Can AI Make Scientific Discoveries?

Marianna B Ganapini

penultimate draft – accepted in Philosophical Studies. Please do not cite

AI technologies have recently shown remarkable capabilities in various scientific fields, such as drug discovery, medicine, climate modeling, and archaeology, primarily through their pattern recognition abilities. They can also generate hypotheses and suggest new research directions. While acknowledging AI's potential to aid in scientific breakthroughs, the paper shows that current AI models do not meet the criteria for making independent scientific discoveries. Discovery is an epistemic achievement that requires a level of competence and self-reflectivity that AI does not yet possess.

Over the past few years, artificial intelligence (AI) technologies (machine learning tools, in particular) have demonstrated their capabilities in various scientific fields, including but not limited to drug discovery, medicine, climate modeling and archaeology. This is usually achieved by AI's astonishing ability to find patterns in the data. AI can also be a valuable tool in hypothesis generation, suggesting new research directions to scientists. And AI can find new scientific facts based on some background theory (Cockburn et al., 2018; Miao 2023; Wang 2023). In particular, automated scientific discovery is the field that – bringing together artificial intelligence, natural science and philosophy – studies the remarkable impact of AI in scientific discovery and how the process of making scientific discoveries can be modeled computationally (Giza 2021). Notably, Newell and Simon (1956) have argued that computer programs can surely go through the mechanism of problem solving and arrive at making scientific discoveries.

Though AI has the potential to make significant contributions to scientific discoveries, in this paper, I raised some questions surrounding the role of AI in scientific research. In particular, I raise some challenges to the claim that AI can, in fact *itself* discover anything. I will not deny that AI can *find* new things and do so reliably. I don't deny for sure that AI can help us achieve scientific breakthroughs (Duede 2023). However, I will argue that *as of now* it does not look like AI models can make scientific discoveries. My claim will be based on the observation that making scientific discovery first requires having the right type of competence. Furthermore, discovery is a kind of epistemic achievement, which necessitates having some assessment on and exercising some control over one's abilities. It is exceedingly difficult for us to determine whether AI has the right competence and it does not look like AI for now is able to assess and control its epistemic activities.

Three quick clarifications before I start. I am not directly interested here in exploring the question: what is a scientific discovery? I am more focuses instead on the question of when an agent (of some sort) counts as a 'scientific discoverer'.[1] Furthermore, I have very little to say here on whether AI has a mind, real preferences or consciousness. In contrast, here I examine *some* of the key features that AI needs to possess to qualify as

[1] This approach, that focuses on an individual as the source of discovery, has been criticized because it fails to recognize the social, distributed element of science (Darden 2006; Copeland 2018: 695; Clark & Khosrowi 2022). However, here I am trying to ask about the role that AI could have in science, beyond just being a tool or aid for scientists. This does not exclude that, in practice, scientific discoveries are a communal affair. Also, the question on whether AI can be an agent of discovery, will be kept separate from the broader ethical and philosophical considerations about agency, autonomy and personhood. In this paper agency in the context of discovery can be defined narrowly, focusing on the functional and causal role that AI plays in generating new knowledge.

the author of scientific breakthroughs. In so far that these may require having a mind, thoughts and reasoning abilities, I will assume (for the sake of the argument) that AI does have those. Finally, I will focus my discussion to the use of deep neural networks, which have recently made impressive scientific findings. I cannot hope to make the case for all of AI in one paper, and deep neural networks seem the most promising AI architecture as of late.

1. Scientific Discoveries

In this paper, I won't tackle the issue of scientific discoveries in general (Michel 2019, 420; 2020) but focus on some of the conditions that allow us to make discoveries in science. I will argue that one of these conditions is that the person discovering needs to be competent in the field of discovery. The second related condition is that – because making a scientific discovery is an epistemic achievement – the 'agent' discovering needs to be someone able to assess her own competence and integrate that information into her reasoning. I will tackle these conditions in turn and then discuss whether AI possesses them.

1.1. Competence

I will start by stating something I believe should be quite uncontroversial: making scientific discoveries is the result of applying one's skills and competence on the subject of discovery. To make this more vivid I will provide a comparison that will hopefully illuminate this point. One day, during the Renaissance period in Rome, a young boy

tripped and fell down a hole. The result was unexpected: he found the Domus Aurea, Nero's long forgotten majestic palace. This child stumbled upon a great piece of archeological discovery but hopefully the reader will agree with me that he did not *discover* the Domus Aurea in the sense he did not make a scientific discovery. This archaeological finding happened by chance and the child had no clue about what he found. Interestingly enough, nobody at that time understood that those ancient ruins actually a Roman building: the entire palace was under ground and covered with dust and dirt and it was kind of hard to figure out what it really was.

Here is a different situation. Again Renaissance: Christopher Columbus left Europe and arrived to America. He mistook it for something else and did not realize he found a new (for the Europeans) continent.[2] There is a lot of controversy on whether Columbus actually *discovered* America, but we can try to set those worries aside for now. My goal here is to prompt the intuition that there is something about his finding that makes it a *possible* candidate to be a scientific discovery (whether or not it is in the end deemed to be an *actual* scientific discovery). In neither one of the two cases (Columbus and the boy), the "discoverer" was actually looking for what they found: their "discovery" was by chance. And, apparently, they never realized the nature of what they found either. And

[2] When Columbus arrived in 1492, the continents of North and South America were already inhabited by indigenous peoples with distinct cultures and civilizations. Those lands were unknown to contemporary Europeans, but the Native peoples had inhabited them for thousands of years. Hence, I acknowledge that saying that Columbus "discovered" America potentially disregards the history and existence of those peoples. And yet looking at it through the lens of the worldview of the late 15th century, Columbus's voyages may well represent a groundbreaking geographical "discovery" in that context, aka the context where science was made.

yet, though the two cases are remarkably similar, we can say that, in the case of Columbus, it sounds more plausible to say that he made an actual scientific discovery whereas the boy merely found something new which *then* acquired scientific recognition (Whewell 1996 [1840]: 189)[3]. There is *something* that is at least closer to a scientific discovery in what Columbus did, whereas I think we can safely say the child is really just stumbling upon an important archeological finding by chance.

