Can Artificial Intelligence Make Scientific Discoveries?

JESSICA SARAH LAUMAN-LAIRSON

DURHAM UNIVERSITY

MASTERS THESIS

Can Artificial Intelligence Make Scientific Discoveries?

Author:

Jessica

LAUMAN-LAIRSON

Supervisor:

Dr. Sara UCKELMAN

Dr. Peter VICKERS

A thesis submitted in fulfillment of the requirements

for the degree of Master of Arts

in the

Department of Philosophy

Department of Philosophy

September 10, 2019

DURHAM UNIVERSITY

Abstract

Department of Philosophy

Master of Arts

Can Artificial Intelligence Make Scientific Discoveries?

by Jessica LAUMAN-LAIRSON

Artificial intelligence is causing a paradigm shift in the scientific method. Traditionally scientific discovery has been a task performed by human scientists using diverse processes from mathematical formalisms and mental representations to abduction and moments of lucky Gestalt shift. After rationalist pursuits of a logic of discovery waned, accounts of scientific discovery largely shifted from philosophy over to psychology, with discovery being increasingly conceptualized as a human psychological process. Yet the application of increasingly successful artificial intelligence programs to science has brought about claims in both philosophy and science that AI is making scientific discoveries. In this thesis I will examine several philosophical conceptions of scientific discovery, particularly Kuhn's taxonomy of puzzle-solving versus revolutionary science. Using these definitions, I aim to clarify the scope of AI's ability to contribute to scientific discovery. I argue that the discoveries which AI can make have distinct characteristics that correlate with Kuhn's notion of the puzzle-solving discoveries that occur during normal science. In contrast, I will argue that AI currently lacks capabilities necessary for a kind of discovery that I call conceptual discovery. Current constraints on the scope of AI's contributions to discovery that I will discuss include the frame problem, lack of a proper representation language, and in particular, the challenge of formalizing abductive and analogical reasoning.

Contents

Chapter 1

Introduction

1.1 Introduction

Artificial intelligence is bringing changes to the scientific discovery process. With its ability to process huge amounts of data quickly, AI can outperform humans in certain tasks that have traditionally been performed by human scientists. All existing AI programs fall under the category "narrow intelligence", in that they are designed to perform domain-specific tasks with high efficiency but lack domain-general reasoning. The goal of "narrow" AI is "creating systems that can perform particular functions that used to require the application of human intelligence (Kurzweil, 2005, pg.72)." My aim will not be to evaluate this distinction. Rather, I will investigate the fact that, within specific domains, AI systems are surpassing human performance at discovery tasks like classification, hypothesis generation, and pattern detection. The question I will address is whether the contributions that AI is making to these areas can be legitimately characterized as AI making scientific discoveries.

Headlines like "AI Trained on Old Scientific Papers Makes Discoveries Humans Missed"(Gregory, 2019) and "Two New Planets Discovered Using Artificial Intelligence"(Texas, 2019) offer an ever more impressive picture of the role AI can play in scientific discovery. In China, Yitu Technology's AI helps radiologists analyze 1.4 billion CT scans per year to catch early signs of

lung cancer. Because AI can process many times more data than radiologists could achieve by hand, it is able to detect patterns that radiologists would miss (Radiology, 2019). AI has also been utilized to avoid visual biases that human radiologists are susceptible to, like inattention blindness [Drew 2013] and satisfaction of search bias [Busby 2017]. AI is quickly developing abilities that a few years ago were considered distinctly human, like the ability to read research papers. In California, an AI system called KnIT read 100,000 research papers in a couple of hours and made new discoveries in biology that were hidden in the information contained in the papers. This discovery included finding 7 p53 kinases (a type of enzyme that prevents cancer) that were discovered in the 10 years after the papers were written, and 2 p53 kinases which were unknown to scientists prior to the AI discovering them (Choi, 2018).

While the "general intelligence" of AI is continually developing, currently a main factor contributing to the success of AI as a tool for scientific discovery is its ability to process large quantities of data quickly. Creating an AI system that has human-level cognitive abilities is an implicit goal in both "weak" and "strong" AI development. Yet AI does not need to perform human-like intellectual exploration in order to produce hugely successful results. Current large language models produce meaningful sentences in a way that mirrors Wittgenstein's concept of language games. For instance, in the case of Wittgenstein's builder and assistant, the assistant speaks a different language than the builder and does not understand what "block" or "slab" means, but learns by trial and error and is able to hand the builder the correct object when the builder says "block" or "slab". Wittgenstein says,

> The language is meant to serve for communication between a builder A and an assistant B. A is building with building-stones: there are blocks, pillars, slabs and beams. B has to pass the stones, in the order in which A needs them. For this purpose they use a

> language consisting of the words "block", "pillar" "slab", "beam". A calls them out; — B brings the stone which he has learnt to bring at such-and-such a call. Conceive this as a complete primitive language [Wittgenstein 1953].

The assistant's ability to complete the task effectively comes from their ability to learn the contexts in which the word is used. AI is like Wittgenstein's assistant——it can encode collections of family resemblances and use these to use a term correctly without understanding what the term means. With their processing speed, AI systems can carry more blocks and slabs in a second than any human could carry in a hundred years, and in doing so can detect more fine-grained patterns than a human can.

Alan Turing said that the question 'can machines think?' is "too meaningless (Turing, 1950, pg.442)" in and of itself to be discussed. Instead, Turing proposed an empirical test as a behavioral measure of intelligence in AI systems. The object of the Turing Test is for an interrogator to correspond with two entities, and to correctly determine which is a human and which is a machine. Questions of both an empirical and conceptual nature rise in response to the Turing test. The first concern is practical and revolves around whether and how a computer might pass the Turing test. The latter questions concern the legitimacy of the Turing Test: should we conclude that a machine is intelligent on based on behavioral, empirical grounds?

The question I investigate is whether AI can make scientific discoveries, not whether they can think. However, these questions could be intertwined to a degree. My thesis could attempt two different ways of answering the question "Can AI make scientific discoveries?" First, I could approach this as a question about the nature of discoverers. This would include questions like "Who is qualified to make a discovery?", and "At what point does a computer stop being a tool (like a microscope) and become more like a discoverer (like a human scientist)?" However, I am going to circumnavigate the

philosophy of mind question of whether computers can think. Instead, I will focus on the structure of discovery, treating human discovery as a product and process with distinct characteristics. The thesis will examine the scope of AI's current ability to contribute to these products and processes.

One reason for focusing on questions about outputs is that AI systems are a black box. Discussing the future of AI from a legal standpoint, Yavar Bathaee (2018) points out that intent and causation, two important measures of human conduct, are difficult to apply to neural networks. The decision making process of the AI can be an impenetrable black box even to the researchers who programmed it (Davoide Castelvecchi 2016). Researchers trying to develop ways to understand the inner working of AI are much like neuroscientists trying to understand the brain (Castelvecchi 2016). Multilayer neural networks are a black box as much as the human brain is, and like the brain, data is not stored in distinct modules. Instead, information is diffused across the network, and this makes understanding the information and how it is used difficult. Castelvocchi says, "Eventually, some researchers believe, computers equipped with deep learning may even display imagination and creativity (Castelvecchi 2016).

I will refrain from claiming any association between my discussion and the "strong"/"weak" intelligence debates. My thesis will not take a stance on the prerequisites of human intelligence and whether AI can achieve this. Instead, I will explore the nature of discovery and what processes it involves. The question that this dissertation will address is what scientific discovery is and whether AI can make scientific discoveries under the definition of discovery which I will aim to establish in the next section.

1.2 What is a Scientific Discovery?

I will first devote a section to developing a philosophically rigorous definition of discovery that will be the framework for establishing the scope of AI's ability to discover. What is a scientific discovery? At first, the answer seems straightforward; surely a discovery is a realization, a new idea, a eureka moment. However, the term describes a range of phenomena, and these are not well organized in the philosophical literature. In very general terms, a scientific discovery can be described as a "process or product of productive scientific inquiry (Schickore, 2014)". The fact that discovery is presented alternately as a process and a product is relevant to my discussion. "Can AI perform X process?" and "Can AI produce X result?" can potentially be very different questions. My definition of discovery will treat it as both process and product, and will evaluate AI's ability to discover on both criteria.

Scientific discovery can generally be understood as the discovery of natural kinds combined with theory about those kinds. Somewhat implicit to the natural kinds thesis is the idea that natural kinds are categories that are found in nature independent of human perception. There is some debate about this; for instance although Kripke and Putnam are realists about essences, there are more intermediary accounts like Rand's that advocate that natural kinds are epistemological and based in human perception though they have a basis in a mind-independent reality. A paradigmatic support of the thesis that natural kinds are intrinsic to nature is the categorization of species. However, this runs into a problem because the way that one understands the concept 'species' affects decisions about how to categorize. For instance, one could categorize by an evolutionary perspective, where birds and reptiles are in the same category because they share a common ancestor, or from a traditional biological classification perspective, where they are of a different kind because reptiles and birds belong to different classes of animal. Other options

for categorization are by ecology or phylogeny. P.D. Magnus says, "A natural kind is a category that scientists are forced to posit in order to be scientifically successful in their domain of inquiry (Magnus, 2012, pg.47)".

It will be significant whether my definition of discovery means that something that exists is uncovered, or means that something new is created. There is philosophical debate about whether discovery is better characterized as invention. Should "discovery" mean (1) a concrete piece of knowledge which exists in the world and has been uncovered by scientists as an archaeologist would discover an artifact or (2) an epistemological development that creates a new way of structuring reality? Under Francis Bacon's account, the external world and internal knowledge are directly linked. The world has preexisting laws and scientists simply find these laws and give a name and description to them. Under this realist view, science is progressive and laws and patterns exist in the data prior to observation. All AI would need to do to make discoveries would be to find these patterns in the data and describe them successfully.

More recently, philosophy of science has turned away from strong realism and the view that discovery is a cumulative process of uncovering objective truths about the world. Instead, many contemporary philosophers view discovery as an active process in which patterns are created rather than found. More emphasis is being placed on the role that language, social values, and existing theoretical frameworks play in shaping discovery [Kuhn 1962].

Though widely less popular in these days than it was in the 19th century, realist views of discovery are making a comeback in some areas of AI research. One example is the BACON system, named after Francis Bacon, which I will discuss in section 2.

Following the work of Piscopo and Birattari, I will label these two ways of conceptualizing scientific discovery as inventionist and discoverist (C. Piscopo, 2013). Discovery under an inventionist lens will describe discovery as

something which is created. Under this definition a discovery invents a reality rather than providing a way to conceptualize a reality that already exists. Discovery under a discoverist lens will describe the realist stance in which discoveries are cases where an objective truth about the world is discovered by scientists.

Many philosophical discussions about discovery examine the process by which theories that explain empirical data are created. However, a neutral definition may be much less specific; discovery might be called a moment of successful science. Kuhn makes a distinction between discovery-that (where a viewer observes a discovered phenomenon) and a discovery-what (a correct theory about what something is, where a viewer creates a useful way of thinking about the phenomenon) (Kuhn, 1962, pg. 55). Kuhn says that both discovery-what and discovery-that are necessary for successful scientific inquiry; it is too simplistic to say that we simply discover a phenomenon without considering the role of our conceptualization of what we see. Discovering a scientific phenomenon "involves recognizing both that something is and what it is (Kuhn, 1962, pg. 55)". For instance, Joseph Priestley discovered the existence of oxygen, but from the conceptual framework of phlogiston theory that he was working from, he assumed oxygen was air with minimal phlogiston. So Priestley's discovery constituted a discovery-that but not a discovery-what. If we want to argue that Priestley discovered oxygen, then this would entail arguing that anybody who first captured an impure sample of oxygen counts as having discovered it. Under this weaker notion of discovery, we are not able to attribute the discovery to Priestley, but merely to that time period. Thus, Kuhn advocates that it is fairly meaningless to attribute discoveries to specific people because they are generally "not isolated events, but extended episodes (Kuhn, 1962, pg. 52, Schindler, 2015, pg.125)".

There is much conversation in philosophy about devising a method by which to decide who is making a discovery and when a discovery is made

(Hudson 2001, pg. 77). Kuhn's account allows one way of making this distinction–we can say that an individual is a discoverer if they make both the discovery-that and the discovery-what (both being the first to observe the phenomenon and the first to correctly conceptualize it. However, Kuhn thinks that this is not a frequent occurrence, and more often a scientific discovery is the product of a scientific community.

A few challenges arise though with the definition of discovery-what. Kuhn's account advocates that the discovery-what occurs when a correct conceptualization of a phenomenon is achieved. Here, Kuhn sounds as though he is taking a more realist standpoint. What is a 'correct' conceptualization? Especially under an inventionist view of discovery, a 'correct' conceptualization of a phenomenon means that a theoretical framework is employed to interpret the phenomenon in a way that is fruitful to scientists' current goals. Kuhn says that "scientific revolutions are…those non-cumulative developmental episodes in which an older paradigm is replaced in whole or in part by an incompatible new one (Kuhn, 1962, pg. 92)." Under his account, a paradigm is accepted because it is persuasive to scientists, not because it is inherently closer to some objective truth. Science evolves, but does not necessarily progress linearly (Kuhn 1962). Under this view, it becomes hard to understand what it means to have a correct conceptualization of a phenomenon. Does Kuhn mean 'correct' in an ontological sense, or would a more pragmatic measure like a paradigm being accepted by scientists be enough to qualify a successful discovery-what?

Similarly, *how* correct does the conceptualization need to be in order to qualify? There are many scientists who we consider to have been the discoverers of a phenomenon, even though their account later turned out to be at least partially incorrect. Schindler (2015) proposes an answer to this. He proposes that a discovery-what of X requires a

> correct conceptualization of X's essential properties that suffice

> (epistemically) to individuate X at a time t (the time of the discovery-what), whereby I take essential properties of X to be those properties of X that are (metaphysically) individually necessary and jointly sufficient for X (Schindler, 2015, pg. 132).

Schindler proposes that this solves the problem because making the argument time-dependent means that not all the essential properties need to be known to individuate X and thus correctly conceptualize X. More properties may be discovered at a later time, and this does not disqualify the original discoverer because they correctly conceptualized X's properties at a time when the properties they conceptualized were sufficient to individuate X. The example Schindler gives is Thomson's discovery of the electron. Thomson is widely regarded as the discoverer of the electron, yet if a discovery required an entirely correct conceptualization than we would not be able to consider Thomson the discoverer, because he was not aware of properties of the electron like wave-like behavior and instead thought of them as corpuscular entities of negative charge. Yet intuitively we do want to argue that Thomson was the discoverer of the electron; he was the first to observe and conceptualize a negatively charged particle at a time when there were not known to be any negatively charged particles. Schindler's account of discovery solves this problem because Thompson's understanding of the particle's properties at time t were enough to individuate the electron at that time by being individually necessary and jointly sufficient for X. Now, negative charge would no longer be sufficient for individuating the electron because we are aware of other negatively charged particles, but in the conceptual environment of the time it was sufficient.

This distinction meets a challenge if we attempt to generalize it to AI, in utilizing it to distinguish which of AI's contributions to science count as discoveries. What is a 'correct conceptualization' when an AI system, rather

than a scientist, is providing the conceptualization? Is there a set of behaviours that an AI can perform or a way that it can structure a result in order to be said to achieve a correct conceptualization? One problem with this is that the concept of a "conceptualization" has assumptions about human psychology built into it, and is thus difficult to define in AI terms. AI does not structure models in terms of laws and theories, as humans do; instead AI uses complex, abstract, black-box models. Newton's laws of motion, if invented in the modern day by AI, would not be in the form of simple laws; they would be abstract computer models that would output extremely specific results but would not have a conceptual element that humans can understand. Many of the epistemic values that scientists value, like explanatory power, interpretability, simplicity, and generality might be disregarded by the AI system in favor of a complex, abstract model that yields highly accurate predictive results. This yields at least two questions that need to be addressed before we can provide a framework for AI satisfying Kuhn's discovery-what requirement.

