# Towards a Philosophy of Science of the Artificial

Nora Ammann

January, 2023

One common way to start an essay on the philosophy of science is to ask: "Should we be scientific realists?" While this isn't precisely the question I'm interested in here, it *is* the entry point of this essay. So bear with me.

Scientific realism, in short, is the view that scientific theories are (approximately) true. Different philosophers have proposed different interpretations of "approximately true", e.g., as meaning that scientific theories "aim to give us literally true stories about what the world is like" (Van Fraassen, 1980, p. 9), or that "what makes them true or false is something external—that is to say, it is not (in general) our sense data, actual or potential, or the structure of our minds, or our language, etc." (Putnam, 1975, p. 69f), or that their terms refer (e.g. Boyd, 1983), or that they correspond to reality (e.g. Fine, 1986).

Much has been written about whether or not we should be scientific realists. A lot of this discussion has focused on the history of science as a source for empirical support for or against the conjecture of scientific realism. For example, one of the most commonly raised arguments in support of scientific realism is the so-called no miracle argument (Boyd, 1989): what can better explain the striking success of the scientific enterprise than that scientific theories are (approximately) true—that they are "latching on" to reality in some way or form. Conversely, an influential argument against scientific realism is the argument from pessimistic meta-induction (e.g. Laudan, 1981) which suggests that, given the fact that most past scientific theories have turned out to be false, we should expect our current theories to face the fate of being proven false (as opposed to approximately true).

In this essay, I consider a different angle on the discussion. Instead of discussing whether scientific realism can adequately explain what we empirically understand about the history of science, I ask whether scientific realism provides us with a satisfying account of the nature and functioning of the future-oriented scientific enterprise—or, what Herbert Simon (1996) called the sciences of the artificial. What I mean by this, in short, is that an epistemology of science needs to be able to account for the fact that the scientist's theorising affects what will come into existence, as well as for how this happens.

I proceed as follows. In Part 1, I explain what I mean by the sciences of the artificial, and motivate the premise of this essay—namely that we should aspire for our epistemology of science to provide an adequate account not only of the

inquiry into the "natural", but also into the nature and coming-into-being of the "artificial". In Part 2, I analyse whether or not scientific realism provides such an account. I conclude that its application to the sciences of the artificial world exposes scientific realism, while not as false, as insufficiently expressive. In Part 3, we briefly sketch how an alternative to scientific realism—pragmatism—might present a more satisfactory account. I conclude with a summary of the arguments raised and the key takeaways.

**Part 1: The Need for a Philosophy of Science of the Artificial**

The Sciences of the Artificial—a term coined by Herbert Simon in the eponymous book—refer to domains of scientific enterprise that deal not only with what is, but also with what might be. Examples include the many domains of engineering, medicine or architecture but also fields like psychology, economics or administration. What characterises all of these domains is their descriptive-normative dual nature. The study of what is is, both, informed and mediated by some normative ideal(s). Medicine wants to understand the functioning of the body in order to bring about and relative to a body's healthy functioning; civil engineering studies materials and applied mechanics in order to build functional and safe infrastructure. In either case, at the end of the day, the central subject of their study is not a matter of what is—it does not exist (yet)—; instead it is the goal of their study to bring it into existence. In Simon's words, the sciences of the artificial are "concerned not with the necessary but with the contingent—not with how things are but with how they might be—in short, with design" (Simon, 1996, p. xii).

Some might doubt that a veritable science of the artificial exists. After all, is science not quintessentially concerned with understanding what is—the laws and regularity of the natural world? However, Simon provides what I think is a convincing case that, not only is there a valid and coherent notion of a "science of the artificial", but also that one of its most interesting dimensions is precisely its epistemology. In the preface to the second edition of the book, he writes:

> "The contingency of artificial phenomena has always created doubts as to whether they fall properly within the compass of science. Sometimes these doubts refer to the goal-directed character of artificial systems and the consequent difficulty of disentangling prescription from description. This seems to me not to be the real difficulty. The genuine problem is to show how empirical propositions can be made at all about systems that, given different circumstances, might be quite other than they are." (Simon, 1996, p. xi)