If you agree with my intuition that Columbus was a little closer to a scientific discovery than the boy, you might also agree that what makes the difference here is that, in the case of Columbus, the discovery was the result of a process that revealed competence in the field of the discovery. The boy knew nothing about Roman buildings or ancient buildings. He knew nothing about archaeology. In contrast, although Columbus had gaps in his knowledge and understanding, he did have some competence in geography, navigation and so on. There seems to be a role played by knowledge and competence in the case of Columbus that allows us to qualify his finding as a possible scientific discovery.[4]

It is time to dig a little deeper on the notion of competence at play here. In particular, there seem to be two ways in which Columbus' discovery revealed or was the result of his competence. For starters, he got to America not completely by chance but by following a reasoning process that revealed understanding of navigation and geography,

[3] Whewell was skeptical there was such a thing as accidentally-made scientific discovery but I will set his worries aside here.

[4] Along similar lines, Clark & Khosrowi (2022) point out how the discovering agent is meant to have "particular qualities/abilities which play a significant role in the discovery."

the kind of skills needed to discover new lands. Second, he would have known he made a new discovery had some background information been different. Had he, for instance, understood that the circumference of the Earth was bigger than he thought, he would have recognized that where he landed could not be the Indies. In this case, he would not only have found a new land but also become aware that he had it. He set out to reach the Indies, but failed at that. Nonetheless, his performance, his actions, and his discovery are partly the result of his competence (and not any competence, but competence in the area that was relevant to the discovery of a new land).

We should try to generalize the conditions mentioned above:

C1) For subject (or group) S' finding[5] F of object Y to count as a scientific discovery of Y, F should be linked to or express S' competence concerning the field of discovery where Y belongs.

It looks like C1 is a necessary condition for scientific discovery. It says that for any scientific discovery of Y, S' finding Y should be linked to their competence in the field of discovery.

Cognitive scientists, linguists and philosophers have been talking about competence and expertise for a while (Ericsson 2018; Watson 2021). For simplicity I will adopt a rough account of competence which should be enough for our purposes: an agent who is scientifically competent concerning Y is someone who has mastered what is taken to

[5] This need not be a first time finding: it is possible that Y was already known but not known *as a new scientific entity*.

be true concerning Y, within the scientific community at that particular period in history.[6] I am aware that this formulation substitutes a mystery (i.e., competence) with another one (i.e., mastery). However, the scope of the paper is not to determine what kind of competence is needed in scientific discoveries, but *how* this competence is reasonably expressed and recognized in discoveries. And there are various possible ways in which a finding can *express* the subject's relevant competence.

One way is to stress the reasoning process that leads to discoveries:

C1-r) F's expresses S' competence concerning the type of object Y if F is at least in part the result of a reasoning process common in the field in which discoveries of Y would take place (at the time in which it took place)

Let me tackle an initial worry. The idea is that if one is competent in a field of inquiry, one should be able to make inferences and those inferences would at least partially cause one to arrive at a particular discovery.[7] Here is an example. In 1928 Dr. Alexander Fleming – after a few days away from his lab –noticed that a petri dish containing Staphylococcus bacteria was overtaken by mold. To his utter surprise, he also observed

[6] Ptolemy was considered an expert astronomer of his time. However, if he were alive today, many of his beliefs and theories would be regarded as false based on our current scientific understanding. This underscores the point that being an expert or being competent does not necessarily require having mostly accurate beliefs or knowledge. Rather, an expert or a competent person is someone who has mastered what is taken to be true within the scientific community at that particular period in history.

[7] This does not exclude that a discovery could also be the result of an eureka moment in which intuition and creativity allow one to envision some new, more successful explanation of some scientific phenomena.

that the mold seemed to inhibit the growth of the Staphylococcus nearby. Further investigation revealed that the mold produced a potent chemical with antibacterial properties. This substance was named "penicillin" by Dr. Fleming. Though the significance of the breakthrough was not understood by the scientific community at first, it eventually was recognized as a world changing discovery. Though one may say that Fleming's discovery was serendipitous, it was also the result of a pattern of thinking that was grounded in the best biology and chemistry of his time. Hence, his finding expressed his competence in biology and chemistry, as one could trace a partial causal path from his expert reasoning to the discovery itself. His stumbling upon a new scientific discovery was fortuitous but not completely by chance.

There is another crucial aspect to consider. Even if we insist that Fleming's discovery was the result of chance, he still had the knowledge and competence to understand the importance of what he found. That suggests that another possible way to understand competence in discovery is based on the idea of explanation:

C1-k) F's expresses S' competence concerning the type of object (they found) Y if, given some key background information, S would be able to explain the scientific relevance (and novelty) of Y.

Condition (C1-k) does not require any direct causal relation between one's competence concerning Y and the act of finding Y. An example will illuminate this point further. Apparently, Gibbons discovered the Fiji crested iguana completely by chance during a screening of the 1980 film *The Blue Lagoon*, when he spotted a new type of lizard and traveled to Nanuya Levu, Fiji, to identify the animal (Gibbons 1981). When he spotted

the new Iguana, Gibbons was not looking for a new species or even doing any science, but simply watching a movie. However, his knowledge and expertise allowed him to make the discovery: his unorthodox 'finding' became a discovery when he was able to explain why the new iguana was a new species.

One way to make sense of (C1-k) and how it differs from (C1-r), is to relate them to the distinction, common in philosophy of science, between the contexts of "justification" and "discovery" (Reichenbach 1938, Popper 2002). The context of discovery is the thinking process that allows to make progress. Some have argued that such a thinking may not even be rational (Strevens 2020). In contrast, the context of justification is where rationality gets employed: it is the phase where a theory gets assessed, the discovery of a new scientific object gets explained and so on.[8] (C1-k) expresses the idea that, no matter how the process of discovery was achieved, any scientific discovery needs to be recognized, justified and understood. An agent of discovery needs to be able to explain and understand the scientific relevance of what they did and found: they employ their own competence to shed light on the novelty and importance of the findings.