The second layer of the distinction Kuhn makes is between discoveries-that which are made before discovery-whats, and discovery-whats which are made before discovery-thats. The latter, what-that discoveries, occur when the theory about the phenomena comes into being before the phenomenon itself is observed. In the example above, a what-that discovery would occur if a correct conceptualization of oxygen were developed before a sample of the gas itself was isolated and observed. These discoveries most fit with Kuhn's idea of normal science, where the discoveries are predictable and resemble the process of puzzle-solving. Normal science is not unexpected or surprising and is a more linear, progressive form of science that aligns more closely with the assumptions of realist, discoverist accounts like those discussed above. Filling in missing elements on the periodic table after the periodic law has already been established is an example of normal science.

In contrast, that-what discovery is surprising and not progressive. A that-what discovery is one in which the phenomena is first observed and then conceptualized. The discovery of oxygen discussed above followed a that-what format; first the phenomena itself was discovered and later a correct conceptualization of it was developed. An important thing to note is that Kuhn does not claim that discoveries-that or that-what discoveries occur without a conceptual element. In Kuhn's view our sensory experience is inevitably influenced by our conceptions. That-what discoveries are different from what-that discoveries not because they do not have conceptualizations attached to them, but because the 'that' is not predicted or expected by our existing conceptualizations (Schindler, 2015, pg. 126). Thus a what-that discovery can overthrow existing paradigms if the conceptualizations which did not predict the 'that" are challenged by the new that-what discovery, and the conceptualizations are connected to a paradigm.

Departing from Kuhn, I will now outline some other philosophers' treatments of discovery. The goal will be to pick out some essential characteristics of discovery.

William Whewell (1840) is one of the earlier philosophers to make a distinction between discovery and other parts of scientific activity. Whewell advocates that discovery is comprised of three parts: the "happy thought", the way the "happy thought" is developed into a hypothesis, and the verification of that hypothesis. In the later philosophy literature, the process of confirmation is generally distinguished from the initial creation and articulation of a new hypothesis. Furthermore, later work often focuses on the search for a logic of scientific discovery, whereas Whewell rejects the idea that discovery has a method which can be formalized. "No maxims can be given which inevitably lead to discovery (Whewell, 1840, pg.186)". He is interested in the psychology of the discoverer and the process by which a "happy thought" becomes integrated into a system of beliefs. Under Whewell's account, the

“happy thought" is not the sole cause of the discovery, but rather is the activation energy that allows the causal pathway leading to the discovery to be initiated. Whewell says that it does not make sense to talk about why the “happy thought" causes the bullet to hit the target, because the “happy thought" is merely the spark, and the discovery can be attributed not just to the spark, but to the gun being loaded and pointed at the target (Whewell, 1840, pg.189).

The second element of discovery is the “colligation" of a set of phenomena into a conceptual system (Whewell, 1840, pg.189). The colligation produces a new conceptual framework as well as recasting the previous ideas and previously observed phenomena. It is a two step, iterative process. First, experiments yield quantitative and qualitative regularities and facts, and scientists' theoretical framework is used to bind together those facts. Second, the theoretical framework is clarified and adapted according to the information that the amalgamation of the new facts yields.

Finally, the third step occurs when the colligation recieves confirmation, and scientists decide whether the colligation yields a useful and plausible hypothesis. Scientists must decide whether the theoretical framework resulting from the colligation of the facts is sufficient to explain the phenomena being considered. This step also involves seeing whether the colligation yielded an outcome which satisfies epistemic virtues like simplicity, explanatory power, ability to make predictions, and coherence with the overall theoretical framework (Ducasse, 1951).

Whewell's account offers a more broad definition of discovery than the accounts that came later; for him, all three steps are part of the discovery process while in later accounts only the “happy moment” or the “happy moment” plus the colligation are elements of discovery, while the verification exists in the category of processes that occur post-discovery.

Bringsjord (2007) works on establishing a formalization of discovery that

allows AI to discover proofs in physics. He distinguishes three deductive tasks that have been attempted to be mechanized:

1. Checking proofs: Decide whether a deduction is sound by looking at its conclusion and premises.

2. Provide proofs: Given premises and a conclusion, decide whether the proof follows from the premises and if it does give a proof.

3. Discovering theorems: Starting with some premises, deduct a conclusion that is sound, provide a proof, and have the conclusion be a useful one (Bringsjord, 2007).

The third task is the one that relates most to the possibility of AI making scientific discoveries. However, despite historical attempts to pin down a logic of discovery, objections from Hempel and others have argued that the aim of formalizing discovery is misguided, and that discovery involves local goals and standards for relevance that cannot be captured in a universal logic, psychological factors like creativity and intuition, and goal-driven, pragmatic considerations that depend on the context of the discovery in question.

In *Science and Method*, Poincare attempts to describe his own discovery process, concluding that it is challenging to define. He says,

> Discovery…does not consist in making new combinations with mathematical entities that are already known. That can be done by any one, and the combinations that could be so formed would be infinite in number, and the greater part of them would be absolutely devoid of interest. Discovery consists precisely in not constructing useless combinations, but in constructing those that are useful, which are an infinitely small minority. Discovery is discernment, selection…One is at once struck by these appearances of sudden illumination, obvious indications of a long course

> of previous unconscious work. The part played by this unconscious work in mathematical discovery seems to me indisputable (Poincare, 1908, pg.79).

Mathematicians' work consists not in coming up with new mathematical facts; it is about recognizing the significant facts. Furthermore, it is about taking an existing set of facts, and seeing them in a new way that reveals new and interesting relationships.

Pólya emphasizes the inductive, creative nature of mathematical reasoning. He says,

> Mathematics is regarded as a demonstrative science. Yet it is only one of its aspects. Finished mathematics presented in a finished form appears as purely demonstrative, consisting of proofs only. Yet mathematics in the making resembles any other human knowledge in the making. You have to guess a mathematical theorem before you prove it; you have to guess the idea of the proof before you carry through the details. You have to combine observations and follow analogies; you have to try again and again. The result of the mathematicians' work is demonstrative reasoning, a proof; but the proof is discovered by plausible reasoning, by guessing...I do not believe that there is a foolproof method to learn guessing. At any rate, if there is such a method, I do not know it...The efficient use of plausible reasoning is a practical skill, and it is learned, as any other practical skill, by imitation and practice (Pólya 1954).

From Peirce to the later work of Norwood Hanson, philosophers have aimed to uncover patterns in the inductive and abductive scientific reasoning that underlies discovery. Pierce and Hanson particularly focused on the

structure of abductive reasoning. The structure of abduction as scientific discovery is as follows, under Hanson's view:

1. A number of unexplained phenomena are observed.

2. There is a specific type of hypothesis that would make the phenomena unsurprising. From such a hypothesis, the phenomena would reasonably follow. Furthermore, a hypothesis of this type would offer some explanatory power.

3. Therefore, a hypothesis of this type has reasonable support justifying its existence and development. Proposing a hypothesis of this type provides an explanation for the phenomena, and this justifies it (Hanson, 1960, pg.104.) Abduction involves forming a hypothesis that does not necessarily follow from the premises but is deemed probable through analogy and inference to the best explanation. If I catch a cold after being sneezed on while riding the metro, I might hypothesize that I caught it from a person who sneezed on me. Much of scientific and common sense reasoning involves abduction (Boyd 1981).

Deduction proceeds from general rule to particular consequence:

(1) A $\Longrightarrow$ B

(2) A

(C) therefore B.

In contrast, inductive and abductive arguments are ampliative and establish new conclusions that were not contained in the premises. For instance, the following is an abductive argument:

(1) A $\wedge$ B, so (using anecdotal evidence and analogy to similar situations) perhaps A $\longrightarrow$ B.

While induction generalizes a premise, abduction proposes a premise.

For Charles Sanders Peirce abduction is "the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction

merely evolves the necessary consequences of a pure hypothesis. Deduction proves that something must be; induction shows that something actually is operative; abduction merely suggests that something may be (Pierce, 1903)." Pierce presents abduction as an inverted modus ponens. Abduction is logically equivalent to the fallacy affirming the consequent:

(1) P $\longrightarrow$ Q.

(2) Q.

(3) Therefore, P

In Pierce's view, we cannot construct new theories by using deduction and induction alone. Deduction makes implicit consequences of a set of premises explicit, and induction generalizes what is already suggested by existing instances. In contrast, abductive reasoning creates novel hypotheses that are not directly derived from established premises. Pierce says, "Abduction consists in studying facts and devising a theory to explain them (Psillos 2011).". Pierce also suggests that abduction involves analogical reasoning. He says, "Abduction, or the suggestion of an explanatory theory, is inference through an Icon (Pierce)", suggesting that abduction proceeds from the phenomenon to be explained and other, more familiar phenomena. For instance, Einstein reasons abductively that gravitational mass can serve as an explanation for inertial effects by using a simple analogy that highlights the empirical indistinguishability of the experience of a person inside of an accelerating elevator in space versus a person in an elevator at rest in a gravitational field (Einstein 1917).

1.3 Two Types of Discovery

Kuhn's well-known description of scientific progression separates scientific activity into "normal science" and paradigm-shifting scientific revolution(Kuhn, 1962, pg.10). Normal science occurs within an established paradigm, and is

roughly analogous to puzzle-solving because it involves finding new solutions using the pre-established rules, goals, and methods of the paradigm. At first glance, this matches the type of discoveries AI systems have historically been capable of. Successful automated discovery projects like BACON (Langeley et al. 1987) have a structure in which the AI system is fed datasets and a set of pattern-recognition rules that together constitute a sort of paradigm. The AI system is then asked to analyze the patterns in this dataset to come up with new conjectures.

An example is the program 'Arrowsmith', which reads through medical databases and finds connections between the contents of different studies, such as linking Reynaud's disease to fish oil as a possible treatment option.'Arrowsmith' reads through medical databases looking for broad interdisciplinary links between medical fields. It uncovers patterns that can lead to entirely new solutions to medical problems. The program is useful bfor discovering overlap between fields that has gone unnoticed by human researchers. One example is the link between the effect of fish oil on circulation and Raynaud's disease that was discovered by Don Swanson in the 1980s (Swanson, 1986). Swanson was reading a book about the Inuits when he noticed that it mentioned that the Inuit's diet was high in fish, and that the high amount of fish oil likely lead to the Inuits having good blood flow and blood vessel cold tolerance. Swanson by chance knew that a disease called Raynaud's disease causes blood vessel restriction, especially in the cold. The existing literature contained no direct link between Reynaud's disease and fish oil, but both were individually identified in different papers as having causal relationships with blood viscosity. It occurred to Swanson that fish oil might improve the symptoms of Raynaud's disease. He found a number of research papers on fish oil improving circulation and a number of papers on Raynaud's disease causing circulation problems, but there were none suggesting that one could be used to treat the other. He wrote a paper proposing

that fish oil might be an effective treatment for Reynaud's disease, and this hypothesis was corroborated by a later study (DiGiacomo 1989).
The program Arrowsmith replicated Swanson's discovery and similar discoveries like finding a link between migraines and magnesium deficiency (Hosanagar 2019; Weeber 2001). Arrowsmith's process did constitute discovery of an unrecognized correlation that suggested a causal relationship, but the discovery is a rote one that can be simplified down to a process of finding a common factor. Raynaud's disease is causally related to reduced circulation, and a diet high in fish oil is causally related to increased circulation. The process of generating the hypothesis that fish oil reduces Reynaud's symptoms amounts to applying the notion of transitivity. It doesn't seem to be a very different activity from something like giving a computer a list of flowers and their colors and having it return a list of the flowers that contain the color yellow. Arrowsmith's process is one of taking a framework and filling in gaps within that framework through extrapolation, much like chemists extrapolating new elements from blank spots in the periodic table.

In contrast, scientific revolutions occur when anomalies accumulate that the existing paradigm is unable to explain, and these anomolies become numerous and pressing enough that a crisis occurs (Kuhn 1962). If a new paradigm is able to resolve these anomalies and answer the questions that the community of scientists find most compelling, then the new theoretical framework replaces the old, bringing new methods, theoretical commitments, and epistemic norms. Kuhn likens paradigm shift to a political revolution or Gestalt shift (Kuhn 1962). While normal science is like having an empty window frame in your house and realizing that you have a window in your garage that perfectly fits that frame, paradigm shift is like realizing that your house is fundamentally flawed and unable to meet your needs even with renovations. Instead of continuing to make incremental modifications, you reconstruct it from the ground up with a different foundation.

Kuhn introduces the idea of the theory-ladenness of observation; that scientists' conceptual environment shapes their perceptions. Hanson says,

> Physical theories provide patterns within which data appears intelligible. They constitute a 'conceptual Gestalt'. A theory is not pieced together from observed phenomena; it is rather what makes it possible to observe phenomena as being of a certain sort, and as related to other phenomena. Theories put phenomena into systems (Hanson 1969, pg. 90).

Scientists don't observe raw phenomena; they view phenomena shaped by the conceptual Gestalt of their paradigm. Under Kuhn's view, scientific revolutions are a Gestalt shift. Rather than just bringing about new empirical results, revolutions restructure old and well-known empirical data. This is why proponents of an old and new paradigm can describe the same phenomena in fundamentally different ways, because the lens of the paradigm affects their basic experience of the world. The theory-ladenness of observation is one source of incommensurability, the fact that paradigm shift can cause characteristics of the old paradigm to be incomparable to the new even in terms of observations, language, basic theoretical commitments, goals and methodologies (Kuhn 1970, pg. 94).

Kuhn's distinction between normal science and revolutionary science is useful when exploring the scope of AI's ability to contribute to science. In normal science, scientists apply existing methods and rules to extrapolate new conjectures from existing theoretical commitments. This is the sort of work that current AI systems used in science can outperform humans at in some contexts. Kuhn says that normal science is a puzzle-solving activity:

> Under normal conditions the research scientist is not an innovator but a solver of puzzles, and the puzzles upon which he concentrates are just those which he believes can be both stated

> and solved within the existing scientific tradition (Kuhn, 1962, pg. 144).

In Kuhn's view, normal science allows a kind of scientific discovery in which the discoverer is like a person playing chess or solving a crossword. The activity of puzzle-solving has a pre-established structure and set of rules and acceptable methods. There is a familiarity about it; rather than something new being created, there is an expectation that an answer already exists and simply needs to be found. An especially interesting point that Kuhn makes is that a puzzle has to have a solution; that is one of the key things that makes it a puzzle rather than, for instance, an essay prompt. The solution is determined by the structure and constraints of the puzzle. In the case of a table-top jigsaw puzzle, there are infinite ways you could arrange the pieces but only one correct way to arrange them. With something like a chess game, it isn't the case that there is only one correct solution for each move, but the moves are still constrained by a clear structure and desired outcome. While the solution to a normal science puzzle may not be "real" or "preexisting" in a scientific realism sense, but it is real relative to the model of the world that is described by the puzzle.

For instance, within the geocentric paradigm operating during Ptolemy's time, his concept of epicycles was a proper solution to the puzzle-solving activity of the time. The existence of epicycles preserved the paradigm's commitment to Aristolean notions like geocentrism and uniform circular motion while also making correct empirical predictions (Kuhn 1962). However, while this solution was one which exemplified a good solution by the standards of the existing paradigm, Ptolemy's epicycles did not correctly capture an objective structure of the world, and are not a good solution when judged by the standards, methods, and theoretical commitments of contemporary physicists.

