Forthcoming in E. Ippoliti, L. Magnani, and S. Arfini (eds.), Model-Based Reasoning, Abductive Cognition, Creativity. Inferences & Models in Science, Logic, Language, and Technology, Springer.

A Neuro-Symbolic Approach to the Logic of Scientific Discovery

Ioannis Votsis (Northeastern University London)

1. Introduction

Scientific discovery is a neglected topic in the philosophy of science (Langley and Arvay 2019). Since around the middle of the last century, the received view has been that discovery is not governed by logic or, more generally, by rationality, but is a largely elusive and inscrutable process (Popper [1935]1959). Thankfully, not everyone has been persuaded by this mystical view (Langley et al. 1987; Cellucci 2013). Given the recent cascade of developments in automation and AI, this means that now more than ever we need to carefully re-evaluate our attitude towards this view. This paper aims to do precisely that by exploring how such developments are reinvigorating the search for a logic of scientific discovery. Neuro-symbolic approaches to AI, it will be argued, offer hope in the reinstatement of the rationalist model, and the automation, of scientific discovery. A neuro-symbolic system is proposed that seeks to integrate several identified desiderata like the ability to detect patterns and to reason. On this proposal, both neural and symbolic methods are employed in the service of automating scientific discovery, but the former are conceived of as subservient to the latter. The rationale for this subservience relation is that any discoveries made need to be integrated into an accessible body of knowledge, and that can more easily be achieved with symbolic methods at the helm.

The paper is structured as follows. Section 2 provides a brief overview of the debate over scientific discovery, touching upon the themes of what logic, if any, best captures it, what heuristics can be employed to assist in discovery and how to make causal discoveries. Section 3 introduces the symbolic vs. neural AI distinction and discusses the strengths and weaknesses of each approach, particularly in relation to the topic of scientific discovery. Section 4 outlines several different types of attempts to find a middle ground through hybrid approaches. Not all of them, it is pointed out, demand the same kind of synergistic relationship between neural and symbolic AI. Section 5 sketches a proposal for a hybrid system that makes use of both neural and symbolic processing but delegates more responsibility to the latter. Section 6 presents a novel heuristic constraint, whose aim is to help reduce the search space of hypotheses. Section 7 concludes with a summary of the main points.

2. Scientific Discovery: A Whirlwind Tour

In what follows, we make a brief foray into the scientific discovery debate.[1] Since the literature on this topic is vast and the space we have at our disposal is little, our foray is necessarily cursory. Still, we hope to cover some key methods and ideas, including inductivism, deductivism, abductivism, heuristics and causal discovery.

As already noted, for many years discovery was an overlooked topic in the philosophy of science. This was due in large part to the efforts of Popper ([1935/1959] 2002), who argued that there is no set of rules, not even rules of thumb, that can systematically lead a scientist to a discovery. In his own words:

> The initial stage, the act of conceiving or inventing a theory [or idea more generally], seems to me neither to call for logical analysis nor to be susceptible of it… there is no such thing as a logical method of having new ideas, or a logical reconstruction of this process. (7-8)

Popper was particularly critical of inductive logic, claiming that induction has no business in scientific investigations. The only logic that we could and should rely on, according to him, was deductive logic, but only in the context of justification, not discovery. If discovery doesn't have a logic, how does it come about? Popper thought by any means possible, including inspiration, dreaming, accident, etc.,

[1] Readers who wish to learn more about this debate may consult Ippoliti and Nickles (2020), Langley et al. (1987), Magnani (2011) and Nickles (2013).

so long as those means are not reducible to some logical or rationalistic recipe.[2] This is made clear in his assertion that "every discovery contains 'an irrational element', or 'a creative intuition', in Bergson's sense" (8).

But we are getting ahead of ourselves. Before we consider the kinds of views that Popper is objecting towards, we need to answer some basic questions about discovery. What does it mean to discover something? Simply put, it means to posit a thing's existence. There is, of course, no guarantee that what we posit actually exists. Hence, Popper, but also the vast majority of other scholars in this debate, is happy to talk of mere invention (see the quoted passage above), as much as he does of the actual discovery of things. We follow this broad usage of the term 'discovery', which includes mere invention.

The next basic question in need of an answer concerns what is being discovered. Besides, hypotheses, theories and ideas, we may also speak of entities, properties and relations, as well as laws, symmetries, principles, events and processes, among others. In other words, whatever can be posited is the object of discovery. Having said this, given that our attempt to understand discovery presented below is heavily logical in nature, it makes sense to simplify by speaking of the discovery of statements or hypotheses. All the preceding posits can, after all, be presented in that form. The caloric posit, for example, can be presented as the following hypothesis: 'There is a special kind of elastic fluid substance, the caloric, that flows from warmer to colder bodies, is virtually imperceptible, a conserved quantity, and subject to a repulsive force (between its particles) and an attractive force (between ordinary matter particles and its particles)'. Henceforth, and unless otherwise noted, we shall speak only of the discovery of statements or hypotheses.