Let me conclude this section by saying that it is reasonable to assume that these two conditions (C1-r and C1-k) are each sufficient for showing competence, though they might not be strictly necessary. They may also *not* be the only possible sufficient conditions for showing competence (both in general and in scientific discovery). However, they seem to be quite central to the process of scientific discovery. Again, take the boy who found the Domus Aurea: he may have been competent in many things, but his finding did not reveal any kind of mastery of archeology. More specifically, he did

[8] Admittedly I may be culpable of what Darden (2006) calls a "simplistic dichotomy" between these two contexts which is common in philosophy of science.

not find the Domus Aurea as a result of a reasoning process typical of the field of archeology. In addition, no (reasonable) amount of background condition would have presumably let him recognize that what he found was a Roman palace. This indicates that he probably did not *have* the right competence to scientifically discover what he in fact found. Hopefully, this example is enough to motivate the intuition that failing to conform to (C1-r and C1-k) is a reasonable indicator that one is not an agent of discovery in science.

2. Scientific Discoveries as Epistemic Achievements

Scientific discoveries are praiseworthy endeavors: we consider them an epistemic achievement, possibly the pinnacle of any scientific activity. Similarly, those who make discoveries are usually praised, and their achievements may be honored with fame. Assigning blame or praise requires assigning a degree of responsibility: the agent subject to epistemic praise for making a discovery, needs to be the type of agent that can exercise a degree of epistemic control over the ways they came to make and/or understand the discovery. This ability is something above and beyond the agent having competence: it is the ability to have *insights about one's own competence*.

To understand this distinction, it is worth looking at some of the work done in value epistemology. Ernest Sosa (2007; 2011) famously draws the distinction between performances that are apt and performances that are meta-apt. A performance that is apt is a performance that achieves as a result of exercising competence. A performance that is meta-apt has something more: it reveals the competent assessment that the performance would be apt. Such an assessment reveals that the subject has good insights into their own competence and on the conditions in which such a competence

can and cannot be successful. To adopt Sosa's famous analogy, an "apt" performance is when an archer successfully hits the target by *competently* exercising their archery skills. It's a skilled performance that achieves the intended goal. However, a "meta-apt" performance has an additional layer. It's when the archer not only hits the target competently, but also has insights that the shot would be successful given the specific conditions at play (e.g., distance, wind, her abilities). This meta-aptness demonstrates the archer has reflective knowledge about the limits of her competence and the kinds of conditions where her skills can reliably achieve the intended result versus not: she does not just unreflectively fire accurate shots, but understand why they are successful.[9]

Similarly, the key idea is that for a discovery to be considered a true epistemic achievement worthy of praise, the discoverer needs to not only possess the requisite competencies, but also have insights into the scope and limits of their own competencies.[10] In other words, it's not enough for an agent to simply perform well by

[9] Sosa argues that an apt true belief amount to knowledge: "its correctness derives from manifesting certain *cognitive virtues* of the subject, where nothing is a cognitive virtue unless it is a truth-conducive disposition" (2009, p. 135. My emphasis). On top of knowledge that p, *reflective* knowledge that p requires also "that under the light of reflection one must be able to defend the reliability of one's sources" for the belief that p (2009, p. 139).

[10] With Sosa, Greco (2010, p.3) claims that knowledge, an apt-performance, is already an epistemic achievement. I do not disagree. However, lack of control over one's performance, namely lack of reflective insight, prevents us from praising the agent *herself*. As Audi puts it, "(as Sosa realizes) knowledge in general cannot be considered to be true belief grounded in virtue – unless perhaps we distinguish what might be called animal virtue, which would be a kind of epistemic *power*, from reflective *virtue*, which would be a trait for which one merits a measure of praise" (2004, p.8). Also, one could employ Riggs' distinction between credit (as attributability) vs. praise. Apt performances should be credited to agents, but they are not praiseworthy (Riggs 2009).

exercising their capabilities in a competent way (an "apt" performance). To be epistemically praiseworthy as the source of a significant discovery, the agent must also demonstrate a reflective grasp of the conditions under which their competencies can successfully achieve the intended goal (a "meta-apt" performance). Without this reflective dimension, the agent cannot be fully praised as the true epistemic source of the discovery because they cannot quite exercise control over the process.[11]

Let me be clear that this requirement does *not* exclude serendipitous discoveries. As I mentioned above, some discoveries are made by chance. When that is the case, however, we expect the agent of discovery to still be able to direct her activities based on her understanding of her competence. Either in the context of discovery or in the context of justification, we would need to see her reflective insights at work. Thus, even if scientific discoveries are frequently recognized and validated post hoc, and agent's understanding of their competence in an area of inquiry should direct her analysis and her justificatory procedures in a way that is worth of praise.[12]

---

[11] In the case of scientific discoveries, they are praiseworthy *qua* exercise of epistemic abilities (and epistemic agency) even though, all things considered, they may not be praiseworthy. For instance, discovering the atomic bomb may have not have been morally praiseworthy, though from a scientific and epistemic standpoint, it surely was worthy of praise.

[12] Scientific discoveries often occur in a context where understanding and competence are distributed across a community. The notion that an individual must fully understand their competence at every step might be too restrictive, especially considering the collaborative nature of modern scientific research. However, here we are trying to understand the role of one agent in this process and whether that agent qualifies as a discoverer at all. I take it that an agent who is blind to what she knows and how she reasons does not qualify as an agent of discoveries.