An example of this would be the addition of new elements like gallium and scandium to Mendeleev's periodic table (Noe, 2002, pg. 33). When Mendeleev proposed a periodic table in 1869 based on properties that appeared relative to the atomic weights of elements, he noticed that there were gaps in the table. He predicted that there existed elements with properties that would fit the pattern of the rest of his table. These elements were not yet known. This is a case of a Kuhnian discovery-what preceding a discovery-that. The properties that Mendeleev predicted for the as yet undiscovered elements turned out to be a good match for the actual recorded properties of scandium and gallium when physical samples of the elements themselves were isolated in 1879 and 1875. However, the filling-in of the periodic table that Mendeleev did was a good puzzle-solution relative to the existing knowledge and theoretical commitments of his time. He believed that elements' physical properties were determined by their atomic mass. Really, element's physical properties are determined by atomic number–the number of protons in the nucleus of the atom, which determines electron configuration. Atomic mass correlated with atomic number, so some of Mendeleev's predictions were correct. However, others were wrong, like Mendeleev's misordering of cobalt and nickel (Thyssen 2014). Mendeleev's puzzle-solutions were good solutions relative to his paradigm, but would not be acceptable to the standards of contemporary chemists.

Another example of puzzle-solving discovery is the discovery of new kinases by an IBM AI that mined research papers to find kinases that have potential to regulate p53 proteins and treat cancer (Simonite, 2013). The discovery process consisted of taking the existing conceptual framework of the papers and explicating new pieces that fit into the puzzle.

The AI system, KnIT, analyzed tens of thousands of research papers about a protein, p-53, which is known to play a role in suppressing tumors (Chen

2016). KnIT analyzed patterns in language used in papers about known kinases that help regulate p-53. By analyzing patterns in the descriptions of different kinases in the literature, KnIT was then able to make predictions about the p-53 regulating ability of new kinases that had been discovered since the date of the most recent research papers in the dataset it had been trained on. KnIT was able to correctly identify several new kinases that do help regulate p-53, by comparing the patterns discovered in the trained dataset to patterns it found in the description of these new kinases in the new research papers (Chen 2016).

KnIT's process resembles Kuhn's notion of a puzzle-solving discovery. The AI system absorbed and systematized patterns from the existing research on protein-kinase interactions, and then used these to extrapolate other resembling patterns that were implicit in scientists' research papers but not explicitly recognized by scientists. Scientists then tested these patterns to see if they were accidental correlations or meaningful connections, and found that several of them were meaningful. KnIT's contribution is analogous to Mendeleev employing the periodic law and existing periodic table to fill in blank spots and predict the existence of as-yet undiscovered elements.

Kuhn calls the tasks that most scientists perform during their careers in normal science "mopping up operations (Kuhn 1962, pg. 24)", which includes (1) gathering facts that the paradigm suggests reveal important ideas about reality or are especially informative; (2) testing and confirming the empirical predictions of the paradigm, (3) doing empirical work like finding physical constants and determining quantitative laws, and (4) extending the explanatory scope of the paradigm by applying the paradigmatic theories to new phenomena and using that new data to either revise and refine the paradigm's theoretical framework or subsume the phenomena under the theories of the paradigm (Kuhn 1962).

In summary, under Kuhn's account, normal science involves working

within a paradigm to apply existing theories, methods, and rules to new phenomena and in doing so refine and fill in gaps in the paradigm. This resembles filling in a puzzle, where the puzzle pieces are already in existence and must be combined in a way that accords with established rules, norms, and agreement about what an acceptable solution is. A paradigm constrains and sets standards for what solutions are acceptable, just as rules in a board game determine what moves and solutions are possible. The paradigm defines the puzzles that need to be solved in the paradigm, the standards for what good science should look like, the exemplars that acceptable puzzle-solutions should be modeled after, etc.

These rules include metaphysical commitments (for instance, particles versus waves, or gravity as a force versus curved spacetime), paradigm-sanctioned methods, and constraints on what questions can be articulated within in the paradigm and are considered important within the paradigm. This puzzle-solving is the sort of activity that AI excels at. Both neural networks and algorithmic AI operate as a tool that solve puzzles whose parameters are established by the input data that is fed to the AI system. In traditional algorithmic AI programs, scientists provide explicit inferential rules as input. In neural networks, scientists train the neural networks on large, curated datasets that reflect the relevant patterns that they want the AI to search for.

Kuhn's latter category, scientific revolution, involves what I will call "conceptual discovery", where anomalies in the observable data lead to a fundamental reinterpretation of what questions are relevant for scientists to pursue, what theoretica commitments scientists share, what methodologies are meaningful, and what puzzle-solutions are acceptable and meaningful. This is in sharp contrast to the normal science activity of filling in gaps in the existing paradigm using tools of the paradigm. Under the category of conceptual discovery fall paradigm shifts like the Copernican revolution, plate tectonics,

the discovery that genetic information is contained in DNA, and Darwin's theory of evolution. In Kuhn's view, "scientific revolutions are... those non-cumulative developmental episodes in which an older paradigm is replaced in whole or in part by an incompatible new one (Kuhn 1970, pg. 92)." Moments of revolutionary science can sometimes be difficult to separate from normal science; there is not a universal structure that seems to characterize revolutionary science, and revolutionary discoveries often emerge from and are founded on discoveries and scientific work that is not revolutionary in the Kuhnian sense. For instance, the paradigm shift to Einstein's general relativity from Newtonian mechanics restructured fundamental assumptions about gravity and spacetime, but was founded on tools, methods, and mathematical formalisms established within the Newtonian paradigm.

Because of this overlap, a bit more effort will be required in defining what I mean by conceptual discovery, as it is not directly equivalent to any term that Kuhn uses. Darwin's evolutionary theory, for instance, is an episode that I will argue counts as a conceptual discovery, yet it did not occur as a result of any type of scientific crisis, which is one of the characteristics of scientific revolution that Kuhn describes.

In response to the problems and vagueness with Kuhn's revolutionary science, some philosophers have offered alternative definitions. For instance, Casadevall and Fang (2016) say,

> We propose a definition of revolutionary science as a conceptual or technological breakthrough that allows a dramatic advance in understanding that launches a new field and greatly influences other fields of science. The Darwin-Wallace theory of evolution therefore qualifies as a revolution because it spawned the new field of evolutionary biology and profoundly influenced diverse fields, including anthropology, theology, sociology, and political science, soon after its publication in 1859. The discovery that

> DNA is the transforming principle of heredity and the subsequent elucidation of its structure also meet our criteria for revolutionary science because they launched the field of molecular biology while transforming the fields of genetics, medicine, and biochemistry(Arturo Casadevall, 2016, pg.2).

However, there are a number of ways in which their definition is not sufficient. First, their account seems to come close to a proposal to categorize scientific discoveries based on the way they are received by members of the scientific community. Second, there are cases where a puzzle-solving discovery might be considered a "technological breakthrough that allows a dramatic advance in understanding." To a materials scientist who is attempting to engineer a semiconductor or has some other vital need for an alloy with a low melting point, the discovery of gallium (which Kuhn and others have categorized as a puzzle-solving discovery) might very well be a technological breakthrough.

The characteristic of conceptual discovery that I argue makes it deserving of its own category is not whether it occurs during a time of scientific crisis or whether it qualifies as a conceptual or technological breakthrough. Conceptual discoveries might occur at a higher rate during periods of scientific revolution but I will not attempt to draw any equivalencies between Kuhn's revolutionary science and the type of discoveries that I will call conceptual discoveries. Instead, the most important feature of a conceptual discovery is that it involves a shift not necessarily in the empirical data itself but in the way the empirical data is conceptualized.

The best illustration of a conceptual discovery that I am aiming to describe comes from Kuhn's discussion of a Gestalt demonstration. He says,

> The subject of a gestalt demonstration knows that his perception

> has shifted because he can make it shift back and forth repeatedly while he holds the same book or piece of paper in his hands. Aware that nothing in his environment has changed, he directs his attention increasingly not to the figure (duck or rabbit) but to the lines of the paper he is looking at. Ultimately he may even learn to see those lines without seeing either of the figures, and he may then say (what he could not legitimately have said earlier) that it is these lines that he really sees but that he sees them alternately as a duck and as a rabbit(Kuhn, 1962, pg.114).

The key characteristic of conceptual discovery is the Gestalt shift; from existing lines forming a new picture and from existing scientific information forming a new paradigm. An indicator of conceptual progress being made is the appearance of mental models (Brewer 2001) and metaphors. According to Brewer, a mental model is "a conceptual framework that provides an explanation for a set of phenomena by postulating a structural relation to another more familiar concept (Brewer 2001, pg. 33). Similarly, both Kuhn and Lakoff and Johnson's work has suggested that the appearance of metaphors is also a characteristic of conceptual discovery. Kuhn discusses this, saying that the logic and puzzle-solving rules that characterize normal science become inapplicable in times of scientific revolution because the fundamental axioms and rules of a science often change during a time of revolution. In part, metaphors might arise because there is not yet language to talk about the new phenomena, and scientists must use the old terms metaphorically in order to express new ideas. Often the linguistic remains of the old terms then become embedded in the new theory, as I discuss in another paper.

While puzzle-solving discovery involves using an existing framework to extend and fill in the framework, conceptual discovery involves making decisions about what the framework should be like. This can involve proposing causal hypotheses to explain observed correlations. It can also involve

making decisions about how to handle data that is incoherent with the existing framework. When contradictory data comes in, scientists must decide whether to use the framework to interpret the data, or use the data to alter the existing framework. For instance, if astrophysicists discover an exoplanet that is violating established expectations of planetary motion, they must decide whether to propose auxiliary hypotheses that explain the anomaly (such as proposing that there is an issue with their measuring apparatus or that the planet has some special feature that explains its irregular motion) or whether to alter the existing laws and theories that govern planetary motion. This decision-making process is a conceptual one, because scientists must consult their goals, their theoretical commitments, their epistemic values, and their intuitions. Making decisions about how to revise existing theoretical frameworks is not governed by the rules contained within the theoretical framework itself.

A useful analogy to describe conceptual discovery is Kuhn's analogy between scientific and political revolutions. In a political revolution, one cannot use the values or institutions of the existing political system to justify moving to the new political system. Kuhn says,

> Political revolutions aim to change political institutions in ways that those institutions themselves prohibit. Their success therefore necessitates the partial relinquishment of one set of institutions in favor of another, and in the interim, society is not fully governed by institutions at all (Kuhn 1962).

When conceptual changes happen, whether political or scientific, the political or scientific framework is weighed against a broader theoretical framework. This broader framework contains things like social values, an agent's

overall body of epistemic commitments, and knowledge about what pragmatic functions the framework needs to serve. Conceptual discovery is distinct from puzzle-solving discovery because it does not follow pre-established rules or methodologies, but consults the broad framework of scientists overall beliefs, needs, and values in order to establish rules and methodologies.

It is useful to distinguish between these two categories of discovery because the former category, puzzle-solving discovery, is closely aligned with the current contributions that AI is making to science. The latter, conceptual discovery, appears to be more aligned with uniquely human abilities. I will argue that AI can make puzzle-solving but not conceptual discoveries. If AI can participate in conceptual discovery, it will do so with limited autonomy; it will be strictly piloted by scientists. These two categories, puzzle-solving and conceptual discovery are not mutually exclusive; rather, they represent two ends of a spectrum describing different ways that scientific discovery can occur. In section 2, I will examine several examples of computer-driven scientific discovery and discuss how they fit into these two categories I have explicated.

1.4 Logic of Discovery

I now come to the question about the parameters of artificial intelligence's ability to contribute to scientific discoveries. Before addressing this question, I will first explore the notion of a logic of discovery.

Prior to Popper, philosophers like Descartes, Bacon, and Mill subscribed to the idea that discovery could be formalized as a set of rational principles and methods, and that a goal of philosophy should be to uncover these rational, truth-tracking principles.

Under this picture, scientific discoveries are instances where rational, methodological, schooled inquiry suddenly pays off, not lucky instances that unpredictably strike individuals. One of the motivations for the search for underlying laws of discovery was the observation that certain individuals, like Einstein and Darwin, seem to be especially good at making discoveries and it seems intuitive that this skill must hinge on a particular quality of their minds (Alai 2004, pg. 1). Uncovering the methodology that gives them this ability would be productive for science because it could provide epistemic norms for scientific reasoning. The goal of philosophers like Bacon, Descartes, and Mill was to uncover normative principles or a rational logic that characterize successful scientific discovery.

Some discoveries seem highly spontaneous and coincidental, like Archimedes' 'eureka!' moment about water displacement upon getting into the bathtub (Biello 2006). Furthermore, Kuhn showed that the notion of scientific progress and what counts as a scientific discovery is paradigm-dependent, and as paradigms shift, so do scientists' beliefs about what methods are epistemically virtuous and which scientific discoveries should be treated as exemplars. If there were a logic of discovery, then discovery would be reduced to following the rules of that logic, and scientific research would be a much more measured and steady process, rather than making leaps and jumps of success and experiencing paradigm shifts (Alai 2004). At the very least, there would be clear qualities and methodologies that make one scientist's reasoning, or scientific method, better than another. Instead, norms for scientific discovery appear to be context-dependent and incommensurable (Kuhn 1962).

So there is a kind of predicament; on one hand it seems that there must be methods and norms that characterize virtuous scientific discovery, but on the other hand, there are a number of reasons to believe that there cannot be universal normative principles that govern the process of scientific discovery.

That there is a logic of discovery is no longer a widely supported view in epistemology. In *The Logic of Scientific Discovery*, Popper famously argues that

> the act of conceiving or inventing a theory seems to me neither to call for logical analysis nor to be susceptible of it. . . the question how it happens that a new idea occurs to a man—whether it is a musical theme, a dramatic conflict, or a scientific theory—may be of great interest to empirical psychology; but it is irrelevant to the logical analysis of scientific knowledge (Popper 1959, pg. 31).

This quote generally captures contemporary philosophy's view on the notion of a logic of discovery. Following Quine's advocacy of a naturalized epistemology, discussion on the nature of scientific discovery have come from psychology more than from philosophy.

Yet, as AI systems contribute more to processes of scientific discovery, it seems that questions about a logic of discovery should make a comeback in order to investigate the scope of what components of the discovery process *can* be formalized. The inauguration of AI as an essential contributor to scientific research necessitates a new question: 'To what extent can a formal, computational system make scientific discoveries?' For a computer to make discoveries requires a formalism of the discovery process.

Reichenbach says that the goal of epistemology is to construct "thinking processes in a way in which they ought to occur...or to construct justifiable sets of operations which can be interrelated between the starting-point and the issue of thought processes (Reichenbach 1938)."

Reichenbach distinguishes between discovery and justification, claiming that the above-mentioned aim can be achieved for the context of justification, but not for the context of discovery; justification can be subjected to logical,

normative analysis, but discovery can only be understood as a psychological process. Reichenbach says,

> The act of discovery escapes logical analysis; there are no logical rules in terms of which a "discovery machine" could be constructed that would take over the creative function of the genius. But it is not the logician's task to account for scientific discoveries...logic is concerned only with the context of justification (Reichenbach 1956).

Popper similarly says,

> My view of the matter...is that there is no such thing as a logical method of having new ideas, or a logical reconstruction of this process. My view may be expressed by saying that every discovery contains 'an irrational element', or 'a creative intuition' in Bergson's sense (Popper 1959).

If AI has the capacity to produce laws, explanatory models, and theories, then this would suggest the existence of a "logic of discovery", a set of normative rules that characterize certain successful cases of discovery. This would not be an exhaustive logic of discovery, because AI only surpasses human scientists in certain narrow contexts of discovery. However, even though AI's contributions to science are narrow, the contributions are still cases in particular contexts in which a "discovery machine" is generating fruitful hypotheses at a faster rate than human creativity is able to. AI systems, even neural networks, rely on a formal architecture, and thus if AI is able to make scientific discoveries then this would indicate that some processes of scientific discovery can be formalized.[1]

[1]Even though AI sometimes provides that appear creative, and even though neural networks are a black box, the underlying mechanism powering AI is still formal and deterministic.

The introduction of AI is creating a chance for an abstract philosophical question to be investigated empirically, through the investigation of the *de facto* discoveries that AI is in fact making (Alai 2004). If this provides evidence that a logic of discovery is possible, the “logic of discovery" will not be a logic in the Reichenbach sense of a universal set of normative rules that govern discovery, but in the pragmatic sense of a set of computational formalisms that produce useful new hypotheses in practice.

