended on the availability of experimental apparatus in similar fashion.) Hence, it might hence be difficult to find any other discipline beyond AI that invented itself and developed its key concepts and theories on the grounds of what computers can do, and that did so across the entire range of possibilities of modelling and simulation.

This specific implementation-dependence of AI models might be counted (and consequently discounted) as a matter of technological fact, but there is a deeper, more conceptual interpretation available: AI's computational concepts and simulation technologies are essential to the discipline because they already purport to represent relevant properties of their target systems in and by themselves. If computers are modelled on some aspects of human cognition, and if, in turn, these aspects are part of what is modelled by computers in AI, there will be a solidly positive analogy between computers and cognition that goes into the foundation of the discipline. It has to be taken for granted even if one does not subscribe to the view that cognitive processes *are* computational processes.

AI's peculiar conjunct of a conceptual claim for a implementation-independence *in principle* of computational models and a particular de facto dependence on their implementation *in practice* makes most sense if one takes the presence of computational structures and processes both in machines and in human minds to be the organising metaphor of AI, or an "archetype" in the Blackean sense. It is an analogy, and above that a material analogy, that has to be accepted in order for the research programme to proceed in the first place. This archetype is established but not explicated in Turing's theory of computation but becomes quite pronounced and elaborated in the Logical Neurone approach (McCulloch and Pitts 1943; McCulloch 1960) and John von Neumann's claim that "the neurons of the higher animals are definitely elements [of digital computing devices]" (Neumann 1945, § 4.2).

Whether this general idea is developed into a strict analogy between the basic principles at work in minds and machines that is reducible to Turing's theoretical principles (which is assumed in AI-4 but not by von Neumann, McCulloch or even Turing himself), or whether it informs the development of further material analogies to guide inquiries into how the mind or the brain works (as in AI-3), is a question that hinges upon the strictness of the interpretation of the computational archetype: does a theory of computation suffice as, and thereby constrain, the theoretical foundation of a computation-based inquiry into cognition, or does the archetype allow for novel computational models and their implementations to be used in the exploration of new domains of phenomena and the development of various theories of cognition? The dynamics of the development of AI after the days of GOFAI, especially within the realm of AI-3 approaches, suggests the latter. If, on the other hand, the diagnosis is correct that Turing's theory of computation itself is based on an exploratory model, namely the human computer, the computational archetype, even in its strict interpretation, will rely on an analogy that is material and pre-theoretic.

The dependence of AI on models as archetypes as such will not distinguish it from other scientific disciplines in their incipient stages, which also develop and coalesce around a central image or metaphor (see, for example, Bensaude-Vincent 2013 and 2001 on synthetic biology and materials science respectively). However, it will be difficult to imagine how such an archetype might define the subject matter and determine the methodology in other sciences to a similar extent as in AI. On some accounts (AI-4 in particular), the guiding metaphor, the theory, the methodology and the subject matter all are known by the analogy of the mind as computer.

The concluding thought of Black's discussion of the importance of models as archetypes is that "Perhaps every science must start with metaphor and end with algebra; and perhaps without the metaphor there would never have been any algebra." (1962, 242) If this observation is to the point, and if the algebra in this context is embodied by AI's computational principles, the algebra will be the metaphor.

## References

Asaro, P. M. (2011). Computers as Models of the Mind: On Simulations, Brains, and the Design of Computers. In: *The Search for a Theory of Cognition. Early Mechanisms and New Ideas.* Ed. by S. Franchi and F. Bianchini. Amsterdam/New York: Rodopi, 89–114.

# *References*

Bensaude-Vincent, B. (2001). The construction of a discipline: Materials science in the United States. *Historical Studies in the Physical and Biological Sciences* 31.2, 223–248.

— (2013). Discipline Building in Synthetic Biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44, 122–129.

Black, M. (1962). *Models and Metaphors.* Ithaca: Cornell University Press.

Boden, M. A. (2006). *Mind as Machine: A History of Cognitive Science.* Oxford: Oxford University Press.