Hence, some level of insights about the agents' competence seems to be a key condition for making a discovery:

C2) For subject (or group) S' finding F of object Y to count as a scientific discovery of Y, S should have reflective insights of their own competence in the field in which discoveries of Y are made[13]

The reference to "insights" is meant to be broad enough to encompass different types states. Having reflective insight does not require awareness or (phenomenological) consciousness. Having reflective insight of one's own competence may mean having metacognition and metaknowledge, or simply being able to produce assessments about or signaling one's competence in a reliable way. And metaknowledge need not require any meta-representation either: for instance, some non-human animals have a sense for what they know without having to represent their mental states (Proust 2010). Similarly, (C-2) does not require one is able to go through a deliberate, fully articulated, explicit assessment of one own's competence: intuitive insights on one's own abilities given the context of performance should be enough to satisfy (C-2).

3. AI and Scientific Discoveries

[13] How reliable need those insights to be? What if a scientist is wrong about what they know and don't know? As mentioned above, Ptolemy probably thought he knew a lot of things about planets and stars, though now we can say that some of his beliefs on the subject were wrong. Recall that competence is contextually situated, though: for his times, Ptolemy was competent in astronomy and his assessments about his own competence were correct.

In this section of the paper, I will discuss whether AI can be an agent of discovery by looking at whether AI matches the conditions mentioned above (C-1 & C-2). Two decades ago, symbolic AI enjoyed widespread popularity, yet AI research has since moved to a different paradigm: machine learning. And within the domain of machine learning, neural networks have emerged as prominent in the last ten to fifteen years. Among neural networks, those characterized by numerous layers, commonly referred to as deep learning architectures, have produced some fantastic results (Zhavoronkov et al 2019). In the context of this paper, the term "AI" primarily denotes deep learning methodologies, as they currently represent the principal cause for excitement surrounding AI's applications in scientific research.[14] One groundbreaking artificial intelligence system is AlphaFold, a deep learning model developed by Google's DeepMind (Abramson at al 2024). This system has demonstrated a remarkable capability to accurately predict the complex three-dimensional structures of proteins at a level of precision that has not been achieved before. AlphaFold could unlock new frontiers in fields like drug development, disease research, and our fundamental understanding of biological processes (Jumper at al 2021). Artificial intelligence techniques can also be employed in the field of archaeology for predictive modeling as well as automated object recognition and detection. Remote sensing data from satellite platforms can be utilized to survey vast geographic areas. Advanced computer vision algorithms trained on archaeological examples appear capable of identifying subtle patterns and features indicative of potential undiscovered archaeological sites. This raises the possibility that

[14] Deep learning models with a lot of layers are usually opaque systems: a 'black box' whose outputs are nearly impossible to explain based on the system's innerworkings (Pasquale, 2015). I will come back to this point.

artificial intelligence systems could lead to the discovery of previously unknown ancient settlements or urban centers, such as a Roman city, by leveraging object recognition approaches to analyze remotely sensed data (Bickler 2021).[15]

The question remains as to whether – beside helping humans make discoveries – AI can itself actually *make* a scientific discovery. If my analysis above is accurate, for AI to be capable of making discoveries, it must demonstrate capabilities similar to those of C-1 and C-2. It cannot simply function as a reliable finder; it must be a finder that exercises its competence and reflects on its abilities.

Let's start with the first one:

C-1) For AI's finding F of object Y to count as a scientific discovery of Y, F should express AI's competence concerning the field of discovery where Y belongs.

This is one way in which this could happen:

> C1-r) F's expresses AI's competence concerning the type of object Y if F is at least in part the result of a reasoning process common in the field in which discoveries of Y would take place (at the time in which it took place)

[15] A similar application in archeology is the ArchAIDE app that leverages AI techniques and automatically recognizes archaeological ceramics from a single photographic image (Anichini et al. 2021).

Let me tackle an initial worry. C1-r) suggests that the AI's reasoning should at least partially mirror or incorporate the typical modes of reasoning, methodologies, and inferential patterns that scientists employ when investigating scientific matters. This may seem a tall order for AI. It is true that many current machine learning techniques, like deep neural networks, operate very differently from human cognition by finding patterns in large datasets using techniques like gradient descent and backpropagation. However, in principle they could mirror human thinking: machine learning algorithms inductively learn generalizable patterns from data samples, similar to how scientists inductively reason from observations to theories. Therefore, while replicating human-level scientific reasoning is certainly a challenge for AI, some existing AI techniques already share important commonalities with our methods of inquiry.[16] We will discuss this point further below, for now I would like to side step this worry.

Another possible way to understand competence in discovery is in terms of explanation (and, possibly, understanding):

> C1-k) F's expresses AI's competence concerning the type of object (they found) Y if, given some key background information, AI would be able to explain the scientific relevance of Y.[17]

[16] It is worth noticing here that symbolic AI techniques can explicitly represent hypotheses, axioms, rules, thus closely mirroring how we reason in science. Bayesian Networks explicitly represent causal relationships using probabilistic graphical models, allowing reasoning about evidence and conclusions.

[17] To satisfy this AI would need to be able to know, understand and so on. If we consider the philosophical approach to knowledge as justified, true belief, the question some ask is

An agent of scientific discovery is usually someone who is understandable to a scientific community so that their discoveries can be framed in terms of concepts, laws, models and current theories that explain their scientific relevance. Similarly, to be an agent of discovery, AI needs to be able to contribute to an understanding of the factors underlying complex scientific phenomena. That means that the AI should provide insight into the causal mechanisms and factors underlying its discoveries – not just showing correlations, but illuminating the explanatory "why" behind the relevance of complex scientific phenomena (Hempel 1965). Relatedly, AI should transmit its findings in a way that at least in part maps to existing scientific concepts, models, and theories that the community is already working with. Arguably, part of what makes a discovery "scientific" is this ability to cohere it with our systematic, accumulating knowledge base and the shared epistemic norms and practices of the scientific method. If the discoveries cannot be properly framed within the scientific method, their scientific status is in question.[18]

Finally, as for C-2, this is how it could apply to AI:

C-2) For AI's finding F of object Y to count as a scientific discovery of Y, AI should have reflective insights of its own competence in the field in which discoveries of Y are made

whether current AI can have *beliefs* at all. Though there is general skepticism around the idea that AI believes anything, I'll put these worries aside here.