1.5 Overiew of AI systems contributing to scientific discovery

Chess has often been used as a tool to experiment with the abilities and methods of artificial intelligence. It provides an environment with bounded conditions: 64 squares, 32 pieces, 6 types of pieces, and a set of fairly simple rules. Despite the simplicity of the setup, there are many different possible game configurations. Even ignoring the possibility of generating infinite games by never declaring a draw, conservative estimate put the number of possible plays at 10^{120} (Shannon 1950). Questions of how to model these possible plays, and what characteristics make a chess player successful (whether human or computer) are very relevant to eventually asking similar questions about larger systems of laws and configurations, like physics. Many research programs consider chess to be a drosophila of artificial intelligence and human reasoning (McCarthy 1997). There is a wealth of research on the psychology of human chess players and the mathematical dynamics of the game. This makes it a good medium for comparing human reasoning with computational inference.

It is significant to note how differently AI and humans approach the task of winning at chess. Some psychologists use the computer as a metaphor

for how the human brain functions, and in many cases this comparison is a useful heuristic tool. In this next section I will explore some of the ways that this comparison falls apart. It was only a decade ago, in the early 2000s, that AI chess programs like Deep Blue began consistently winning against human chess champions (Kasparov 2010). However, Deep Blue can compute one billion moves per second while psychology studies suggest that the best human players are only capable of around 500 (Hartston 1996). So, assuming that computing power and some other characteristics together determine how good a chess player is, it seems that Deep Blue is qualified in computing power and underqualified in something else, while humans are underqualified in computing power and qualified at something else. Chess-playing AI programs exhibit an interesting mixture of intelligent and unintelligent behavior. For instance, in some games, Deep Blue has made moves which seem very much like short-sighted blunders that a top human player would not make whereas in other games Deep Blue has played as though it had an intuitive understanding of the layout on the board. In one game against the AI, Gary Kasparov said that only a human grandmaster would be able to play the winning game that the AI had played, but the move it had made that he praised as being psychologically strategic ended up being due to a bug that caused Deep Blue to accidentally choose a non-optimal move.

Deep Blue is a traditional algorithmic AI program, in which it has been programmed with explicit heuristics rather than evolving its own model through reinforcement learning as a neural network like Alpha Go does. Thus, Deep Blue's success is a function of its processing power plus the efficacy of the explicit rules in its programming. In many cases, these rules plus its processing power allow its evaluation of moves to far surpass a human player. The cases where it exhibits seeming irrationality are instances where the rules selected by its programmers fail to be optimal for a particular board

configuration. For instance, Deep Blue might be programmed to avoid configurations that put its king at risk, but sometimes a move that is non-optimal with regard to risk can be optimal with regard to the psychology of the opponent if the move is unexpected or follows a pattern that is non-standard in chess.

In illustrating the above differences, chess can serve as a useful model for investigating some aspects of how AI's contributions may be similar to or differ from humans' in the realm of scientific discovery. However, chess is not representative of all discovery contexts. In chess, the rules of the game set clear parameters on what moves are possible. There is an extremely large, but still finite set of possible games that can be played, and some of these game plays are probabilistically better than others relative to the goal of winning the game. Even though there is not objective, foundational justification for a given move that DeepBlue makes, because the success of each move depends on the psychology of the human opponent and their ability to successfully respond to the move, there is objective data about which board configurations satisfy the objective heuristics that have been programmed into Deep Blue. For instance, part of Deep Blue's programming includes weighting of the different chess pieces' importance, with the queen and king having a high weight relative to a pawn, for instance. Deep Blue's programming includes a rule which tells it to prioritize protecting the higher-weighted chess pieces. Even non-algorithmic, deep learning AI systems like Alpha Go operate within set parameters and relies on the simple heuristic of making the move that has the highest probability distribution.

In contrast, scientific discovery involves both setting the parameters and conducting searches within those parameters. Scientists both formulate the problem, set the range of acceptable answers to that problem, and iteratively alternate between searching within those parameters and altering the parameters.

In the following section, I will examine several examples of AI programs either designed to replicate past discoveries or used in the pursuit of contemporary theory generation. For each of these examples, I will analyze the extent and nature of AI's ability to contribute to the discovery process.

Before deep neural network technology advanced in the 2010s, most AI was algorithmic and operated on the basis of explicit programmed rules. Many of the algorithmic AI systems in science were aimed at replicating the discovery of empirical laws. Under this category fall programs like GLAUBER and BACON (Langley et al. 1985). These programs look for numerical relationships between variables. For instance, GLAUBER takes as input a set of objects and predicates about particular chemical properties, and from this extrapolates abstract categories like 'salt' and general laws about those categories (Langley 1985). For example, GLAUBER outputs generalizations like the fact that an acid and a base neutralize each other in the form of rules like 'For all x, if x is an acid then it reacts with sodium hydroxide (NaOH) to form a salt.'

These AI systems are essentially pattern-recognition machines, and their reliance on inductive reasoning reinvigorates questions about the nature of inductive reasoning in science. Is inductive reasoning captured by the kind of activity that GLAUBER engages in?

Hempel argues that the view that theory can be objectively extrapolated from observation is misguided. He characterizes this as the 'narrow' view of induction, which can be attributed to accounts like the following:

> If we try to imagine how a mind of superhuman power and reach, but normal so far as the logical processes of its thought are concerned...would use the scientific method, the process would be as follows: First, all facts would be observed and recorded, without selection or a priori guess as to their relative importance. Second, the observed and recorded facts would be analyzed, compared,

> and classified, without hypothesis or postulates other than those necessarily involved in the logic of thought. Third, from this analysis of the facts, generalization would be inductively drawn as to the relations, classificatory or causal, between them. Fourth, further research would be deductive as well as inductive, employing inferences from previously established generalizations (Wolfe 1924).

Hempel argues that views like that described above fundamentally misunderstand the relationship between theory and observation. First, he points out that scientists do not impartially gather empirical facts; their background knowledge and intuition necessarily guides the method and selection of evidence. Second, he points out that even selection of a question to investigate does not sufficiently narrow the method and scope of data collection:

> The question as to the causes of lung cancer does not by itself determine what sorts of data would be relevant-whether, for example, differences in age, occupation, sex, or dietary habits should be recorded and studied (Hempel 1966).

Scientists must rely on their local knowledge and background assumptions in order to select relevant data. Hempel concludes, theories and hypotheses are “not mechanically inferred from observed “facts"; They are invented by an exercise of creative imagination (Hempel 1966).”

Popper, too, takes the stance of denying that inductive inference is a philosophically meaningful notion, saying, "Induction, i.e. inference based on many observations, is a myth. It is neither a psychological fact, nor a fact of ordinary life, nor one of scientific procedure (Popper 1963, pg. 53)." His support for the claim that induction is a myth comes from two premises. The first mirrors Hempel’s: he claims that observations themselves require and

implicitly have a theoretical framework, and thus it is not possible for a person to start with a set of observations and objectively generate a theory from them.

This observation makes salient the fact that, in cases of the inductive AI systems like GLAUBER, a theoretical framework is being fed to the AI system as input. The data that GLAUBER takes as input is already selected and represented by scientists, and thus includes a theoretical framework. GLAUBER does not infer from observation to hypothesis; it infers from pre-programmed theoretical framework and structured, selected data to hypothesis. This is a rote puzzle-solving activity.

Popper's second and more famous objection is that acceptance of theories should not be based on instances that confirm the theory. Popper proposes that the problem of induction should be resolved by scientists instead starting with hypotheses, making falsifiable predictions from those hypotheses, and working to refute those through observation. GLAUBER is unable to follow Popper's refutation model as it has no capacity to differentiate between disconfirming evidence and absence of evidence (Langley 1985, pg. 7). In contrast, chess-playing AI programs like Deep Blue use a search tree in which they proceed down a path in the tree and backtrack if they encounter a falsifying result on that path. The way that this path-following and backtracking is structured, it essentially amounts to a number of conjectures being tested and refuted until an optimum conjecture is settled upon (Bolc, 2012).

In contrast to programs like GLAUBER that aim to replicate historical discoveries in science, some AI programs that have proved fruitful in contemporary discovery are those that automate lab experiments. Chemists at the University of Glasgow have made several significant discoveries in chemistry by using AI programs to automate chemistry experiments. Designing AI systems that control the running of the experiments allows more time for human scientists to focus on larger conceptual aspects of the problem

they are working on. In addition, it has been found that having AI run the experiments reduces research costs because the AI programs are on average better at deciding which experiments to run, leading to less wasted time and materials. In addition, AI is now being used to read through academic papers across broad or disparate fields and uncover information that helps scientists decide what new hypotheses to consider and test experimentally. One example of an AI that has been used to automate experiments is the 'Robot Scientist', Adam, developed in a collaboration between researchers at Aberystwyth University and Cambridge University in 2009. The program is comprised of both robotics (arms, grips to clasp flasks and equipment), and a set of four computers which forms the computing mechanism. The robot can design experiments based on hypotheses that it has come up with and physically test the experiments. A Stanford AI researcher observed that the Robot Scientist was operating at the level of a graduate student. The Robot Scientist is said to be the first machine to make a new discovery in science independently of human scientists. Of course, as I discussed earlier the structured nature of the input data means that scientists' background assumptions factor into the output, even if the internal component of the inference process is performed by the AI independently. Adam's discovery was new information about genes in the genome of *Saccharomyces cerevisiae*. The program independently generated a hypothesis, designed experiments to test the hypothesis, and then physically conducted those experiments using its inbuilt robotics, and developed a conclusion about the results. The researchers then tested the conclusion to confirm that it was both correct and novel scientific information.

A final category of discovery task that I will discuss before I move on to the question of what type of discoveries AI can make is data mining. Given the ever-expanding breadth of scientific publications as the population of scientists grows much more slowly, there is a niche for computer programs that

can go through and find statistical connections between research papers. For example, in medicine, correlations between health problems and medications can be examined to discover side effects, and in medical research correlations between certain genes and diseases can be captured. These data mining AIs usually work on a machine learning framework, either a neural network which is trained on data sets or a more independent genetic algorithm.

One example of a data mining AI is the Warmr program. Unlike most data mining programs, Warmr falls under the category of inductive logic programming, which is a sort of hybrid of machine learning and logic programming. Inductive logic programming is structured as follows: (1) The program is provided with background knowledge and a data set comprised of positive and negative examples. The program will then create a hypothesis in the form of a set of logical propositions from which the positive examples follow but the negative examples do not. It differs from many standard data mining programs in that it captures more complex relational structure between data points rather than simple associations between individual data points. In capturing more abstract relational mappings, Warmr is analogical rather than simply associative. In humans, analogical reasoning works by noticing relational structure and generalizing it, and in doing so sometimes ignoring more simple direct associations. Warmr was used to make hypotheses about which chemicals are carcinogenic for rodents using an existing database of background knowledge and examples (King 2001).

Chapter 2

AI Can Make Puzzle-solving Discoveries

2.1 Existing AI Programs Make Puzzle-solving Discoveries

When headlines describe new scientific discoveries that an AI has made, the AI being described is often a supervised machine learning algorithm (Hosanger, 2019). Supervised AI programs work with labeled datasets where each input is already associated with an output, and the AI program analyzes patterns between the input and output data to generate a function that can make predictions for new inputs. A supervised AI program works by taking a database, say, of patients taking prescribed pharmaceutical drugs, the patients' known preexisting medical conditions and general information about their health, and the health problems that they are diagnosed with from the time of taking the medication through their life. The algorithm would process this large dataset, and would compute functions that can predict longterm side effects of different medications.

Overall, this is an example of AI performing a pattern-recognition job that humans can do, but much faster and more efficiently. Automating pattern recognition tasks has a number of potential benefits, such as allowing

scientists to focus on conceptual tasks and also simply by conducting more pattern-recognition analysis than human scientists are practically capable of doing. The patterns that AI uncovers in such tasks is a discovery in the sense that it was previously unknown. AI finding that certain types of hormonal birth control correlate with a higher rate of cancer would likely be labeled a "discovery" by newspapers. How would philosophers classify it?

For Kuhn, it seems that it would constitute a discovery-that. For Kuhn, a discovery-that identifies that a phenomenon or correlation exists, without correctly conceptualizing the nature of that phenomenon or relationship. AI's discovery that there is a correlation between certain types of hormonal borth control and increased incidence of cancer would establish that a relationship may exist, but not what the nature of that relationship is. Discovery-thats are usually followed by discovery-whats, like chemists trying to conceptualize the chemical mechanism underlying combustion.

So while an AI-discovered correlation between certain hormonal birth controls and cancer would arguably be new information (even if implicitly contained in the input data that was fed to the AI program), it would not count as a discovery-what under Kuhn's account. A discovery-what would established by scientists providing a theory that accurately explains the correlation. Another feature of supervised machine learning that constrains the scope of its contribution to discovery is that the input data is cleaned up and labeled by researchers before it is presented to the AI program. In this sense, the task is mechanical: you could replicate it by putting items of a certain shape on a conveyor belt and having them tumble down into a basket if they fit a certain shape criterion. Supervised AI is provided with labeled blocks and tools for reading block shape. Using the training dataset, it adjusts its model to minimize prediction error. The model that AI generates for the block-sorting task is entirely determined by the initial architecture that it is programmed with and the contents and labeling of the training dataset.

In contrast to supervised and reinforcement learning AI, unsupervised AI programs uncover patterns in datasets without direct human intervention in the form of labeling and positive/negative reinforcement (Raza 2018). The notion of relevance that the AI uses to build the model comes not from explicit labeling or positive reinforcement, but from the input data itself and the AI's similarity metric. In these cases, the notion of relevance that is implicit in the initial dataset's representation and the AI's similarity metric shapes the types of patterns that the AI identifies. The choice of data and representation of data in the initial dataset thus influences what types of data clusters AI generates.

Thus, unsupervised learning is mechanical in the same sense that supervised reinforcement learning is. Scientists' background assumptions are implicit in the initial selection and representation of input data as well as the AI's similarity metric, and the AI model merely extrapolates further conjectures from those theory-laden representations and similarity metric.

Another category of AI programs are unsupervised reinforcement learning programs. There are now various AI programs, such as Google's Alpha Go (Hosanagar 2019), that don't have labeled data sets generated for them by researchers, but instead do their own labeling of data based on criteria they iteratively extrapolate from positive reinforcement on training sets. For instance, while Deep Blue has explicit rules in its algorithm that its programmers developed in collaboration with expert chess plays, Alpha Go instead plays against itself to generate its own model. These types of AI programs are not limited to playing board games; they have been used in chemistry and the pharmaceutical industry.

One such AI program is the 'AI Physicist' developed at MIT (Wu 2019). The AI Physicist is designed to imitate the methods human physicists use to discover physical laws. One of the biggest strengths that makes the AI Physicist unique in comparison to other AIs is that the it is designed to look for

multiple small theories rather than one large theory. This solves one weakness that other AIs face; many AIs when given a large body of data look for a single theory that explains the whole data set. With large complicated data sets like virtual worlds with physical laws, searching for a single generalizable theory can be inefficient because the data results vary in different contexts. For instance, (Wu 2019) points out that the dynamics of a double pendulum are most efficiently formalized by separating the complex system into two separate domains, the upper and lower portions of the pendulum. The AI Physicist took the double pendulum's trajectory as input and was then able to separate the dynamics into two separate domains, one domain describing the dynamics of the upper portion and one describing the dynamics of the lower portion of the pendulum (Wu 2019).

The AI physicist is also programmed to include epistemic values like Occam's razor, theory unification, and predictive power. These principles guide the direction of its model generation. It relies on reinforcement learning and remembers solutions that have worked in the past and applies these to future datasets. The developers tested the AI physicist on forty artificial worlds with unfamiliar laws of physics (Wu 2019). Their aim was to see if the AI would be able to discover the laws of physics for each world by simply observing the phenomena of each world. The AI physicist was able to derive accurate laws for 92.5% of the tested worlds.