As such, one of the things we want from our epistemology of science—a theory of the nature of scientific knowledge and the functioning of scientific inquiry—is to provide an adequate treatment of the science of the artificial. It ought, for example, to allow us to think not only about what is true but also what might be true, and how our own theorising affects what comes into being (i.e., what comes to be true). By means of metaphor, insofar as the natural sciences are typically concerned with the making of "maps", the sciences of the artificial are interested in what goes into the making of "blueprints". In the philosophy of science, we then get to ask: What is the relationship between maps and blueprints? On one hand, the quality of our maps (i.e., scientific understanding) shape what blueprints we are able to draw (i.e., the things we are able to build). At the same time, our blueprints also end up affecting our maps. As Simon puts it: "The world we live in today is much more a man-made, or artificial, world than it is a natural world. Almost every element in our environment shows evidence of human artifice." (Simon, 1996, p. 2).

One important aspect of the domain of the artificial is that there is usually more than one low-level implementation (and respective theoretical-technological paradigm) through which desired function can be achieved. For example, we have found several ways to travel long distances (e.g. by bicycle, by car, by train, by plane). What is more, there exist several different types of trains (e.g. coal-powered, steam-powered or electric trains; high-speed trains using specialised rolling stocks to reduce friction, or so-called maglev trains which use magnetic levitation). Most of the time, we need not concern ourselves with the low-level implementation because of their functional equivalency. Precisely because they are designed artefacts, we can expect them to depict largely similar high-level functional properties. If I want to travel from London to Paris, I generally don't have much reason to care what specific type of train I end up finding myself in.

However, differences in their respective low-level implementation can start to matter under the 'right' circumstances, i.e., given relevant variation in external environments. Simon provides us with useful language to talk about this. He writes: "An artifact can be thought of as a meeting point—an 'interface' in today's terms—between an 'inner' environment, [i.e.,] the substance and organization of the artifact itself, and an 'outer' environment, [i.e.,] the surroundings in which it operates. If the inner environment is appropriate to the outer environment, or vice versa, the artifact will serve its intended purpose." (Simon, 1996, p. 6). As such, the (proper) functioning of the artefact is cast in terms of the relationship between the inner environment (i.e., the artefact's structure or character, its implementation details) and the outer environment (i.e., the conditions in which the artefact operates). He further provides a simple example to clarify the point: "A bridge,

under its usual conditions of service, behaves simply as a relatively smooth level surface on which vehicles can move. Only when it has been overloaded do we learn the physical properties of the materials from which it is built." (Simon, 1996, p. 13).

The point is that two artefacts designed with the same purpose in mind, and which in a wide range of environments behave equivalently, will start to show different behaviours if we enter environments to which their design hasn't been fully adapted to. Now, their low-level implementation (or 'inner environments', in Simon's terminology) starts to matter.

To further illustrate the relevance of this point, let us consider the field of artificial intelligence (AI)—surely a prime example of a science of the artificial. The aspiration of the field of AI is to find ways to instantiate advanced intelligent behaviour in artificial substrates. As such, it can be understood in terms of its dual nature: it aims to both (descriptively) understand what intelligent behaviour is and how it functions, as well as how to (normatively) implement it. The dominant technological paradigm for building AGI at the moment is machine learning. However, nothing in principle precludes that other paradigms could (or will) be used for implementing AGI (e.g., some variation on symbolic AI, some cybernetic paradigm, soft robotics, or any number of paradigms that haven't been discovered yet).

Furthermore, different implementation paradigms for AGI imply different safety- or governance-relevant properties. Imagine, for example, an AGI built in the form of a "singleton" (i.e. a single, monolithic system) compared to one built as a multi-agent, distributed system assembly. A singleton AGI seems more likely to lack interpretability (i.e. behave in ways and for reasons that, by default, remain largely obscure to humans), while a distributed system might be more likely to fall prey to game-theoretic pitfalls such as collusion (e.g. Christiano, 2015). It is not at this point properly understood what the different implications of the different paradigms are, but the specifics of this must not matter for the argument I am trying to make here. The point is that, if the goal is to make sure that future AI systems will be safe and used to the benefit of humanity, it may matter a huge deal which of these paradigms is adopted, and to understand what different paradigms imply for considerations of safety and governability.

As such, the problem of *paradigm choice*—choices over which implementation roadmap to adopt and which theories to use to inform said roadmap—comes into focus. As philosophers of science, we must ask: What determines paradigm choice? And: How, if at all, can a scientist or scientific community navigate

questions of paradigm choice "from within" the history of science?