[18] Famously Kuhn talks about 'accumulation of anomalies' that lead to paradigm shifts: anomalies and failures to fit into the current paradigm can still fit into the intellectual endeavor of science.

As experts in fields like archaeology need to be able to assess their own competence, evaluate the validity of their answers based on that competence, gauge their confidence, and identify ways to address gaps in their understanding, an AI system should ideally possess similar capabilities. This condition requires AI to have reflective insights of its own abilities. As previously said, this condition does not necessitate consciousness or fully developed metacognitive representations. Whatever insights AI may have vis-à-vis its own competence, it may not reflect what happens with humans either (Kammerer and Frankish 2023). However, some degree of introspection and metaknowledge seems important for AI to be able to direct its inquiry. Absent those, it is hard to imagine how AI could be seen as an agent of discoveries at all.

4. Can AI meet the requirements?

The main question of this paper is whether AI can make scientific discoveries. In this section, I would like to present a few possible hurdles that we face when answering that question in the positive. In particular, I believe that, even for high functioning models, it is quite difficult to assess whether they satisfy (C1). And for making sure current models meet (C2), we need to endow them with the abilities to assess their own uncertainty, which is a challenging task as well.

Let's take these points in turn. AI can be used to advance many sectors of scientific endeavor, as we saw above. However, one may wonder whether it really possesses the competence one needs to make scientific discoveries. The answer seems to be obvious,

at least at first blush: if one adopts a form of reliabilism about competence, one may say that an AI is competent in some field of inquiry if it reliably succeeds in making predictions and classifications in that field of inquiry. I believe this is not a good strategy and this section is devoted to show why even an apparently reliable model may lack competence.

Let me begin by reminding the reader that I am solely referring to data-based models with multiple layers (Knüsel and Baumberger 2020). What the models learn - their ability and competence - come from data. Furthermore, while deep neural networks demonstrate exceptional performance on numerous tasks, these models struggle to generalize their learning to examples that deviate from the data they were trained on. And data can be biased, insufficient or missing in a way that is hard to detect: even when seemingly reliable, those models may actually fail to have the required knowledge and competence.

To make these worries more clear, let's use an illustrative example of how things can go wrong when training machine learning models. Imagine you want to train a model to recognize cats, and you are building your training dataset. However, for some reason, the images you choose are only of white cats. In this scenario, you are training and testing your model on data that lacks diversity and comprehensiveness. The goal is to build a model that understands cat images in general, but your training data is limited to only white cats. In this case, your model might be reliable at recognizing white cat images, but it lacks knowledge about cats of other colors or breeds. More precisely, while your model may perform well when tested on white cat images (the data it was trained on), it will likely struggle to generalize to identify cat images outside of that narrow training distribution. The correct predictions made by your model on cat images would not be based on a comprehensive understanding of 'cats' in general; they would solely

rely on the knowledge of whatever is present in the training data. Hence, the model's reliable performance in spotting *cats* is the result of luck, namely what happens when the model is tested and deployed only on inputs that match the training distribution. However, once the model encounters cat images that deviate from the training data (e.g., black cats), it will fail to recognize them as cats.

This highlights a significant issue: the correct predictions made by the model are not based on knowledge of *cats*, but rather on the biases and limitations of the training data to only particular types of cats. So based on its performance one may think that the model is competent about cats, when in fact it is only competent about white cats.

Now imagine that your AI is making a discovery about a new type of white cat. The fact that the AI model has insufficient data, hinders its ability to know about cats in general. The finding of a new type of white cat is matching exercise, not an actual scientific discovery. As we said above, C-1 requires the discoverer to have competence in the field of inquiry. But if a model only knows white cats, its competence is too narrow because it does not have any real idea on what cats are. If an AI model lacks a robust grounding – even in the training data – in the fundamental nature and defining characteristics of the concept in question (in this case, "catness" or what makes a cat a cat), then any proclaimed "discovery" of a new type of cat is a finding based on superficial features, rather than a substantive insight based on grasping of the subject matter.

The key question we should be asking when dealing with AI is: when the AI provides good predictions and assessments, are those outputs based on its competence of the subject matter? Or are the predictions merely a result of chance or insufficient knowledge, stemming from training on incomplete or biased data? If an AI model is

trained on insufficient or skewed data that fails to capture the full scope of the domain, that will hinder the model's outputs. The model may excel at making predictions on examples similar to its training data, but it will lack the foundations required to generalize its understanding to novel scenarios or edge cases. Therefore, a purely reliabilist assessment of a model based on its track record is not enough to establish competence.

Before concluding this section, let me consider a reliabilist rejoinder here. The reliabilist may argue that all we need to assess competence is to (1) make sure the model in question is reliable about some Y (e.g., cats), and (2) it is trained on a set of data that represents Y enough to be sure that the model's knowledge is not limited to some subset of Y (e.g., white cats). There are rigorous validation and testing processes many AI systems undergo, and these may often be able to limit the gaps in the training data. [19]

Unfortunately, that does not work either. The issue I am raising in this section is not simply about having a representative data set. The issue is that we do not know what the model is latching onto when making its (reliable) assessments. This epistemological worry is grounded on the fact that black box models often leverage hidden features: patterns in the data that are not apparent to human observers. While this capability can be advantageous, it also means that errors based on these hidden features are

[19] This solution is actually too optimistic. The gaps in a model's training do not merely reflect a lack of specific knowledge (about, say, black cats are). The point goes deeper: models do not seem able to generalize. And it is still not clear how much training data they would need to be able to actually acquire the right competence and ability to generalize. Some scientists believe we need to train these models on so much more data and at some point, they will 'get it'. Others believe it is an architectural issue. As long as this is not solved, we cannot really attribute them competence only based on reliable performance.