The AI Physicist's weighing of epistemic values is a strength of the program. An important component of automated discovery is to formalize the ability to derive multiple coherent theories from one dataset and to be able to judge how to conduct theory choice based on principles that scientists use. If the phenomena of a physical world can yield either two complicated theories that are limited in scope or four more small but simple and explanatory theories, a good AI needs to be able to make a judgment about whether to choose the two theories or the four theories to represent the world, and weigh this

decision against other epistemic values.

Aside from the numerous practical applications of AI to discovery processes, there has been extensive research conducted by Langeley et. al. (1985) that aims to design AI systems that can reproduce historical discoveries in science and mathematics. This research is focused not on developing useful ways to use AI, but on demonstrating that AI can replicate historical discoveries. Pat Langeley (1985) explores several examples in her 2001 work *The Computational Support of Scientific Discovery*, and joins with Herbert Simon and others to publish *Scientific Discovery: Computational Explorations of Creative Processes* (1987). Their program has the goal of showing that discovery is a rational process that can be formalized. The first of their programs was named "BACON" after Francis Bacon, who is often called the father of empiricism and is a notable advocate of mechanizing the scientific method.

Langeley et al.'s computational discovery project aims to design AI programs that engage in "heuristic selective search (Langley 1987, pg.5)." Heuristic selective search uses programmed heuristics to selectively narrow down the number of possible solutions to explore in a large problem space. A problem space is all of the possible arrangements of data within the constraints of the domain. For a program searching for the constituents of a chemical reaction that forms water, it might be all of the plausible chemical pathways that are allowed by the initial setup that the scientists have designed. In chess, the problem space is all of the ways that chess pieces can be configured on a chess board over n possible games, where n equals all possible games (Shickore 2018).

The problem space has a set of states; each state is one way that the constituents of the problem space can be arranged within the problem space. A random layout of chess pieces on a chess board would be one state of the problem space. The initial state of the problem space is important because it determines the future states, and the goal state is important because

it constrains what intermediary states are viable. The programs are programmed with heuristics that guide the selective search process. This is a puzzle-solving activity because the goal is to solve the puzzle of how to start with the initial state and get to the goal state. It is a kind of orienteering process. The path is constrained by operators which dictate how new states emerge from the prior state, and path constraints which act as roadblocks to make the number of available states a finite number (Shickore 2018). To reach the goal state, the program applies the operators to the initial state, and the resulting states from that state, until it has eventually explicated all of the possible states within its search space and found an optimum path that leads from the initial state to the goal state.

Langeley (2000) claims that scientific discovery can be broadly sorted into five types of process. First, the generation of taxonomies serves to group things into basic categories so the categories can be sorted or manipulated. Elements are sorted into gases or metals, organisms are sorted into genus and species, and notions of symmetry specify groupings in physics. Then comes the proposal of qualitative laws, which determine interactions between categories such as interactions between certain classes of chemicals. Another stage is the generation of quantitative laws. Finally, structural and process models, in which descriptive accounts are expanded into deeper mechanistic or explanatory models. Langeley claims that each of these five stages have been achieved with AI programs (Langeley 2000). The remainder of this section will summarize the progress of Langeley et al.'s programs.

First I will discuss AI that generates useful taxonomies, one of the simpler stages of scientific discovery according to Langeley. Astrophysicists work to sort stars into taxonomies by the characteristics that are accessible to us from the visible light they emit, which is captured by satellites like IRAS. The AI system AutoClass developed by Cheeseman (1988) was applied to the data from IRAS to see if it could sort data into useful taxonomies (Langeley 2000,

pg. 398). The system puts objects into classes depending on the random pre-established definition of each class. It then updates the descriptions of the classes based on the objects that it has added. The program adds classes to the initial programmed number of classes until classes with minimal variation exist (Langeley 2000, pg. 398). This process yielded 77 classes of stars, which were then sorted into more high-level clusters by repeating the automated sorting process on the class descriptions themselves. Because the result was different from the one commonly used in science and the results seemed useful (Cheeseman 1989), the results were published in an astronomy journal. The main points of interaction between the scientists and the AI was when the scientists structured and cleaned up the data before feeding it to the system, ran the program a second time on the classes themselves, and interpreted the results.

The program that Langeley et al. have developed to evaluate qualitative laws is called GLAUBER. It works by taking qualitative facts and outputting a set of classifications and some rules about the classes and their members. For instance, chemical classifications were historically sorted based on their observable attributes: salts dissolve in water, acids taste sour and dissolve metals, etc. In addition to this, they were evaluated on how they interact, such as salts being formed when acids interact with a base (Langeley et al.). GLAUBER replicates these discoveries and is named after the chemist, Johann Glauber, who helped discover the interactions between acids and bases.

GLAUBER receives as input sets of predicates, paired attributes, and values. GLAUBER sorts through its database and searches for cases where things have the same predicate and value for an attribute (Langeley et al. 1985). When GLAUBER discovers similarities, like multiple objects having the attribute "tastes sour", it creates a category labelled with this common attribute and stores the similar objects in the class. GLAUBER would notice that the sour-tasting class and the sodium hydroxide reacting class share the

same attributes and would be combined in a class together, one stored under the taste category and the other under the reacting category. This would eventually yield the chemicals sorted into the categories that we recognize ourselves.

Langley et. al. claim that GLAUBER's results are essentially the results that early chemists came to themselves when they found the patterns of salts and acids (Langeley et al. 1985). GLAUBER explicates the pattern in which acids and alkalis react to form salts, which is essentially the same result that early chemists achieved. Furthermore, when presented with more data, GLAUBER comes up with the more complex result that there is a category of elements, bases, that react with acids to form salts (Langley et. al.pg 466). If using empirical data and coming up with a qualitative law is sufficient grounds for a discovery, then GLAUBER does seem to be capable of replicating this early discovery in chemistry.

To prove that AI can replicate the discovery of quantitative laws, Langley et. al. set one of their BACON programs up to prove Proust's law of constant proportions, Dalton's law of multiple proportions, and Lussac's law of combining volumes (Langeley et al. 1985). The three laws were precursors to the atomic masses of the elements being computed. The first law states that for a chemical compound X, say water, the ratio of the component elements relative to one another is a constant for that compound X. The second law says that when two elements form different compounds together, their combined weight always equals some multiple of their individual weights. The third law says that gases combine in ratios that are integers when applied to their volume.

The program BACON searches through a data set to find which variables are dependent on each other, and in which way. Taking one independent variable at a time, it computes through and specifies the relationship between other variables and that variable. It then moves on to another independent

variable and repeats the process, this time also seeing whether the variable relates to the conclusion that it drew with the last variable. In the case of the research that led to the above laws, the independent variables were the elements and chemical compounds, and the dependent variables were the ways of measuring those elements and compounds: weight and volume. BACON examines a dataset in which it can see the ingredients and effects of a chemical reaction and their weights relative to one another. Different compounds with nitrogen (N) and oxygen (O) as their constituents would yield different weight ratios for weight of O divided by weight of N: NO with respective weights of 1.14 and 1 would yield 1.4, N_2O with weights of .57 and 1 would yield .57, NO_2 with 2.28 and 1 would yield 2.28. BACON assigns these numbers as intrinsic properties of the individual compounds (Langeley et al. 1985). Next, BACON would note that the ratios all are divisible by .57, yielding the integers NO = 2, N_2O = 1, and NO_2 = 4 (Langeley et al. 1985).

This is equivalent to Dalton's law of multiple proportions (Langeley et al. 1985). Carbon forms either CO or CO_2 by combining with oxygen in different proportions. X grams of carbon can react with either Y grams of oxygen or 2Y grams of oxygen. The two options for grams of oxygen that can react with X grams of carbon form the ration 2:1. Dalton's proposal was that this result represents the fact that one carbon atom can react with either one or two oxygen atoms. BACON's discovery process is essentially inductive reasoning over a data set.

The discovery system DALTON, named after the chemist John Dalton, creates structural models of molecular reactions. It takes into account two numbers; the number of molecules of each type and the number of atoms in each molecule (Langeley et al. 1985). Not knowing how many of each molecule there are, it begins with one of each molecule and proceeds from there. If the model faces problems later on in the process the AI will go back

and alter the initial assumption about the number of molecules.

The information that DALTON receives from the human scientists is akin to that which a chess-playing AI receives about the rules of chess and the desired outcome. DALTON is told that it can not violate certain laws, for instance the law of conservation: that mass cannot be created or destroyed in a closed system despite any chemical reactions or changes of state. Furthermore, DALTON is programmed with information about the solution it is searching for. Langeley says, "The conservation operator tells Dalton that the water molecule must be composed of one h particle and one o particle, and that the final model must have the form ((h)(o)$\longrightarrow$ (ho))(Langley pg. 457)." The solution in a sense already exists, and DALTON must search through the possibilities to find an option that does not violate the constraints that it has been programmed to reject. DALTON operates on a search tree framework, in which searching involves going down all of the different paths within the problem space, constrained by certain search heuristics, and backtracking every time it reaches a dead-end (Langley 1985, pg.457). The process that Langley describes is identical to a robot mechanically walking a maze and placing a do-not-enter marker over each path that ended up yielding a dead end. This process is an effective way to solve a puzzle. But is it a scientific discovery?

There has been much philosophical discussion about the rationalist account of scientific discovery proposed by Langley and Simon. Their account treats discovery as problem solving that occurs with the aim of finding numerical relationships within a data set. Langeley et al. say, “The research is mainly limited to finding a set of mechanisms that is sufficient to account for discovery (Langley et al. 1987, pg.4).”

Further, Simon says,

> The claim, then, that the processes of scientific discovery are normal problem-solving processes is a claim that scientific discovery follows the four principles...First, its basic method is selective (heuristic) search. Secondly, it uses both general and domain-specific heuristics. Thirdly, means-ends analysis, a heuristic of broad applicability, plays an important role in analysis and reasoning. Fourthly, effectiveness in discovery depends heavily on processes of recognition, making use of tens of thousands of productions that index memory with familiar and recognizable cues characterizing common problem situations (Simon, 1992, pg.8).

One piece of the evidence they provide for the argument that the programs are performing discovery is that the programs are capable of coming up with new, unobservable properties, and that this amounts to theorization. Version four of BACON, for example, comes up with values for the conductance of different wires, but there is no instrument which can directly measure conductance(Pat Langley, 1987, pg.129). The argument is that BACON is discovering processes which are not directly measurable/observable and are thus are theoretical, and therefore BACON is generating theory.

Langeley et al. say

> On the basis of the experience with these programs, it seems reasonable to claim that the mechanisms of scientific discovery can, indeed, be subsumed as special cases of the general mechanisms of human problem solving. To be sure, there may be essential novelty hidden in those aspects of problem solving that lie outside the range of the programs. But given the evidence of behavior that we have reviewed here, bare claims that such novelties exist are not convincing. There seems to be no present reason to believe that

> any aspects of scientific discovery must remain indefinitely beyond the powers of heuristic search, or that the discoveries of human scientists cannot in time be explained within the information processing paradigm for problem solving (Langeley et al. 1981).

In the case in which discovery is being defined as puzzle-solving discovery, this seems acceptable. Under the puzzle-solving definition, discovery means searching for an acceptable solution within a structured framework, akin to correctly putting together a jigsaw puzzle. However, the functions performed by BACON, GLAUBER, and DALTON do not seem to qualify as cases of conceptual discovery. BACON, for instance, works to rediscover Ohm's law, which

> ...relates the current I of an electric circuit to its voltage V and its resistance R. The law may be stated as $I = V/R$. In physical terms, the voltage is associated with the battery used in the circuit, while resistance is associated with the wire (Langeley 1980).

When computing this relationship, BACON discovers properties that are implicit in the input data and it does not propose an explanatory theory. BACON does not provide any knowledge about the nature of conductance nor what causes the different wires to have different conductance abilities. Furthermore, although conductance itself is unobservable in this case, the information about conductance that BACON generates is derived directly from observable properties that are provided as input. BACON is not really proposing a new, unobservable property but is just isolating a constant which already exists as a direct consequence of the observable data (Alai 2004, pg. 24).

Again, with Langley and Simon's program Stahl, it seems that Stahl's results can only be counted as scientific discovery when discovery means puzzle-solving, not conceptual, discovery. Mario Alai (2004) argues that

while Stahl yields a correct interpretation of the oxygen reaction, it would be impossible to conclude that Stahl is actually discovering oxygen theory. Alai says "Stahl...simply applies [oxygen theory]: it accomplishes what Kuhn would have called normal science tasks (Alai, 2004, pg.5)." Alai characterizes this as a simple and mechanical inference process. In contrast, discovery that occurs outside normal science restructures paradigms rather than performing mechanical deduction tasks within a paradigm.

I will discuss an overview of some objections that have been raised against Langeley et al.'s claim that the AI programs above are making scientific discoveries. Critics claim that Langeley et al.'s account is inadequate because it disregards the social, historical, and psychological elements of discovery. Taking these factors into account creates a complex picture which Langeley et al's account lacks. Another criticism, coming from Donald Gillies, is that Langeley and Simon's programs replicate discoveries that have already been made by scientists but have not managed to discover any new laws(Gillies, 1996. Furthermore, he points out that the programs have not yielded and results which are actually useful to science.

Another objection, which I touched on earlier in this paper, is that discovery is thought to rely on scientific intuition and creativity, and intuition and creativity seem incapable of being formalized. Simon responds to this by claiming that intuition involves scientists using their background knowledge and experience to recognize a new fact. He argues that the apparent mysteriousness and spontaneity of intuition is not due to the process itself being mysterious and unformalizable, but instead is merely due to humans lacking epistemic access to the contents of their reasoning processes. Simon says,

> When pressed for information about the method of solution, the respondent may reply, "I just used my intuition", or "it's based on my experience", or "the answer just came to me". There is no

> reason to doubt the truth of these replies. They are just what we could expect if the solution were obtained by an act of recognition: that is, if some cue in the stimulus situation evoked a recognition of something already familiar in the mind of the respondent, and thereby gave access to information previously stored in memory. In my paper I show that recognition is a well-understood process in psychology that has been simulated effectively by the EPAM program (Feigenbaum & Simon, 1984). It is also well known that a person can report what he or she has recognized, but not what features of the stimulus allowed it to be discriminated from other possible stimuli. The discrimination process is subconscious, hence not reportable. Recognition is 'intuitive', or better, intuition is simply recognition (Simon 1992).

Simon claims that processes which we tend to refer to as "intuition" are logical processes of recognition that can in theory be modeled computationally.

Another objection against Langeley and Simon is that if their account purports to resolve the problem of induction, the impossibility of providing foundational justification for any inductive law, then their account is misguided. Simon (1992) replies that the goal of their account is not to provide an ultimate logic of discovery that yields true, infallible theories. Rather, their goal of formalizing discovery is to replicate the fallible human activity of coming up with laws that fit the data that is observable at a given time.

Under Langeley and Simon's account, scientific inductive reasoning is reduced to the generation of taxonomies, qualitative laws, and quantitative laws. These tasks are not so much concerned with theory; rather, they are the fleshing out of details within a theory. Langley claims that the fact that some of their programs find laws about unobservables means that there is an element of theorization that the programs achieve. However, Gobet et.

al. (2017) argue that the discovery of qualitative and quantitative laws can include discoveries about observables or unobservables but is still a more or less separate category from theory (Gobet et. al. 2017, pg. 4).

For Gobet et al., "A theory is an underlying explanation, accounting for a set of observations by means of a causal process. For example, Newton's theory of gravitation explains Kepler's observations by means of a deeper, causal principle. The theory of evolution by natural selection, conceived by Darwin and Wallace, explains a wide range of observations regarding organisms' adaptations, and forms the basis of the modern science of behavioural ecology (Fernand Gobet, 2017, pg.5)." Though they generally present theory as a belonging to a separate category than inductively-derived qualitative and quantitative laws, they do acknowledge that sometimes there is some overlap between the categories. For instance, Newton's theory of gravity is both an empirically-derived law and a theory with explanatory power (Fernand Gobet, 2017). They also point out a distinction between discovery of a theory that results from direct observation like the discovery that DNA is a double helix and discovery of a theory that is more conceptual and for which observing is not equivalent to discovering like Darwin's theory of evolution(Fernand Gobet, 2017, pg.34).