This is where our discussion of the appropriate epistemology for the sciences of the artificial properly begins. Next, let us evaluate whether we can find satisfying answers in scientific realism.

**Part 2: Scientific Realism of the Artificial**

Faced with the question of paradigm choice, one answer that a scientific realist might give is that what determines the right paradigm choices comes down entirely to how the world is. In other words, what the AI researcher does when trying to figure out how to build AGI is equivalent to uncovering the truth about what AGI, fundamentally, is. We can, of course, at a given point in time be uncertain about the 'true nature' of AGI, and thus be exploring different paradigms; but eventually, we will discover which of those paradigms turns out to be the correct one. In other words, the notion of paradigm choice is replaced with the notion of *paradigm change*. In essence, the aspiration of building AGI is rendered equivalent to the question of what AGI, fundamentally, is.

As I will argue in what follows, I consider this answer to be dissatisfying in that it denies the very premise of the science of the artificial we have discussed in the earlier section. Consider the following three arguments.

First, the answer by the scientific realist seems to be fundamentally confused about the type-signature of the concept of "AGI". AGI, in the sense I've proposed here, is best understood as a *functional description*—a design requirement or aspiration. As discussed earlier, it is entirely plausible that there exist several micro-level implementations which are functionally-equivalent to (i.e. depict) generally intelligent behaviour. As such, by treating "the aspiration of building AGI [as equivalent] to the question of what AGI is", the scientific realist has implicitly already departed from—and thus failed to properly engage with—the premise of the question.

Second, note that there are different vantage points from where we could be asking the question. We could take the vantage point of a "forecaster" and ask what artefacts we should expect to exist 100 years from now. Or, we could take the vantage point of a "designer" and ask which artefacts we want to create (or ought to create, given some set of moral, political, aesthetic, or other commitments). While it naturally assumes the vantage point of the forecaster, scientific realism appears inadequate for taking seriously the vantage point of the designer.

Third, let's start with the assumption that what will come into existence is a matter of fact. While plausible-sounding at first, further inspection reveals prob-

lems with this claim. To show this, let us consider the tri-partite characterisation proposed by Drexler (2018). We want to distinguish between three notions of "possible", namely: (physically) realistic, (techno-economically) plausible, and (socio-politically) credible. Beyond such facts as the fundamental laws of physics (i.e., the physically realistic), there are other factors—less totalising and yet endowed with some degree of causal force—which shape what comes to be in the future (e.g., economic and sociological pressures).

Importantly, the physically realistic does not on its own determine what sorts of artefacts come into existence. For example, paradigm A (by mere chance, or for reasons of historical contingency) receives differential economic investment compared to paradigm B, resulting in its faster maturation; or inversely, it might get restrained or banned through political means, resulting in it being blocked, and eventually forgotten. Examples of political decisions (e.g. regulation, subvention, taxation) affecting technological trajectories abound. To name just one, consider how the ban on human cloning has, in fact, stopped human cloning activities, as well as any innovations related to making human cloning 'better' in some way.

The scientific realist might react to this by arguing that, while the physically realistic is not the only factor that determines what sorts of artefacts come into existence, there is still a matter of fact to the nature and force of economic, political and social factors affecting technological trajectories, all of which could, at least in principle, be understood scientifically. While I am happy to grant this view, I argue that the problem lies elsewhere. As we have seen, the interactions between the physically realistic, the techno-economically plausible, and the socio-politically credible are highly complex and, importantly, *self-referential*. It is exactly this self-referentiality that makes this a case of paradigm *choice*, rather than paradigm *change*, when viewed from the position of the scientist. An adequate answer to the problem of paradigm choice must necessarily consider a "view from inside of the history of science", as opposed to a "view from nowhere". After all, the paradigm is being chosen by the scientific community (and the researchers making up that community), and they are making said choice from their own situated perspective.

In summary, it is less that the answers provided by the scientific realist are outright wrong. It rather appears as if the view provided by scientific realism is not expressive enough to deal with the realities of the sciences of the artificial. It cannot usefully guide the scientific enterprise when it comes to the considerations brought to light by the sciences of the artificial. Philosophy of science needs to do better if it wants to avoid confirming the accusation raised by Richard Feynman—that philosophers of science are to scientists what ornithologists are to birds; namely irrelevant.