harder to detect. One major concern with black box models is their potential reliance on spurious correlations: these models might learn to associate irrelevant features with the target variable simply because those features appear frequently in the training data. Hence, unbeknownst to us, high performing black box models may be inadvertently learning to rely on irrelevant features in the environment. That means that a model may be performing well but for the wrong reasons.[20]

4.1 AI and C1-r

In this section I ask whether – when AI finds some previously unknown scientific object, process or theory – it does so at least in part as a result of a reasoning process[21] common

[20] Another significant issue that limits a model's competence, is the noise or errors in the training data. Suppose you realize the need to expand your training examples and start collecting images of cats from different angles, colors and breeds. Even with this diverse data, the labeling process (where you identify each image as containing a cat or not) can inadvertently introduce noise. For example, a human labeler might incorrectly label an image containing both a cat and a dog as just a "cat". As long as you feed the model with examples that match the training data, it may perform well. However, in real-world scenarios where the model encounters images containing cat with dogs, the noise in the training data can cause the model to make incorrect predictions, such as labeling images of cats and dogs as simply cats.

[21] One worry this may not be the case is that, condition (C1-r) above talks about the reasoning process leading up to scientific discoveries. One may be skeptical that AI is able to reason at all. Even if we adopt a very liberal view of what this reasoning amounts to, it does seem to involve some sort of representations of – what we may call – scientific 'facts'. Several accounts of explicitly reject the possibility of AI having representations. For instance, Giere (2010) advocates an intentional conception of representation in science that necessitates considering scientific agents and their intentions. Similarly, Suárez (2004)

in the field in which discoveries of that type of object, process or theory would take place. This may seem a tall order for deep learning models, as we said. It seems that the most promising AI models work very differently than human reasoning, as they are able to extract patterns and learn correlations among huge amount of data.

To be sure, humans learn from experience through a process of trial and error, observation and feedback. This learning is often reinforced by successes and failures, leading to improved understanding and skills over time. Supervised learning in machine learning involves training a model on a labeled dataset, where the model learns to make predictions or classifications based on the data it has seen. This is not too dissimilar from how humans learn from experience, either. Reinforcement learning, where an agent learns to make decisions by receiving rewards or penalties, also closely mirrors human learning from consequences. Finally, even in science human ingenuity involves making judgments without explicit reasoning, often based on subconscious pattern recognition and prior experience. Deep learning models, particularly neural networks, can make complex predictions that may seem unintuitive, involving making unexpected associations.

That said, however, with current black box models we struggle to understand what is really going on under the hood: their internal processes remain largely inscrutable, obfuscated by the complex interplay of millions of parameters and non-linear transformations. Black box models like deep neural networks consist of numerous

---

defends the idea that *only* intentional agents can produce inferences (Boge 2021; Tamir and Shech 2022; cf Sullivan 2023). On these views at least, AI does not reason. Though this is a valid concern, I will put it on one side here.

layers and connections, where each layer applies various transformations to the input data. These transformations are influenced by millions of parameters that adjust during training to minimize errors. The intricate adjustments and interactions among these parameters contribute to the model's ability to make accurate predictions but also render the process opaque. The non-linear nature of these transformations further complicates interpretability: non-linearity allows models to capture complex patterns in data, but it also means that small changes in input can lead to disproportionately large and non-intuitive changes in output, making it hard to trace the logic behind specific predictions. The inscrutable nature of opaque AI systems obscures the causal pathways that underlie their predictions, making it challenging to ascertain the origin and foundation of what they generate (Khalifa 2017; Creel 2020). Without clear insights into what is going on in the model, it is hard to see how AI can match a requirement such as C1-r.

To be sure, this epistemological issue is in a sense orthogonal to whether an AI model has the right competence and performs the right reasoning when operating within science. Nonetheless, in asking the question of whether AI models can make discoveries, we are faced with the problem of determining their level of competence in relation to our current scientific practice, and C1-r highlights one way in which that competence could be expressed. Unfortunately, in black box models we do not have a clear view on the reasoning patterns that brought the model to output a certain solution (Duede 2023).[22]

[22] We have technical tools that promise to look into backboxes and gain some insights on their functioning and even abilities (Gunning et al 2019; Barredo et al 2020; (Wexler 2017; Guidotti et al 2018). Creel (2020) highlights an important perspective on the nature of opacity in AI systems by arguing that depending on the type of opacity (overall functional opacity, structural opacity, or specific run opacity) different Explainable AI methods can be

Absent that, we need to at least suspend judgement on whether high performing models satisfy C1-r.

4.2 AI and C1-k

Whereas AI can *help* in the process of discovery (Duede 2023; Tamaddoni-Nezhad et al. 2021), it rarely offers an explanation for why such finding is a scientific discovery. As they stand, deep neural nets could be successful in making predictions and categorizations but they do not provide explanations of the right kind (Boge 2022; Boge et al 2022). In particular, deep neural networks do not provide the logical or causal steps that link these results to scientific laws or principles: they might identify that, e.g., certain pixel patterns are associated with a disease in medical imaging, but cannot explain the biological or medical reasoning behind the association. Let me give another example: though outstanding in their own right, the predictions made by AlphaFold still require human interpretation and validation. Scientists must integrate these predictions into existing knowledge, design experiments to test them, and provide the broader theoretical context that gives meaning to these structures. AlphaFold can be seen as part of a hybrid model where AI and human scientists collaboratively engage in the justification process. AI provides data-driven insights and preliminary validations, while humans integrate these insights into broader scientific narratives. Thus, while

applied to make AI models transparent enough to meet the specific goals of scientists. However, none of these techniques is for now able to fully address the issues mentioned above (Awotunde et al 2022).

AlphaFold aids in discovery, the context of justification remains largely a human endeavor. Thus, AlphaFold does not satisfy C1-k.