In later papers, Langley et al. (2006; 2019) have expanded their account of the components of scientific discovery to include process models, which capture deeper structural relations rather than just qualitative and quantitative regularities. These are more explanatory than merely descriptive. However, Langeley sees these as belonging to the later stages of science, not to the initial discovery phase.

I will argue that Langeley et al.'s formalisms of discovery are sufficient for certain puzzle-solving tasks, but not for conceptual discovery. I support Langeley et al.'s argument that the team's computational discovery algorithms

are replicating instances of scientific discovery in the form of a problem-solving activity. However, I replace their premise that scientific discovery is problem solving with my own premise that problem solving constitutes *one* type of scientific discovery. I conclude that the programs that I discussed above can perform puzzle-solving discovery under the definition I specified in Section 1, but they cannot perform conceptual discovery. BACON, DALTON, GLAUBER, and STAHL are able to find regularities like quantitative laws and taxonomies *within* a theoretical framework, but they are not capable of making decisions about choice of theoretical framework or able to interpret evidence to recognize when revisions to the theoretical framework may be warranted.

Another account that aims to establish the computability of scientific discovery comes from Ioan Muntean (2014) and a study conducted by Schmidt and Lipson (2009). Muntean defines a type of scientific discovery which he calls bottom-up discovery and argues that bottom-up discovery is computable. Furthermore, he argues that some level of creativity and autonomy can be achieved by the computer programs conducting such discoveries. In particular, his argument focuses on the use of computer programs that are designed to operate using concepts of biology. Called genetic algorithms, these programs perform iterative search within a problem space for the best solution. These programs are structured around the idea of evolutionary constraints and organism adaptation. The aim of the research program is to find a way to generate algorithms that adapt and evolve to changes in domain constraints.

Mimicking an evolutionary process in nature, genetic algorithms begin with a population of individuals in an initial state that then either die or propagate into the next generation according to the processes of natural selection, sexual selection (reproduction of the fittest), recombination, and mutation (Muntean 2014, pg. 3). The program stops when a pre-programmed

level of success is reached or when all of the individual entities evolve into to be identical entities. This strategy has been used by Schmidt and Lipson (2009) to find laws and persisting properties in physics. In their programs, the individual entities can be mathematical formulas, models, or heuristic methods. The environment that the individuals exist within is a set of empirical data. When applied to the empirical data environment, some individuals do not "survive"–they are inconsistent or mathematically unsound. In other cases, an individual fits well with the data and is allowed to survive and are cross-combined with some other successful individual. In running this type of program through many iterations, Schmidt and Lipson discovered that very 'fit' individuals come into existence that provide optimal models of the data (Muntean 2014).

One thing to note is that the automation of scientific discovery I have described is not achieved without significant involvement of human developers. The involvement raises some questions: which decisions are automated and which are being performed by the human scientists? Moreover, is the work that is being performed by the AI sufficient to qualify the AI as discoverer? The computational discovery AI programs cannot achieve all of the activities that are associated with scientific discovery. For instance, even more autonomous programs like genetic algorithms cannot invent new instruments for measuring empirical data, change methodologies without a framework being provided by the programmers, or design experiments. Furthermore, scientific discovery has historically not occupied organized and premeditated problem spaces. In science performed by humans, part of the discovery process is deciding what the goal state is and what the and what the parameters of the problem space are.

Relevant to the question of whether puzzle-solving within pre-defined parameters counts as scientific discovery is the discovery game Foldit. Foldit is a platform that facilitates collaboration between non-scientist humans to

make scientific discoveries as a puzzle-solving activity (Cooper 2010). Foldit presents users with a 3D model of a protein and challenges them to configure the protein into a structure that uses minimal energy. The user solves the puzzle, the solution is judged by the energy state of the protein's configuration, and the user receives a numerical score. In competing to get high scores in a game that is presented to them as a puzzle to solve, humans with no scientific background are part of a collaboration to discover protein structures that have not yet been discovered by scientists. Foldit essentially is a highly productive scientific community comprised mainly of non-scientists, who, if asked, would not claim to be doing anything more than playing a game.

Yet, despite not being scientists or replicating any of the characteristics that have been associated with the traditional views of the discovery process in the philosophy of science literature, these gamers are extremely productive discoverers who have generated results that have been published in scientific journals (Cooper 2010). The puzzle that is presented to the gamers is not highly structured; it has vague aims like "find a good configuration that has a low energy score" and the gamers do not receive any feedback besides a single immediate numerical score (Cooper 2010). The platform has no goal of educating the players about science. The users are useful for scientific discovery by virtue of their puzzle-solving ability, not their knowledge of science. So, the gamers are not being trained to be scientists who understand protein structure; instead their role in the platform is identical to that of an AI that learns by reinforcement learning. They are not communicating with the scientists, they are not being told why their answer is plausible or implausible; they are merely being provided with vague constraints and rewarded more highly for answers that satisfy certain heuristics. Just like in a neural network and human scientist dynamic, the puzzle-solving computer is a black box to the human scientist and the human scientist is a black box to

the puzzle-solving computer. The only information that is being passed between them is the formal structure of the puzzle, the result, and the reward for a better result (Khatib 2011).

Like AIs that use evolution strategies instead of reinforcement learning, Foldit has one method called "evolving" that allows users to work together. In this system, users join together in groups and share their solutions with other members of the group. As the solution passes around between members of the group, each tries to improve it. If it is successfully improved upon the new solution is marked as an evolved solution. The solution that the group submits for scoring is the solution that has the highest overall score, and it can either be one of the evolved solution or a solution that was generated by just one member.

Foldit has two main types of puzzles. The first, and more common, type of puzzle is called a prediction puzzle, where the amino acids are fixed and the user merely manipulates the way that the protein is structured in 3D. The second puzzle is a design puzzle, and is a puzzle where the gamers can actually change the amino acids. The first is like giving gamers a jigsaw puzzle that can be configured in multiple ways and asking them to find the most aesthetically pleasing configuration. The second is like having a jigsaw puzzle that has multiple options for each piece, and allowing a player a little more freedom in choosing pieces to make the best design.

The interesting aspect of the Foldit platform is that, as is the case with many current AI programs, entities that have no scientific knowledge are making incredible scientific discoveries. Most recently, Foldit gamers have generated the structure of an enzyme that relates to AIDS that scientists have been in search of for over 10 years (Khatib 2013). The enzyme is a significant step towards curing AIDS. The time that it took for gamers to come up with the model for the enzyme was only three weeks.

Can these Foldit gamers make scientific discoveries? Although these discoverers are not trained scientists, it seems undesirable to deny that they are discoverers simply because they lack formal scientific training. Whewell would probably disagree. He says, "previous condition of the intellect, and not the single fact, is really the main and peculiar cause of the success. The fact is merely the occasion by which the engine of discovery is brought into play sooner or later. It is, as I have elsewhere said, only the spark which discharges a gun already loaded and pointed; and there is little propriety in speaking of such an accident as the cause why the bullet hits its mark (Whewell 1840, pg. 189)." While the bullet hitting the mark may be the result of the previous condition of the intellect, in the case of puzzle-solving discovery, a good condition of the intellect need not be equivalent to a broad knowledge or understanding of science. Conceptual understanding is not required for making a scientific discovery when the discovery in question is a puzzle-solving discovery as defined in Section 1, not a conceptual discovery.

Chapter 3

Prerequisites for AI Making Conceptual Discoveries

In the previous section I aimed to show that AI can perform puzzle-solving discovery. In this section I will discuss three elements that I will argue are essential to conceptual discovery. These are abductive reasoning; a representation language that incorporates fuzziness, paraconsistency, and predicate logic; and an ability to make and use conceptual metaphors. I will argue that these are the preconditions to AI making conceptual discoveries and I will examine the current abilities of AI to satisfy each precondition.

3.1 Formalizing Abductive Reasoning

Abductive reasoning seems to be essential to conceptual scientific discovery. Scientific discovery does not merely involve extrapolating implicit consequences of existing empirical observations, as programs like BACON do. Instead, scientific discovery often depends on scientists abductively reasoning to formulate a theory that provides a plausible explanation for the empirical evidence. This theory may not be a perfect fit for the data. For instance, Newton hypothesized that gravity explained planetary motion, even though this hypothesis made imperfect predictions about Mercury's orbit. Newton knew that Mercury's elliptical path was not fixed; it was rotational. This did

not fit with his hypothesis about gravity, but the theory was still the best option available to explain the total body of empirical evidence available at the time (Curiel 2018).

Kuhn does not explicitly discuss abductive reasoning. However, implicit to his argument is the claim that paradigm shifts, like the shift to the heliocentric model of the solar system, rely on abductive reasoning, not just inductive reasoning. Scientists use abduction to propose a compelling explanatory framework for anomalies that the old paradigm is unable to explain, which can lead to paradigm shift even when the new paradigm is incoherent with or unable to explain some phenomena. Kuhn says,

> Almost from the start of his electrical researches, Franklin was particularly concerned to explain that strange and, in the event, particularly revealing piece of special apparatus [The Leyden jar]. His success in doing so provided the most effective of the arguments that made his theory a paradigm, though one that was still unable to account for quite all the known cases of electrical repulsion. To be accepted as a paradigm, a theory must seem better than its competitors, but it need not, and in fact never does, explain all the facts with which it can be confronted (Kuhn 1962, pg. 17).

In order for AI to make conceptual discoveries, it will need to be able to replicate abductive reasoning: application of epistemic values and domain-specific goals to choose which data are most pressingly in need of explanation and production of an explanatory framework that explains those data points while also satisfying scientists' other epistemic commitments. However, formalizing abductive reasoning in computers is a significant challenge. There is even contention among philosophers about how to define abduction.

Even in Pierce's work, some important characterization of abduction is left ambiguous, or seems to be approached differently in different parts of his work. For instance, in one paper Pierce says that "it must be remembered that abduction, although it is very little hampered by logical rules, nevertheless is a logical inference asserting its conclusion only problematically or conjecturally, it is true, but nevertheless having a perfectly definite logical form (Pierce, 1931, pg. 188). Yet in the same paper, his definition seems to differ and he calls abduction "a flash..an act of insight(Pierce, 1931, pg.180)". Pierce subtly changes his characterization of abduction over the course of his career. Kapitan (1992) responds to the differences between Pierce's earlier and later work on abduction by suggesting that Pierce actually intended that abduction should mean a dual process, a generation and then selection of hypotheses. Kapitan calls the first part "abductive discovery" and the second "abductive preference."

Since acknowledgement of the importance of abduction has spurred AI research in formalizing abduction, abduction has come to have a number of slightly differing definitions within the AI literature as well. In most general terms, AI research in abduction is more in alignment with "abductive preference" than "abductive discovery". The number of expert systems that aim to use abduction for tasks like diagnosis and problem solving are generally set up with the following structure. The logic-based ones have a theory T that formalizes the constraints of the domain that the program will be working in, a number of formulas X for the different effects that might be observed within the system and the way they might appear, and a set of possible hypotheses Y that could explain the effects. The expert system searches for a formula from the set Y that is consistent with the domain theory and from which the relevant effect in the set X could logically follow.

Even in the ones that are not logic based, the process can be described as selective search to choose a theory, rather than generation of new theories.

For this reason, the abductive abilities of current AI programs are sufficient for puzzle-solving, where decisions are made within a theoretical framework, but are not sufficient for conceptual discovery, where the theoretical framework itself is restructured or refined.

3.2 The Frame Problem

Closely related to the need for AIs that can perform abductive reasoning is the frame problem (Shanahan 2008, 2016). On the AI research side, the frame problem is concerned with trying to find a way to logically model the effects of an action without a lot of unnecessary, obvious, or non-effects being included in that model. On the philosophy side, the frame problem is concerned with how to explain our ability to act based on relevant information without having to sift through all the non-relevant information. It is also concerned with how the reasoning process can be streamlined to still result in the correct effects but not include irrelevant information.

Essentially, the frame problem describes the fact that it is difficult to create logical formulas that describe the effects of an action without including a mass of axioms describing what does not change as a result of the action. Shanahan (2016) explains this with an example.

Take the two following formulas "painting x causes x to be some color" and "moving x causes x to be in some position". Let's say the color of object x is blue and the position of object x is in the car. Now let's say the object x has the actions "painted yellow" and "moved to porch" acted upon it. Common sense says that the object is now yellow and on the porch. However, Mccarthy and Hayes (1969) showed that classical predicate logic successfully concludes that the object is positioned on the porch, but does not conclude that the object is colored yellow. This is because the logic itself does not successfully represent the fact that the color of object x does not get changed

by the object x being moved (Shanahan 2016). This common sense principle would need to be included as separate frame axioms, which would specifically state that color should not change when the object moves and position should not change when the object is painted unless there is a specific reason that a change should occur. However, this is a hugely cumbersome and inelegant way of solving the problem. For one, many systems would become many times their original size if they included frame axioms for every case where they needed this principle (Morgenstern 1996). Furthermore, it signals a problem in what logic can accomplish and what it is. It is not really modelling our reality if we need to constantly include external axioms to make it obey the principles of our reality. Most logicians do not want to think of logic as a utilitarian structure that is returning the right results only in cases where masses of frame axioms have been constructed by hand to prop it up and stop it going astray at every action.

Thus, the solution to the logical frame problem is to find a way to formalize the idea that actions should not cause assumed changes to a property of an object unless there is an explicit reason that the action would change the property (Morgenstern 1996). The problem with this is that classical logic allows more conclusions as more axioms are added, and this makes it difficult to express a law like the idea that most properties don't change with most actions, which is a rule that has an unknown number of exceptions. For instance, in our earlier example, perhaps some avid painter paints a car so vigorously that the car rolls down a hill.

Though these practical solutions exist, some philosophers consider the idea that the frame problem that arises in AI research might signal "a new, deep epistemological problem—accessible in principle but unnoticed by generations of philosophers (Dennett 1987)." Fodor says, "the frame problem goes very deep; it goes as deep as the analysis of rationality (1987)." The

answer to the frame problem seems to lie in finding a way to program a notion of relevance into such programs, but relevance is fundamentally context-dependent. There is no universal set of axioms that give rise to a flexible, wide-ranging notion of relevance. Of course, in a very narrow system working on puzzle-solving discoveries, a notion of relevance that works for the particular task can be pinned down. For an AI to make conceptual discoveries, however, a more flexible formalization of relevance is needed, especially as true scientific discovery-making is often about discovering relevance where none was expected. A new scientific theory often arises when some previously known data takes on a new relevance which gives rise to a new way of looking at preexisting facts and data (Kuhn 1962). An AI program that cannot recognize relevance in different contexts will be deficient in one of the main things science requires.

3.3 A Proper Representation Language

The architecture underlying deep learning AI systems is different from a traditional logic system as they rely on heuristics and probabilistic reasoning. There has been some discussion about whether the structure of deep learning systems is fundamentally different from and irreconcilable with mathematical and logical language (Desjardins-Proulx 2017, pg. 1).