Boltzmann, L. (1902). Model. In: *Encyclopaedia Britannica.* Ed. by D. M. Wallace, A. T. Hadley, and H. Chisholm. 10th ed. Vol. 30. London: Adam and Charles Black, The Times, 788–791.

Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese* 195, 5339–5372. DOI: `https://doi.org/10.1007/s11229-018-01949-1`.

Burian, R. (1997). Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938-1952. *History and Philosophy of the Life Sciences* 19, 27–45.

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences* 36.3, 1–73. DOI: `10.1017/S0140525X12000477`.

Copeland, B. J. (2000). The Turing Test. *Minds and Machines* 10, 519–539.

— ed. (2004). *The Essential Turing.* Oxford: Oxford University Press.

Dreyfus, H. L. (1979). *What Computers Can't Do. A Critique of Artificial Reason.* New York/London: Harper & Row.

Feigenbaum, E. A. and J. Feldman, eds. (1963). *Computers and Thought.* New York: McGraw Hill.

Fisher, G. (2006). The Autonomy of Models and Explanation: Anomalous Molecular Rearrangements in Early Twentieth-Century Physical Organic Chemistry. *Studies in History and Philosophy of Science Part A* 37.4, 562–584.

van Fraassen, B. C. (1980). *The Scientific Image.* Oxford: Clarendon Press.

Galison, P. (2017). The Pyramid and the Ring. A Physics Indifferent to Ontology. In: *Research Objects in their Technological Setting.* Ed. by B. B. Vincent, S. Loeve, and A. Nordmann. History and Philosophy of Technoscience 11. London/New York: Routledge, 15–26.

Gelfert, A. (2016). *How to Do Science With Models: A Philosophical Primer.* Cham: Springer.

Guala, F. (2002). Models, Simulations, and Experiments. In: *Model-based Reasoning: Science, Technology, Values.* Ed. by L. Magnani and N. Nersessian. New York: Kluwer, 59–74.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge/London: Cambridge University Press.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica.* D 42, 335–346.

Hartmann, S. (1996). The World as Process: Simulations in the Natural and Social Sciences. In: *Simulation and Modelling in the Social Sciences from the Philosophy of Science Point of View.* Ed. by R. Hegselmann, U. Mueller, and K. G. Troitzsch. Dordrecht: Kluwer, 77–100.

Hegselmann, R. (2017). Thomas C. Schelling and James M. Sakoda: The Intellectual, Technical, and Social History of a Model. *Journal of Artificial Societies and Social Simulation* 20.3, 15. DOI: `10.18564/jasss.3511`.

Hertz, H. (1899). *The Principles of Mechanics. Presented in a New Form.* With an introduction by H. v. Helmholtz. Translated by D.E. Jones and J.T. Walley. London: Macmillan.

Hesse, M. B. (1966). *Models and Analogies in Science.* Notre Dame: University of Notre Dame Press.

Hodges, A. (1983). *Alan Turing: The Enigma.* New York: Simon & Schuster.

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.* Oxford: Oxford University Press.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep Learning. *Nature* 521, 436–444.

Lenhard, J. (2015). *Mit allem rechnen – zur Philosophie der Computersimulation.* Berlin/Boston: de Gruyter.

Marr, D. (1977). Artificial Intelligence – A Personal View. *Artificial Intelligence* 9, 37–48.

— (1982). *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information.* Reprint edition 2010. Cambridge: MIT Press.

McCarthy, J. (1960). Recursive Functions of Symbolic Expressions and Their Computation by Machine. *Communications of the ACM* 3.4, 184–195.

McCulloch, W. S. (1960). What Is a Number, that a Man May Know It, and a Man, that He May Know a Number? *General Semantics Bulletin* 26/27, 7–18.

McCulloch, W. S. and W. Pitts (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biology* 5.4, 115–133. ISSN: 0092-8240. DOI: 10.1007/BF02478259.

Moor, J. H. (1976). An Analysis of the Turing Test. *Philosophical Studies* 30, 249–257.

Morrison, M. (1999). Models as Autonomous Agents. In: *Models as Mediators. Perspectives on Natural and Social Science.* Ed. by M. S. Morgan and M. Morrison. Cambridge: Cambridge University Press, 38–65.