The limits of this kind of technology seem to rest in part[23] on the fact that they are black boxes: AI's opacity seems to limit its ability to satisfy C1-k and engage in the context of justification (Creel, 2020). In addition, if we adopt a more demanding reading of C1-k, what AI is lacking is ability to *understand* its outputs and their relevance for science. Skow (2018) maintains that to understand why Q one also needs to appreciate *how* Q happened, or *why* it happened. According to Hills (2016) understanding often requires grasping expectations and being able to anticipate and explain outcomes under different conditions. This implies an ability to connect explanations to a broader conceptual framework or theory.

Having some understanding seems an important step in taking part in the context of justification which is key to being able to be an *agent* of discovery in science. This lack of explainability and related understanding is thus a significant barrier to deep neural networks' acceptance as full-agents of discovery within the scientific community, where the ability to engage in the context of justification is important. [24]

4.3 AI and uncertainty

[23] Sullivan (2022) argues that the real issue is not that these models are inscrutable but that they fail to link up with the real target of inquiry and can at best provide possible explanations but not actual ones.

[24] Michel (2019) argues that for scientific discoveries we need an element of external recognition too. It is difficult to imagine this recognition can happen with black boxes, whose patterns of inference are invisible and inaccessible.

For something to count as a genuine epistemic achievement by an agent, there seems to be an implicit requirement that the agent has some level of ability to reason about their own knowledge and uncertainties. Indeed, having insights of one's own strengths and epistemic shortcomings, allows one to direct own's inquiry toward the goal of truth while restraining from committing to something likely to be false. This insight allows an agent to recognize the limits of their current competence; identify gaps, anomalies or areas requiring further inquiry; deliberately formulate questions/hypotheses to reduce uncertainty; evaluate the strength of evidence for or against hypotheses. Similarly, having insights about its competence for AI means having (at least) the ability to determine gaps and errors in what it learnt from the training data and assess its certainty or uncertainty when producing an output (Soleimany et al 2021).[25] That would be a step in the direction of allowing an AI agent to purposefully "exercise control" over the discovery process, making its discovery a true achievement.

Now, some machine learning experts may argue that AI systems do have a way to assess confidence in their outputs through the use of probability distributions. When an AI model is deployed, it typically provides predictions in the form of probability distributions over the possible output classes. Machine learning practitioners often interpret these probability values as a measure of the model's confidence in its predictions.

However, there is a growing realization that these probability distributions may not accurately represent the model's true confidence or uncertainty (Moloud et al 2021). Consider again a scenario where a model is trained on a limited dataset of only white cats, and then encounters examples of black cats during inference. The model may

[25] This approach comes from a MIT lecture by scientist Alexander Amini https://www.youtube.com/watch?v=toTcf7tZK8c

confidently predict that a black cat image does not contain a cat, assigning a high probability (e.g., 0.8) to the "not cat" output. In this case, not only is the model's prediction incorrect, but its assigned probability distribution is also misleading as it is expressing high epistemic confidence in an erroneous output. This highlights a fundamental issue: the model lacks the ability to gauge its own limitations and express an appropriate model's uncertainty.

This shows the need for uncertainty estimation in these models. That is, there is a need for AI systems to develop robust uncertainty estimation capabilities that go beyond simple probability distributions. An ideal system should be able to: assess its own knowledge and uncertainty levels during the training phase, identifying areas where it has strong or weak knowledge based on the quality and comprehensiveness of the training data. What's more, the model should, during inference, provide not only predictions but also well-calibrated epistemic uncertainty estimates that accurately reflect the model's confidence or lack thereof, given the input data and the scope of its knowledge.[26] In contrast, if the model cannot assess its confidence in the outputs it provides, then we can argue that it does not truly have a way to evaluate whether its predictions are correct or incorrect.

[26] One potential solution that has been proposed involves the development of metacognitive architectures for AI systems (Bergamaschi et al 2021). In such an architecture, a metacognitive module would have access to both the outputs and the uncertainty estimates of various sub-components (e.g., black-box models, white-box models, rule-based systems), and make decisions about which one to adopt.

There is important research being done in this area. Techniques such as ensemble methods are designed to compute uncertainty in the model (epistemic uncertainty) by looking at the variance across stochastic models with same hyperparameters and trained on same data. Using ensemble methods is however extremely difficult and requires a lengthy process, making it hard to adopt. This issue concerning ensemble techniques limits models' ability to assess the limitations of their own knowledge: if they cannot use it, and do so quickly, it is hard to imagine they can direct their activities based on their uncertainty estimation. This raises skepticism about whether current AI involved in discoveries systems is in fact in a position to figure out if they possess genuine competence and provide reliable confidence assessments for their outputs. Unless ensemble techniques are adopted, these systems are currently "blind" with respect to their own abilities and the confidence they should have in their outputs. If this self-reflection and ability to assess one's own gaps is a necessary condition or an important characteristic of being an agent of discovery, then "discoveries" made by these systems may just blind findings, requiring human experts to recognize and validate the significance of those results.

While progress is being made in areas like uncertainty estimation and metacognitive architectures (Kadavath et al., 2022; Lin, Hilton and Evans 2022), significant challenges remain. Developing AI systems with true expertise and the ability to make genuine scientific discoveries requires not only robust competence but also the capabilities to assess, reason about, and address gaps in that knowledge. Without these it does not seem we have enough to grant AI models the status of scientific discovers.

5. Conclusion

AI can help humans make scientific discoveries and is a wonderful tool to conduct scientific research. AI is a reliable *finder* in many cases as it can find new patterns, spot new objects and species, analyze trends in novel and exciting ways. The question is, however, whether AI can discover anything *by itself*. Given the challenges and limitations we have discussed, and the lack of clear solutions to address them, it is reasonable to conclude that, at least for now, AI systems are not truly making scientific discoveries because they do not seem to have the necessary competence and the ability to estimate their epistemic status when reaching conclusions.

Acknowledgment: I extend my sincere gratitude to Allan Hazlett and Carl Craver from the Philosophy Department at Washington University in St. Louis for their valuable comments and feedback on an early draft of this paper. I am also grateful for the insightful discussions with participants at the "Artificial Researchers and Scientific Discoveries" conference at the University of Düsseldorf which enriched the development of this work.