Next, we consider whether a different epistemological framework—while holding onto as much realism as possible—appears more adequate for the needs of the sciences of the artificial.

**Part 3: A Pragmatic Account of the Artificial**

So far, we have introduced the notion of the science of the artificial, discussed what it demands from the philosophy of science, and observed how scientific realism fails to appropriately respond to those demands. The question is then: Can we do better?

An alternative account to scientific realism—and the one we will consider in this last section—is pragmatic realism, chiefly originating from the American pragmatists William James, Charles Sanders Peirce, and John Dewey. For the present discussion, I will largely draw on contemporary work trying to revive a pragmatic philosophy of science that is truly able to guide and support scientific inquiry, such as by Robert Toretti, Hasok Chang, and Rein Vihalemm.

Such a pragmatist philosophy of science emphasises scientific research as a practical activity, and the role of an epistemology of science as helping to successfully conduct this activity. While sharing with the scientific realist a commitment to an external reality, pragmatism suggests that our ways of getting to know the world are necessarily mediated by the ways knowledge is created and used, i.e., by our epistemic aims and means of "perception"—both the mind and scientific tools, as well as our scientific paradigms.

Note that pragmatism, as I have presented it here, does at no point do away with the notion of an external reality. As Giere (2006) clarifies, not all types of realism must subscribe to a "full-blown objective realism" (or what Putnam called "metaphysical realism")—roughly speaking the view that "[t]here is exactly one true and complete description of 'the way the world is'" (Putnam, 1981, p. 49). Pragmatic realism, while rejecting objective or metaphysical realism, remains squarely committed to realism, and understands scientific inquiry as an activity directed at better understanding reality (Chang, 2022, pp. 5, 208).

Let us now consider whether pragmatism is better able to deal with the epistemological demands of the sciences of the artificial than scientific realism. Rather than providing a full-fledged account of how the sciences of the artificial can be theorised within the framework of pragmatic realism, what I set out to do here is more modest in its ambition. Namely, I aim to support my claim that scientific realism is insufficiently expressive as an epistemology of the sciences of the artificial by showcasing that there exist alternative frameworks—in this case pragmatic

realism—that do not face the same limitations. In other words, I aim to show that, indeed, we can do better.

First, as we have seen, scientific realism fails to adopt the "viewpoint of the scientist". As a result, it collapses the question of paradigm choice to a question of paradigm change. This makes scientific realism incapable of addressing the (very real) challenge faced by the scientist; after all, as I have argued, different paradigms might come with different properties we care about (such as when they concern questions of safety or governance). In contrast to scientific realism, pragmatism explicitly rejects the idea that scientific inquiry can ever adopt a "view from nowhere" (or, a "God's eye view" as Putnam, 1981, p. 49 puts it). Chang (2019, p. 10) emphasises the "humanistic impulse" in pragmatism:

> "Humanism in relation to science is a commitment to understand and promote science as something that human agents do, not as a body of knowledge that comes from accessing information about nature that exists completely apart from ourselves and our investigations."

This aligns well with the need of the sciences of the artificial to be able to reason from the point of view of the scientist.

Second, pragmatism, in virtue of focusing on the means through which scientific knowledge is created, recognises the *historicity* of scientific activity (see also Vihalemm, 2012; Chang, 2019). This capacity allows pragmatic realism to reflect the historicity that is also present in the science of the artificial. Recall that one central epistemological question of the sciences of the artificial concerns how our theorising affects what comes into existence. As such, our prior beliefs, scientific frameworks and tools affect, by means of 'differential investment' in designing artefacts under a given paradigm, what sort of reality comes to be. Moreover, the nature of technological progress itself affects what we become able to understand, discover and build in the future. Pragmatism suggests that, rather than there being already a predetermined answer as to which will be the most successful paradigm, the scientist must understand their own scientific activity as part of an iterative and path-dependent epistemic process.