Nagel, E. (1961). *The Structure of Science.* New York: Harcourt, Brace & World. Chap. 12: Mechanistic Explanation and Organismic Biology.

Neumann, J. von (1945). *First Draft of a Report on the EDVAC.* draft report. Philadelphia: Moore School of Electrical Engineering, University of Pennsylvania.

Newell, A. (1980). Physical Symbol Systems. *Cognitive Science* 4, 135–183.

Newell, A., J. C. Shaw, and H. A. Simon (1963). Empirical Explorations with the Logic Theory Machine: A Case Study in Heuristics. In: *Computers and Thought.* Ed. by E. A. Feigenbaum and J. Feldman. New York: McGraw Hill, 109–133.

Newell, A. and H. A. Simon (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM* 19.3, 113–126.

Nordmann, A. (2011). Science in the Context of Technology. In: *Science in the Context of Application.* Ed. by M. Carrier and A. Nordmann. Dordrecht: Springer Netherlands, 467–482. DOI: 10.1007/978-90-481-9051-5_27.

Pylyshyn, Z. (1980). Computation and Cognition: Issues in the Foundations of Cognitive Science. *The Behavioral and Brain Sciences* 3, 111–169.

Ribe, N. and F. Steinle (2002). Exploratory Experimentation: Goethe, Land, and Color Theory. *Physics Today* 55.7, 43–49.

Ringle, M. (1979). Philosophy and Artificial Intelligence. In: *Philosophical Perspectives in Artificial Intelligence.* Ed. by M. Ringle. Atlantic Highlands: Humanities Press, 1–20.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review* 65.6, 386–408.

Rumelhart, D. and J. McClelland, eds. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* 2 vols. Cambridge/New York: MIT Press.

Russell, S. and P. Norvig (2010). *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River: Prentice Hall.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks* 61, 85–117. DOI: https://doi.org/10.1016/j.neunet.2014.09.003.

Soler, L., S. Zwart, M. Lynch, and V. Israel-Jost, eds. (2014). *Science after the Practice Turn in the Philosophy, History, and Social Studies of Science.* London: Routledge.

Sprevak, M. (2017). Turing's Model of the Mind. In: *The Turing Guide: Life, Work, Legacy.* Ed. by B. J. Copeland, J. Bowen, M. Sprevak, and R. Wilson. Oxford: Oxford University Press, 277–285.

Steinle, F. (1997). Entering New Fields: Exploratory Uses of Experimentation. *Philosophy of Science* 64, S65–74.

Suppes, P. (1960). A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences. *Synthese* 12, 287–301.

Tarski, A. (1953). A General Method in Proofs of Undecidability. In: *Undecidable Theories.* Ed. by A. Tarski, A. Mostowski, and R. M. Robinson. Amsterdam: North-Holland, 1–35.

Thagard, P. (2014). Cognitive Science. In: *Science after the Practice Turn in the Philosophy, History, and Social Studies of Science.* Ed. by L. Soler, S. Zwart, M. Lynch, and V. Israel-Jost. London: Routledge, 531–542.

Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* s2-42, 230–265.

— (1948). *Intelligent Machinery: A Report by A.M. Turing*. Tech. rep. London: National Physical Laboratory.

— (1950). Computing Machinery and Intelligence. *Mind* 59, 433–460.

Warwick, K. and H. Shah (2016). *Turing's Imitation Game. Conversations with the Unknown.* Cambridge: Cambridge University Press.

Waters, C. K. (2007). The Nature and Context of Exploratory Experimentation: an Introduction to Three Case Studies of Exploratory Research. *History and Philosophy of the Life Sciences* 29.3, 275–284.

Whitby, B. (1996). The Turing Test: AI's Biggest Blind Alley? In: *Machines and Thought.* Ed. by P. Millican and A. Clark. Vol. 1. The Legacy of Alan Turing. Oxford: Clarendon Press, 53–62.

Winsberg, E. B. (2010). *Science in the Age of Computer Simulation.* Chicago: University of Chicago Press.