References

Abramson, J., Adler, J., Dunger, J. et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*.

Anichini, F. et al. (2021). The automatic recognition of ceramics from only one photo: The ArchAIDE app. Journal of Archaeological Science: Reports, 36, 102788.

Awotunde, J.B., Adeniyi, E.A., Ajagbe, S.A., Imoize, A.L., Oki, O.A., & Misra, S. (2022). Explainable artificial intelligence (XAI) in medical decision support systems (MDSS): Applicability, prospects, legal implications challenges. The Institution of Engineering and Technology, London, 45-90.

Barredo Arrieta, N., Díaz-Rodríguez, J., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion, 58*, 82-115.

Bergamaschi Ganapini M., Murray Campbell, F. Fabiano, L. Horesh, Jonathan Lenchner, Andrea Loreggia, Nicholas Mattei, Francesca Rossi, Biplav Srivastava, and Kristen Brent Venable. (2021) Thinking fast and slow in ai: the role of metacognition. In International Conference on Machine Learning, Optimization, and Data Science.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. Minds and Machines, 32(1), 43–75.

Boge, F. J., Grünke, P., & Hillerbrand, R. (2022). Minds and Machines special issue: Machine learning: Prediction without explanation? *Minds and Machines, 32*(1), 1–9.

Clark, E., & Khosrowi, D. (2022). Decentring the discoverer: How AI helps us rethink scientific discovery. Synthese, 200, 463.

Cockburn I. M., Henderson R., Stern S. (2018). *The impact of artificial intelligence on innovation*. National Bureau of Economic Research.

Copeland, S. (2018). 'Fleming leapt upon the unusual like a weasel on a vole': Challenging the paradigms of discovery in science. Perspectives on Science, 26(6), 694–721.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*(4), 568–589.

Duede, E. (2023). Deep learning opacity in scientific discovery. Philosophy of Science, 90(5), 1089-1099.

Gibbons, J. (1981). The biogeography of Brachylophus (Iguanidae) including the description of a new species, B. vitiensis, from Fiji. Journal of Herpetology, 15 (3), 255–273.

Giza, Piotr (2021). Automated discovery systems, part 1: Historical origins, main research programs, and methodological foundations. Philosophy Compass 17 (1):e12800.

Greco, J. (2010). Achieving knowledge. Cambridge: Cambridge University Press.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.Z. (2019). XAI-Explainable artificial intelligence. Sci Robot, 4(37), eaay7120.

Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z.H., DasSarma, N., Tran-Johnson, E., & Johnston, S. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Khalifa, K. (2017). Understanding, explanation, and scientific knowledge. Cambridge University Press.

Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. Studies in History and Philosophy of Science Part A, 84, 46–56.

Hempel, C. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Hills, Alison, 2016, "Understanding Why", *Noûs*, 50(4): 661–688.

Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. arXiv preprint arXiv:2205.14334.

Miao, Q W. Zheng, Y. Lv, M. Huang, W. Ding and F. -Y. Wang, "DAO to HANOI via DeSci: AI Paradigm Shifts from AlphaGo to ChatGPT," in IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 4, pp. 877-897, April 2023

Michel, J.G. (2019). How are species discovered? Declarative speech acts in biology. Grazer Philosophische Studien, 96(3), 419–441.

Michel, J.G. (2020). Could machines replace human scientists? Digitalization and scientific discoveries. In B.P. Göcke & A. Rosenthal-von der Pütten (Eds.), Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences (pp. 361–376). Paderborn: mentis/Brill.

Moloud, A., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications, and challenges. Information Fusion, 76 , 243–297.

Newell, Allen & Simon, H. A. (1976). Computer Science as Empirical Inquiry: Symbols and Search. Communications of the Acm 19:113-126.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Cambridge, Massachusetts: Harvard University Press.

Popper, H. (2002). The logic of scientific discovery. Routledge Classics.

Proust, J. (2010). Metacognition. Philosophy Compass, 5(11), 989–998.

Reichenbach, H. (1938). Experience and prediction: An analysis of the foundations and the structure of knowledge.

Robert, A. (2004). Intellectual virtue and epistemic power. In J. Greco (Ed.), Ernest Sosa and his critics (pp. 1–16). Malden, MA: Blackwell.

Skow, Bradford, 2018, "Against Understanding (as a Condition on Explanation)", Grimm 2018: 209–231, *Making Sense of the World: New Essays on the Philosophy of Understanding*, New York: Oxford University Press.

Soleimany, A. P., Amini, A., Goldman, S., Rus, D., Bhatia, S. N., & Coley, C. W. (2021). ACS Central Science, 7(8), 1356–1367.

Sosa, E. (2007). A virtue epistemology. Oxford: Oxford University Press.

Sosa, E. (2011). Knowing full well. Vol 1 and2. Princeton, NJ: Princeton University Press.

Strevens, M. (2020). The knowledge machine: How irrationality created modern science. Liveright.

Sullivan, E. (2022). Understanding from machine learning models. British Journal for the Philosophy of Science, 73(1), 109–133.

Sullivan, E. (2023). Do ML models represent their targets? Philosophy of Science, 1–14.

Tamaddoni-Nezhad, A., Bohan, D., Afroozi Milani, G., Raybould, A., & Muggleton, S. (2021). Human–machine scientific discovery. In S. Muggleton & N. Charter (Eds.), Human-like machine intelligence (pp. 297–315). Oxford University Press.

Wang, H., Fu, T., Du, Y. et al.(2023) Scientific discovery in the age of artificial intelligence. Nature 620, 47–60.

Watson, J. (2021). Expertise: A philosophical introduction. Bloomsbury Academic.

Wexler, J. (2017). Facets: An open-source visualization tool for machine learning training data. Google Open Source Blog.

Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology, 37, 1038–1040.