Some philosophers draw parallels between AI and the cognitive science concept of System 1 reasoning. In psychology, Dual Process Theory describes the idea that when reasoning, humans use two systems. System 1 is an intuitive gut-feeling system based on patterns learned through experience, and System 2 is rule-based, systematic reasoning relying on conscious selection of rules and heuristics (Kahneman 2011). The first process is subconscious and implicit, whereas the second process is explicit and consciously-performed,

and can change as we consciously choose different models and axioms to perform reasoning with. Further distinctions that have been discussed include the idea that System 1 is involuntary, unconscious, allows rapid reasoning, is structured as holistic top-down reasoning, and is not verbal or linked to a language (Kahenman 2011). System 2 is more general and applicable across multiple domains, is based on explicit rules and axioms/facts, is oriented as bottom-up reasoning based on underlying assumptions, and is language and model-based. Neural networks behave much more like System 1. For instance, a deep learning AI that distinguishes between images of bicycles and cars will have a first layer that identifies small segments of lines, a second layer that identifies how two small segments of a line are oriented in relation to each other, a third layer that identifies more complex shapes based on the lines relation to each other, and a layer that identifies even more complex shapes from those complex shapes. The idea is to have many layers and have the first layers analyze very simple patterns and the latter layers analyze increasingly complex patterns. This method has great potential to be very useful for modelling or finding patterns in complicated systems that are beyond the scope of mathematical formalisms and mental models. However, it is somewhat different from a formal system of logic or mathematics. Whereas formal logic involves a bottom-up approach with a structure where a skeleton/foundation which is filled in with consistent theorems, a neural network is top-down structured entity where unstructured data is analyzed and a skeleton/foundation is created. The practice of mathematical and formal logic correlates with System 2. The idea of combining formal logic and mathematical systems with deep learning neural networks would suit the potential goal of attaining a system that incorporates both of these modes of reasoning. This would be very useful (and perhaps necessary) for making scientific discoveries with AI because both of these modes are fundamental

to scientific reasoning. "Effective scientific reasoning requires [that] individuals must understand how to assess what is currently known or believed, develop testable questions, test hypotheses, and draw appropriate conclusions by coordinating empirical evidence and theory (Morris 2012, pg. 62). They show that this coordination especially manifests in a balance between instinct and explicitly learned models and strategies. Minsky (1986) says,

> For generations, scientists and philosophers have tried to explain ordinary reasoning in terms of logical principles with virtually no success. I suspect this enterprise failed because it was looking in the wrong direction: common sense works so well not because it is an approximation of logic; logic is only a small part of our great accumulation of different, useful ways to chain things together (Minsky 1986, p. 167).

In order to unify traditional logic with deep learning, it is necessary to develop a common representation language that is compatible with both artificial intelligence algorithms and formal logic systems (Desjardins-Proulx 2017). Furthermore, in order to harness the impressive power of AI to process data and make predictions, we would require a method for formalizing existing scientific knowledge that is compatible with AI processing. AI projects that aimed to formalize inductive reasoning performed on examples and background knowledge have achieved discovery of new rules(Muggleton et al. 1994), but the logical representation language that they use is too inflexible to handle the way scientific theories can be partially true, overlapping, or inconsistent (Desjardins-Proulx 2017). In other cases, inconsistencies can mysteriously yield useful results, as is the case with Kirchoff's approximation of light behaviour moving through an aperture (Vickers 2011). In summary, AI will not be able to make conceptual discoveries until an appropriate representation language is achieved.

Another desideratum that has been suggested is the integration of probability theory, predicate logic, and fuzzy logic (Proulx et al. 2017). Simple predicate logic type structure has been used for many expert systems, like DENDRAL. However, simple predicate logic has its limits. Basic predicate logic does not allow for a case to be true most but not all of the time unless the formula explicitly specifies it. The formula "all swans are white" would be violated if there were a single instance of a black swan. This is problematic for representing scientific theories because these are a set of rough generalizations that are only usually true. A structure needs to be added so that a system retains the formula "all swans are white" even if applying actual values to its variables occasionally violates it. Yet the system also needs to acknowledge and record that the formula has been violated (Desjardins-Proulx 2017).

By adding a probabilistic element, systems can become more complex and more representative of the real world. In such a system, formulas in the system are weighted based on how likely they are. Application of a concrete value makes a formula more or less likely rather than true or false. One suggestion for achieving this is with Markov logic (Philippe Desjardins-Proulx, 2017). In Markov logic, logical formulas have weights according to how likely they are. If a possible world refutes a formula in the knowledge base the world is not impossible, but just less likely. A possible world is more likely if it violates less formulas. Another benefit of Markov logic is that it can accommodate contradicting formulas, which makes it good for integrating two knowledge bases. Consider the following example.

Suppose we wanted to design an algorithm which could discover how a specific population of birds has evolved traits to suit their environment after a staple of their diet, a native berry bush, suffered from a disease and mostly died off fifty years ago. Some of the first steps we might perform would be to find a way to create a basic representation of the knowledge we had about the

environmental pressures on the population, a list of the general phenotypes of the population, whether various traits are dominant or recessive genetically, and the degree to which environmental pressures encourage or discourage certain traits. Paralleling the example Proulx et al.(2017) gives with species interactions, I will show how Markov logic is useful for representing such a system. The first step in our representation would be to suggest that all organisms have a set of traits, n. Suppose our organism is a bird and one trait is beak length, a second is diet, and a third is stomach acid composition. Existing observations might be described by the following rules:

1. First, let's make one of our axioms be that most short-beaked birds do not eat sunflower seeds. $\forall$x(HasShortBeak(x) $\Rightarrow$ $\neg$ EatsSunflowerSeeds(x))

2. $\forall$x (HasLongBeak (x) $\Rightarrow$ EatsSunflowerSeeds)

3. $\forall$ x (If HasLongBeak (x) $\Rightarrow$ $\neg$ HasStrongStomachAcid(x)).

4. $\forall$ (HasShortBeak (x) $\Rightarrow$ $\neg$ HasStrongStomachAcid(x))

After further empirical study, it turns out that (1) and (4) are not always true; sometimes short-beaked birds do eat sunflower seeds and have strong stomach acid. In an ordinary predicate logic system this would be problematic, but using Markov logic the statements would be weighted and it might emerge as a theorem of the system that some birds have short beaks and do eat sunflower seeds, and that these birds happen to be the only ones with strong stomach acids for breaking down the seed shells. Instead of just being correct, (1) and (4) are only usually correct, and this provides a useful new refinement of our representation of the bird population.

Markov logic has been used to good effect in several applications of AI to science. For instance, Yoshikawa (2009) used it in identifying temporal relations between events. Brouard et al. propose Markov logic as a way to refine inferences about gene regulatory networks in systems biology (Brouard et al. 2013).

One issue though, which shows that Markov logic alone is not quite enough

to model systems in science, is that the predicates need to come out either true or false when actual values are assigned to them. The predicates above are not very good at seeing spectrums or ranges of traits; they require simple binary results (Proulx et al. 2017). For instance, with beak length, we would have to decide on a random length that would be the threshold for "HasLongBeak" in order for the predicate to be able to return a true or false value. This could be the downfall of our whole program; suppose we chose a number that was a bit too high and the predicate returned false even though there was a significant group of long-beaked birds who hovered just under the threshold? Perhaps there would be a whole group of birds with elevated but not threshold-level stomach acid and beak length who would be disregarded by the program. The binary requirement is problematic because the threshold that we choose for a predicate to be true affects what results we get.

What is needed is a way for the structure of our propositions themselves to be as flexible and multifaceted as our range of probabilities for them being true. This is where fuzzy logic is useful. This is generally achieved by using a zero to one continuum rather than a Boolean result and is known as probabilistic soft logic. In this case beak length would exist on a spectrum rather than designation of an arbitrary cutoff being required.

A final feature of the representation language needed to express scientific theories in AI is paraconsistency. In performing reasoning processes on data sets, an AI program will eventually encounter inconsistencies in its information repository. Much research in AI has been done that focuses on eliminating inconsistencies (Moss and Sleeman, 2012). However, some philosophers note that capacity for inconsistency can be an unavoidable and sometimes even a fruitful component os science. In some cases, very successful theories might have inconsistency buried within them. For instance, (Frisch, 2004) says that classical electrodynamics experiences contradictions within

its four foundational assumptions. Furthermore, he says that any attempts to fix these inconsistencies leads to serious conceptual problems (Frisch, 2004, pg. 525). His conclusion is that inconsistencies might be considered acceptable within science in certain cases. This is not the only example of an inconsistent scientific theory being incredibly successful; Kirchoff's theory of light diffraction is another example. In order to perform conceptual discovery, AI systems need to be able to handle inconsistency.

Scientific paradigms are often idealizations that highlight certain questions and phenomena while obscuring others to make problems more tractable (Kuhn 1962). For AI to be able to make conceptual discoveries, it must be able to revise underlying assumptions, even those initially regarded as probable, and sometimes start from premises that are flawed but potentially useful. Rigid adherence to only highly probable truths may prevent novel discoveries, as seen historically.

An appropriate representation language is an initial precondition for AI making conceptual discoveries. Integrating all three of these desiderata (paraconsistency, fuzziness, and predicate logic) would be a first step towards creating such a language.

3.4 Formalizing Conceptual Metaphor

Lakoff and Johnson's Conceptual Metaphor Theory suggests that metaphor is fundamental to everyday communication and perception (Johnson et al. 1979). Their research investigates the centrality of metaphor to human thought and language. They note that prior to their work, *Metaphors We Live By*, metaphor was often treated as rather superficial and literary accoutrement of speech.

This mirrors the current situation in AI research on metaphor and analogy in AI research. Much of the research that is being done involving metaphors

in AI is about how we can teach computers to understand metaphors in speech, in particular in the field of chatbot help-desk AIs that need to form appropriate responses to metaphorical language used by humans. Currently, AI research into metaphor is concerned with making chatbot AI programs respond appropriately to metaphors used in human speech. These lingual metaphors in speech are just mere figments of the deeper conceptual metaphors that cause them to pop up in language, according to Lakoff and Johnson. Yet much of the AI research on metaphor focuses on enabling AI to have the capacity to translate metaphorical sentences into literal sentences (Massey 2017).

Lakoff and Johnson propose that our conceptual models of the world are fundamentally metaphorical. The foundational claim of Lakoff and Johnson's Conceptual Metaphor Theory (CMT) is that

> ...metaphor is pervasive in everyday life, not just in language but in thought and action. Our ordinary conceptual system, in terms of which we both think and act, is fundamentally metaphorical in nature... [metaphors can] create a reality rather than simply to give us a way of conceptualizing a preexisting reality (pg. 144).

One example of a conceptual metaphor that they give is 'argument is war'. In our surface-level language, this conceptual metaphor emerges as a number of lingual by-products: 'You attacked the contradiction in my argument', 'I couldn't defend my argument against your objections, 'I win every argument with him', 'Don't shoot down my argument before you hear my conclusion!'. Lakoff and Johnson make the case that these statements are not just a literary, verbal comparison between 'argument' and 'war'. Rather, the 'argument is war' metaphor is actually embedded in the way we think about arguments; it cannot be removed without the actual conceptualization of 'argument' being affected. The structure of an argument reflects the metaphor.

If a certain line of reasoning is not helping us gain ground, we take up a new line of attack. A view is something to be defended, and a bad argument is something to be proved false and thus removed from the opponent's arsenal (Lakoff and Johnson, 1979 pg. 5).

Lakoff and Johnson point out that the conceptual metaphors that are embedded in society and in scientific communities operate as lenses that shape the interpretation of new phenomena we encounter. A conceptual metaphor, by which we conceptualize one concept in terms of another, has the dual effect of obscuring other facets of a concept. Sometimes the fact that the metaphor obscures alternative conceptualizations can be quite difficult to spot. Lakoff and Johnson share Michael Reddy's example of the metaphor 'language is a conduit' (Reddy 1979). Under this metaphor, concepts, knowledge, and meaning are objects. Language is a container that these objects can be placed into, and communication is the act of conveying these objects to other people. Reddy proposes that many the lingual expressions we use to talk about communication and language use this metaphor. A few examples of such expressions are: 'How can I get this concept across to you?', 'That was my idea, I gave it to you', 'that podcast is always putting new ideas in my head', 'your speech should pack more meaning into less words', 'the idea is good, but it's really buried under all the dense pedantry' (Reddy 1979).

It can be difficult to recognize such phrases as having a metaphorical basis because they are so embedded in our language and culture (Lakoff et al. 1979). The fact that a phrase or concept *is* metaphorical does not mean that we are at all aware of the metaphor even when we're using it. This 'language as conduit' metaphor would perhaps fit under (MacCormac, 1983)'s definition of a paraphor because it is so widely established and orthodox that nobody notices that they are using them. Paraphors are metaphors that have become so paradigmatic that they are treated as literal and the fact that

they are metaphors is not acknowledged. As MacCormac says, "By forgetting that theories presuppose basic metaphors and thereby by taking theories literally, both scientists and theologians create myths (MacCormac, 1983, p. 17)." Metaphors may be inevitable in our science and thinking, but it is important that they be subjected to critical evaluation to analyze the purpose they serve and what assumptions are entailed by the metaphor. For instance, the 'language as conduit' metaphor assumes realism about the thoughts we communicate and assumes that language does not alter the ideas that we communicate (Kudriavtseva 2015, pg. 116). This objectivist view presumes that language exists to communicate preexisting thoughts, and is characteristic of our Western societies (Schlesinger 1991, pg. 8; Kudriavtseva 2015, pg. 116; Lakoff et al. 2003, pg. 187). Under the objectivist view, metaphor would be a superficial act of comparison between two words whose meanings are already fixed.

However, there is good reason to think that language is context dependent. Lakoff and Johnson give the example with the phrase "Please sit in the apple juice seat". However, taken in different contexts the sentence can give rise to meaning, such as if there is a table set for breakfast with different glasses of juice at each place and the host is instructing a guest on where to sit (Lakoff and Johnson 1979 pg. 13). The non-objectivist view is supported in discussion of science as well; Kuhn says "I would hazard the guess that the same interactive, similarity-creating process which Black has isolated in the functioning of metaphor is vital also to the function of [conceptual] models in science (Kuhn 1970, pg. 415)."

One challenge for the prospect of formalizing Conceptual Metaphor Theory is that many of the metaphors that shape science are grounded in pre-scientific cultural and physical experiences (Lakoff et al. 1979). AI does not have access to culture or physical experience. The account given by Lakoff and Johnson runs counter to the sort of foundationalist, rationalist view that

characterizes much current AI research. Lakoff and Johnson's account claims that metaphors are heuristic device that we use to create understanding. Under their account, truth and meaning come from human psychology, culture, and context-dependent goals, not from any objective foundational source. Meaning is not disembodied; the truth of a statement depends on the conceptual context in which the statement is applied (Lakoff et al. 1979).

Analogy and metaphor play crucial roles in scientific hypothesis generation, theory choice, and experimental design, though their essential role sometimes goes unrecognized. Examples of metaphor in theory-building can be found in Darwin's analogy between artificial and natural selection (Darden 1982) and interpretations of quantum mechanics (Lauman-Lairson, forthcoming). Even in cases where these metaphors turn out to be wrong, they are often useful to the theory-making process and can yield empirically successful theories.

The points I discussed can be summarized as follows: the meaning of a theory is nontrivially interwoven with its metaphorical and historical context and a pattern that endures in the building of scientific theories is the use of metaphor.

If we subscribe to the idea that many conceptual advances in science utilize conceptual metaphors, then being able to make relevant metaphors is an important competency for AI to have if it is to make conceptual discoveries.

Under Conceptual Metaphor Theory, metaphors are key to conceptual models (and even to mathematical formalisms in the account given by Henry Poincare). For instance, Buchdahl says, "the intellectual satisfaction involved in explanations requires more than the establishment of merely formal connections between laws; but that the theory enables the laws which it explains

to be 'visualized'; it traces an analogy, more or less close, between the phenomena expressed by the laws and some other phenomena usually of a mechanical nature with which we are familiar in everyday experiences (Buchdahl 1964, pgs. 159-160)." This idea of developing new theories by using metaphor to relate new phenomena to preexisting, visualizable processes is one that comes up repeatedly in reflective accounts that scientists give of their work. The accounts always specify that what is being strived for "must involve...picturable physical mechanics...processes that can be pictures (Cushing 1991, pg. 341)" and "an explanation which is a reduction to more familiar notions ((Campbell 1991, pg. 157)." In particular, metaphors are essential for providing explanations, for in the absence of an understanding of a new phenomenon we must relate it to existing conceptual frameworks. "That is the essence of metaphor—an unusual juxtaposition of the familiar and the unfamiliar (MacCormac 1985, p. 9)."