Lastly, consider how the sciences of the artificial entail a 'strange inversion' of 'functional' and 'mechanistic' explanations. In the domain of the natural, the 'function' of a system is understood as a viable post-hoc description of the system, resulting from its continuous adaptation to the environment by external pressures. In contrast, in design, the 'function' of an artefact becomes that which is antecedent, while the internal environment of the artefact—its low-level implementation—becomes post-hoc. It appears difficult, through the eyes of a scientific realist, to

fully accept this inversion. At the same time, accepting it appears to be useful, if not required, in order to develop an epistemology of the sciences of the artificial on its own terms.

Pragmatic realism does not face the same trouble. To exemplify this, let us take Chang's notion of *operational coherence*—a deeply pragmatist yardstick for scientific inquiry—which he describes as "a harmonious fitting-together of actions that is conducive to a successful achievement of one's aims" (Chang, 2019, p. 14). Insofar as we are able to argue that a given practice in the sciences of the artificial possesses such operational coherence, it is compatible with pragmatic realism. What I have tried to show hereby is that the sciences of the artificial, including the 'strange inversion' of the role of 'functions' which it entails, is fully theorisable inside the framework of pragmatic realism. As such, unlike scientific realism, the latter does not fail to engage with the sciences of the artificial on its own terms.

To summarise this section, I have argued, by means of three examples, that pragmatic realism is a promising candidate for a philosophy of science within which it is possible to theorise the sciences of the artificial. In particular, I have invoked the fact that the sciences of the artificial requires us to take the "point of view of the scientist", to acknowledge the iterative, path-dependent and self-referential nature of scientific inquiry (i.e., its historicity), and, finally, to accept the central role of 'function' in understanding designed artefacts.

**Conclusion**

In Part 1, I laid out the case for why we need a philosophy of science that can encompass questions arising from the sciences of the artificial. One central such question is the problem of paradigm choice, which requires the scientific practitioner to understand the ways in which their own theorising affects what will come into existence.

In Part 2, I considered whether scientific realism provides a sufficient account, and concluded that it doesn't. I listed three examples of ways in which scientific realism seems to be insufficiently expressive as an epistemology of the sciences of the artificial. Finally, in Part 3, I explored whether we can do better, and provided three examples of epistemic puzzles, arising from the sciences of the artificial, that pragmatic realism—in contrast with scientific realism—is able to account for.

While scientific realism seems attractive on the basis of its explaining the success of the natural sciences, it does not in fact present a good explanation of the success of the science of the artificial. How, before things like, say, planes, computers, or democratic institutions existed, could we have learnt to build them if all that

was involved in the scientific enterprise was uncovering that which (already) is? As such, I claim that the sciences of the artificial provide an important reason for why we should not be satisfied with the epistemological framework provided by scientific realism with respect to understanding and—importantly—guiding scientific inquiry.

---

**References**

Boyd, R. N. (1983). On the current status of the issue of scientific realism. *Methodology, Epistemology, and Philosophy of Science: Essays in Honour of Wolfgang Stegmüller*, 45–90.

Chang, H. (2019). Pragmatism, perspectivism, and the historicity of science. In *Understanding Perspectivism* (pp. 10–27). Routledge.

Chang, H. (2022). *Realism for Realistic People*. Cambridge University Press.

Christiano, P. (2015). On heterogeneous objectives. *AI Alignment Forum*. `https://ai-alignment.com/on-heterogeneous-objectives-b38d0e003399`

Drexler, E. (2018). Paretotopian goal alignment. Talk at EA Global: London 2018.

Drexler, E. (2019). *Reframing Superintelligence*. Future of Humanity Institute.

Fine, A. (1986). Unnatural attitudes: Realist and antirealist attachments to science. *Mind*, *95*(378), 149–177.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, *48*(1), 19–49.

Putnam, H. (1975). *Mathematics, Matter and Method*. Cambridge University Press.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge University Press.

Simon, H. (1996). *The Sciences of the Artificial* (3rd ed.). MIT Press.

Soares, N., & Fallenstein, B. (2017). Agent foundations for aligning machine intelligence with human interests: A technical research agenda. *The Technological Singularity: Managing the Journey*, 103–125.

Toretti, R. (2000). 'Scientific realism' and scientific practice. In E. Agazzi & M. Pauri (Eds.), *The Reality of the Unobservable*. Kluwer.

Van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press.

Vihalemm, R. (2012). Practical realism: Against standard scientific realism and anti-realism. *Studia Philosophica Estonica*, *5*(2), 7–22.