I will discuss some of MacCormac's categorizations of metaphors and their uses because it will be useful for creating a more focused discussion about what kinds of metaphors AI needs to be able to make. MacCormac discusses three main facets of metaphor which will be useful for my later sections, as well as a distinction between two different types of metaphor. These are (1) the strength of a metaphor, (2) the process by which a linguistic metaphor becomes a descriptive, conceptual metaphor, and (3) the direction that a metaphor moves in. I will first discuss Cormac's argument about the strength of a metaphor. A metaphor says that two things are similar. MacCormac points out that it is not usually the case that a whole object/phenomena is similar to a whole other object/phenomena, but rather that some aspect of one object/phenomena is similar to some aspect of the other object/phenomena. The similarity is drawn between properties of things, not things themselves (If such a distinction can even be made!). MacCormac states that objects or phenomena possess properties, which he calls referents,

and a metaphor pairs referents from one object/phenomena with referents of another object/phenomena. The referents can either be analogous of disanalogous, and the ratio of the number of paired referents that are similar to the number of paired referents that are not similar determines the strength of a metaphor, or its expressive weight. Of course, then the strength of a metaphor is dependent on how many paired referents each object has that are not dissimilar, and a correlation could be seen between how many referents has and the strength of the metaphors that can be made from it, since more referents will usually result in a low ratio of similar to not similar numbers of referents. Furthermore, the number of referents is dependent on what is known or believed about the object. So, it would actually be the case that things which less is known about would be easier to make strong metaphors with. For instance, the metaphor 'DNA is a computer code' was very useful to early descriptions of DNA and engendered a whole species of language that revolves around the metaphor: "DNA editing", DNA operator, DNA copy error". When less was known about DNA, conceptualizing it in terms of computing was a useful metaphor. However, as both DNA research and computing research has advanced, it has been suggested that 'DNA is a computer' is no longer a useful metaphor because the number of referents that are not analogous.

The above demonstrates a case in which the strength of a metaphor can weaken as the number of referents that the two entities possess increases. When two entities are viewed through a simplistic lens that reduces their number of referents it can lead to a stronger metaphor.

The second of MacCormac's points is a distinction between three types of metaphor. He describes diaphors to be literary, linguistic metaphors in which there are not many referents, or there is a low ratio of analogous referents to disanalagous referents. MacCormac would probably categorize Lakoff and

Johnson's classical example 'Their relationship was a block of ice' as a diaphor. The metaphor draws a comparison between the two entities which describes a single property of the relationship very well, but it does not achieve many analogous pairs of referents.

In contrast, an epiphor has a high number and ratio of analogous to disanalogous referent pairs. These are the metaphors that really describe phenomena and overturn new ways of conceptualizing things. These are also the metaphors more associated with science rather than literature and poetry, though as MacCormic points out, diaphors do become epiphors and vice versa.

Finally, there are paraphors. These are metaphors that have become so paradigmatic that they are treated as literal and their metaphorical roots become obscured. MacCormac calls these metaphors dangerous to science, and points out that "By forgetting that theories presuppose basic metaphors and thereby by taking theories literally, both scientists and theologians create myths. (MacCormac, 1985, p. 17)" This will be a useful idea to note, because one would expect that AI would be especially susceptible to treating metaphors as fact; it would be interesting to consider how you would program a computer to use metaphorical thinking to draw its conclusions, yet then not treat as literal the conclusions it has drawn. MacCormac also points out that many of our paraphors are treated very much as literal facts, except when the spotlight of scrutiny is shone on them directly, and then they figuratively hold up their hands with the defense "I'm just a metaphor!".

It is important to make the distinction here between epiphors that have achieved literal meaning and paraphors. Paraphors are not taken literally because scientific enquiry has determined through empirical investigation and debate that the metaphor is in fact best understood literally. Rather, paraphors are cases where a metaphor is so commonly known that it is almost subconsciously referred to literally during any work in a subject.

MacCormac's discussion of the danger of uncritically treating metaphors as literal revolves around paraphors. However, he points out that not only is the main paraphoric metaphor a potential problem, but the supplementary metaphors that accompany that main metaphor pose another risk.

Previous sections have concluded that AI is currently making puzzle-solving discoveries but not conceptual discoveries. I have argued that conceptual metaphors are an essential part of paradigm shift and abductive reasoning.

Crucial to the question of whether AI can make and use conceptual metaphors is the question of how conceptual metaphors can be formalized. Some recent studies have been working at the intersection of AI and metaphor, simile, and analogy to investigate AI's ability in this area. However, this research has not extended to include the types of conceptual metaphors that operate in scientific discovery. Much of the existing research on AI and metaphor is concerned with linguistic applications. The first program that I will discuss is a chatbot that assists people in troubleshooting computer issues.

The program MIDAS (Metaphor Interpretation, Denotation, and Acquisition System) was developed by James Martin in the 1990s. MIDAS was developed to improve a program that answers questions about the operating system Unix, for users who are experiencing problems with it. MIDAS uses a map to link concepts in a network of metaphors. A phrase that the program might come across when interacting with a user is "Okay, I am in Mozilla Firefox now". To correctly interpret this statement, MIDAS must know that 'physically occupying a region' is a conceptual metaphor for 'using a computer process' (Barnden, 2008). This is formalized by the metaphor map, which creates three conceptual metaphors. These are:

4. Physically being inside a space = using a program within a computer.

From this conceptual metaphor come two constituent conceptual metaphors; namely:

5. the physical space doing the act of enclosing = the program within the computer that is being used. And

6. the thing that is being enclosed in the space = the user of the computer program.

MIDAS would not interpret 'I am in Mozilla Firefox' literally because (1) the program Mozilla is not represented in the program's database as being a conceptual space and (2) MIDAS is programmed to prefer to apply conceptual metaphors even if a literal reading of the statement is coherent with other information in its database.

The real power of MIDAS comes from the fact that it can extrapolate on its existing map of conceptual metaphors to correctly interpret new statements that it has not come across before. While it cannot conduct a creative process in which it creates new parallels between things, it can extend existing links. The two techniques it uses are similarity extension and core extension (Martin 1992). The former method works by using existing parallels between concepts, and then inferring that the metaphor that works for one also applies to the other one. For instance, the computer might come across the phrase "I am in a phone call" and find in its repository a parallel between 'phone call' and 'computer program' that comes from them both being tagged as a certain type of process that a human can engage in. Taking this knowledge, MIDAS might take the conceptual metaphor 'physically being inside a space = using a program within a computer" and the parallel between phone call and computer program to conclude that a reasonable conceptual metaphor is 'talking on the phone is being inside a physical space', and that the human is likely not literally located inside of the phone call, but is merely talking on the phone.

The second method is core extension. It is called this because it only works when two concepts are related to one another on a core level, like a state being the result of an action. Other very direct relationships would

also work, like equality or 'if this than not that'. One application would be to use the link between cause and effect to understand results of actions. For instance, if a user asks how to get into Mozilla Firefox, the metaphor interpretation process needed is more complex than just knowing the conceptual metaphor 'being inside a space = using a program on a computer'. However, MIDAS can handle this by knowing the aforementioned conceptual metaphor and some basic information about physical spaces like the fact that you can move into and out of them. These are the so-called core concepts; get-into is core-linked to being-inside-of. MIDAS needs to know that 'get into' means the action of entering a physical space, and that the result of the action 'get into' would be 'being inside of a space'. Using these two things, MIDAS can then reason that the user is asking what action results in them being inside the space, which equals using the program. The result MIDAS would come up with as an interpretation of 'how do I get into Firefox?' would be 'How do I commence using the program Firefox?' The new conceptual metaphor which MIDAS creates is 'Getting into a space = commencing use of a computer program' (Barnden, 2008).

One problem with MIDAS is that the metaphors that a metaphor map includes is dependent on what metaphors are already in the system and how the database of mappings was initially set up. It is easy to imagine two MIDAS AIs with initial starting definitions and mappings that seem similar to a human, creating very different results as their evolution spirals out in different directions. It also seems that MIDAS can create mappings between concepts that are not meaningfully related even though the way they are represented in MIDAS' database makes them seemingly similar. The contents of MIDAS's initial repository of knowledge could have a significant impact on what later metaphors it makes.

Representing Lakoff and Johnson's CMT with computer systems is a complex task, and there are some big differences between the different efforts that

have been made to achieve this. One issue that comes up is a lack on consensus on what core metaphors should be used and how extensive this set of foundational metaphors should be. Lakoff and Johnson "claim that a small set of generalized metaphors structured in a hierarchy provides the framework for metaphor interpretation and creation (Jakubowsky 1999)."

In contrast, some researchers like Tony Veale and Mark Keane, in building AIs to perform metaphor interpretation, have taken Lakoff and Johnson's basis to mean that a very loose, small set of metaphors should be started with, and a process of "conceptual scaffolding" should be used in place of an extensive starting set of core metaphors. In this approach, the computer interprets a metaphor in a very broad, vague way, and then this interpretation becomes more sharp and precise in subsequent steps.

The type of conceptual metaphor theory that needs to be incorporated into AI research should have the goal of allowing AIs to develop model-based theories (Brewer 2001). A model-based theory is "a conceptual framework that provides an explanation for a set of phenomena by postulating a structural relation to another more familiar concept (Brewer 2001, pg. 33). There is a good deal of discussion in science about the idea that new, revolutionary science occurs by use of metaphors because the old language is not sufficient for describing the new phenomena and thus must be described through analogies from more conceptually accessible phenomena. For instance, Buchdahl discusses the importance of explanation and says, "the intellectual satisfaction involved in explanations requires more than the establishment of merely formal connections between laws; but that the theory enables the laws which it explains to be 'visualized'; it traces an analogy, more or less close, between the phenomena expressed by the laws and some other phenomena usually of a mechanical nature with which we are familiar in everyday experiences (Buchdahl 1964, pgs. 159-160)." Cushing says, "understanding of physical processes must involve picturable physical mechanics

and processes that can be pictures (Cushing 1991, pg. 341)." The idea that the generation of new theories requires a conceptual, metaphorical element, and not just a formalized process using language in traditional ways comes up repeatedly in science. Campbell (1991) says that new theories "can be derived from simplicity and generalization, from an explanation which is generalization as well as from an explanation which is a reduction to more familiar notions (Campbell 1991, pg. 157)." The idea that emerges is that (1) discovering new theories depends upon being able to visualize new phenomena in terms of more familiar, picturable notions, (2) metaphors are an essential part of the process of visualizing new phenomena in terms of existing ideas. I propose that formalizing conceptual metaphor could provide a framework by which computer discovery process could come to more closely resemble the process (and thus hopefully the results) of human conceptual discovery processes.

Chapter 4

Conclusion

As artificial intelligence plays an increasing role in scientific discovery, questions arise about the scope of AI's ability to contribute to scientific discovery. I have discussed Kuhn's distinction between the puzzle-solving of normal science and the conceptual shifts of revolutionary science. In surveying the current literature on AI's contributions to scientific discovery, I argued that AI can make puzzle solving discoveries but not conceptual discoveries. In section 3, I discussed some of the prerequisites for AI making conceptual discoveries such as a proper representation language, an ability to perform abductive reasoning, a solution to the frame problem, and a structure that takes into account conceptual metaphor theory. Finally, I have make an argument for why each of these is a prerequisite for AI making conceptual discoveries, and I have explored the progress that is being made in each of these areas.

Bibliography

Alai, Mario (2004). "Artificial Intelligence, Logic of Discovery, and Scientific Realism". In: *Minds and Machines* 14.

Appenzeller, Tim (July 2017). "The AI Revolution in Science". In: 10.

Arturo Casadevall, Ferric Fang (2016). "Revolutionary Science". In: *American Society for Microbiology*.

Barnden, John (2008). "Metaphor and Artificial Intelligence: Why They Matter to Each Other". In.

Bolc, Leonard (2012). *Computational Models of Learning*. Springer Science Business Media.

Bringsjord, Y. Yang (2007). *Mental Metalogic: A New, Unifying Theory of Human and Machine Reasoning*.

Brouard, C (2013). "Learning a Markov Logic network for supervised gene regulatory network inference." In.

C. Piscopo, M. Birattari (2013). "Invention vs. Discovery". In: *Encyclopedia of Creativity, Invention, Innovation, and Entrepreneurship*.

Choi, Byung-Kwon (2018). "Literature-based automated discovery of tumor suppressor p53 phosphorylation and inhibition by NEK2". In: *Proceedings of the National Academy of Sciences*.

Ducasse, C.J. (1951). *Whewell's Philosophy of Scientific Discovery*.

Fernand Gobet Peter Sozou, Peter Lane Mark Addis (2017). "Computational Scientific Discovery". In.

Frisch, M. (2004). "Conceptual Problems in Clasical Electrodynamics". In.

George Lakoff, Mark Johnson (1980). *Metaphors We Live By*. University of Chicago Press.

Gillies, Donald (1996). *Artificial Intelligence and Scientific Method*. Oxford University Press.

Gregory, Madeleine (July 2019). "AI Trained on Old Scientific Papers Makes Discoveries Humans Missed". In.

Hanson, Norwood Russell (1960). "Is There a Logic of Scientific Discovery?" In: *Australasian Journal of Philosophy* 38.

Hosanger, Kartik (2019). *A Human Guide to Machine Intelligence*.

Jakubowsky, Karen (1999). "Metaphor and Understanding: The Work of Lakoff and Johnson and Natural Language Processing". In.

Kuhn, Thomas (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kurzweil, Ray (Sept. 2005). *The Singularity is Near: When Humans Transcend Biology*. Penguin.

MacCormac, E. (1983). "Scientific Metaphors as Necessary Conceptual Limitations of Science". In.

Magnus, P.D. (2012). *Scientific Enquiry and Natural Kinds*. Springer.

Moss, Laura and Derek Sleeman (2012). "Detecting and resolving inconsistencies between domain experts' different perspectives on classification tasks". In.

N., Ensmenger (2012). "Is chess the drosophila of artificial intelligence? A social history of an algorithm." In.

Noe, Keichii (2002). "The Structure of Scientific Discovery: From a Philosophical Point of View". In.

Pat Langley Herbert Simon, G. Bradshaw J. Zytkow (1987). *Scientific Discovery, Computational Explorations of the Creative Processes*. MIT Press Cambridge.

Philippe Desjardins-Proulx Timothee Posot, Dominique Gravel (2017). "Scientific Theories and Artificial Intelligence". In.

Pierce, Charles (1931). *The Collected Papers: Abduction and Pragmaticism*.

Pierce, Charles S. (1903). "Harvard Lectures on Pragmatism: Lecture VI". In.

Poincare, Henri (1908). "Science and Method". In.

Radiology, European Society of (2019). "What the Radiologist Should Know about Artificial Intelligence". In: *Springer Berlin Heidelberg*.

Schickore, Jutta (2022). "Scientific Discovery". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2022. Metaphysics Research Lab, Stanford University.

Schindler, Samuel (2015). "Scientific Discovery: That-whats and what-thats". In: *Ergo* 2.

Simon, Herbert A. (1992). "Scientific Discovery as Problem Solving". In: *International Studies in The Philosophy of Science* 6.

Swanson, Don R. (1986). "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge". In: *Johns Hopkins University Press* 30.

Texas, University of (Mar. 2019). "Two New Planets Discovered Using Artificial Intelligence". In: *Phys.org*.

Thyssen, Pieter and Koen Binnemans (2015). "Mendeleev and the Rare-Earth Crisis". In: *Philosophy of Chemistry: Growth of a New Discipline*. Ed. by Eric Scerri and Lee McIntyre. Dordrecht: Springer Netherlands, pp. 155–182. ISBN: 978-94-017-9364-3. DOI: 10.1007/978-94-017-9364-3_11. URL: https://doi.org/10.1007/978-94-017-9364-3_11.

Turing, Alan (1950). "Computing Machinery and Intelligence". In: *Mind* 5.

Whewell, William (1840). *The Philosophy of the Inductive Sciences, Founded Upon Their History*.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Ed. by G.E.M. Anscombe. New York, NY, USA: Wiley-Blackwell.