Getting back to the logics that Popper excluded from playing any legitimate role in discovery, let us consider each of these in turn, beginning with deductivism. On this view, hypotheses are generated via deductive inferences. Here's a toy example: From the assumptions 'If all patients with virus x, exhibit symptom y' and 'If all patients with symptom y, exhibit symptom z', we can infer the hypothesis 'All patients with virus x, exhibit symptom z'. Despite Popper's misgivings, the approach has played a significant role in the history of scientific discovery in at least two ways. First, indirectly through mathematics. Arguably, mathematical discoveries can be made, at least when rationally reconstructed, purely via deductive methods. Since successful scientific theories are formulated (or at least formulable) in the language of mathematics, the discovery of the mathematical form of those theories is amenable to a deductive methodology. Second, it is a well-known fact that some notable figures in the history of science employed deductive inferences to help make discoveries. Newton ([1687] 1999), for example, famously proclaimed "[w]hatever is not deduced from the phenomena… [has] no place in experimental philosophy. In this philosophy, particular propositions are inferred from the phenomena, and afterwards rendered general by induction" (Book III, 'General Scholium').[3] This is no straightforward derivation of hypotheses from the phenomena.[4] Rather, as Worrall (2000) compellingly argues, this is a derivation that employs both evidence about the phenomena as well as general background assumptions (in this case about causal inference) to derive something like a hypothesis.

Deductivism is not a popular approach in discovery. The main reason for its unpopularity has to do with the fact that deduction is, by its very nature, non-risky, and, therefore, non-ampliative, i.e. not

[2] This seems to exclude logic (both inductive and deductive) as one of those methods through which discoveries may be made.

[3] Beyond such instances of the employment of deductive inferences to make discoveries, it is worth noting that Popper's falsificationism is a kind of deductivism, though it is not of course an account of scientific discovery, as he explicitly denounces logic in the context of scientific discovery.

[4] For any broad-scope hypothesis this would be impossible, as any evidence we have about the phenomena would never suffice to entail the hypothesis. See the point below about the non-ampliative character of deductive inferences.

content-increasing. After all, for the truth of the conclusion to be guaranteed by the truth of the premises, as it does in deductively valid inferences, it must be the case that the conclusion asserts nothing that is not already found in the premises. The elimination of risk is thus only obtained at the expense of being unable to create new content.[5] Although this is the main problem faced by deductivists, there are certainly various others, including the problem of irrelevant or trivial inferences (Votsis forthcoming), the mismatch with the non-monotonicity of actual human inferences, and the problem of determining the right (if any) deductive logic. There is no suggestion here that these problems are insuperable.

Next up is inductivism. Roughly speaking, this view holds that the generation of hypotheses is, or at least ought to be, achieved via inductive inferences. Unlike deductive inferences, inductive ones are ampliative. That is, they produce new content. Although there are precedents of this approach in antiquity, it really takes off in the work of early modern philosopher Francis Bacon, particularly in his *Novum Organum*. Bacon urges the formulation of hypotheses via inductive inference on the basis of, and after, (passively) observing the world. The inference he has in mind is enumerative induction, where, on the assumption that there are no observed exceptions, one generalises from some instances of a given type of object having a specific property to all instances of that type of object possessing that property. For example, from the assumption 'All observed patients with virus x, exhibit symptom y', we can infer the hypothesis 'All patients (observed and unobserved) with virus x, exhibit symptom y'. Inductivism has come a long way since then – for a survey see Norton (2005). It now incorporates data from experimental as well as observational studies. It also goes beyond enumerative induction to embrace all sorts of inductive inferences, including induction-to-the-next instance, retrodiction, analogical reasoning, Bayesian reasoning, causal reasoning, inferences from observables to unobservables, direct inference, and, more broadly, statistical inferences. In fact, some of these types of inferences/reasoning, e.g. causal reasoning and inferences from observables to unobservables, may best be construed under an abductivist approach – we return to abductivism below.

Several problems afflict the inductivist approach. These include the problem of induction, the argument from the underdetermination of theory by evidence, the old evidence problem, the problem of selecting priors, the problem of distinguishing causal relations from mere correlations, and the problem of choosing what to observe or experiment on without theoretical (i.e. top-down) guidance. Some of these problems are linked to specific versions of this view, e.g. the old evidence problem is only a problem for Bayesians, while others are more general, e.g. the underdetermination argument, afflicting not just different varieties of inductivism, but sometimes also other approaches to scientific discovery. As before, there is no suggestion here that these problems are insurmountable. Even the most entrenched are being gradually worn down by innovative methods, e.g. the problem of induction is being attacked by simulation-based optimality results (Schurz 2019).

A third and final logical approach to discovery is abductivism. This approach emphasises abductive inferences, i.e. ampliative inferences that are constrained by explanatory considerations (Lipton 1991). Such inferences take evidential statements as premises and offer hypotheses that attempt to best explain those statements as conclusions. For example, from evidential statements 'All known patients with virus x, exhibit symptom y' and 'All known patients with virus x, exhibit symptom z', and the supposition that there is no better explanation of that evidence, we can infer the hypothesis that 'Virus x is the cause of symptoms y and z'. Unsurprisingly, this type of inference also goes by the name 'inference to the best explanation'. Like inductive and deductive inferences, abductive ones are widespread in science and everyday life. The real question, as always, is whether they should play a role in scientific discovery.

[5] The exact opposite happens with induction: novel content is gotten at the expense of increased risk.

Relying as it does on ampliative inferences, abductivism inherits many inductivist problems, e.g. the underdetermination of theory by evidence argument. Having said this, the whole point of abduction is that it places additional constraints on inductive inferences to secure more reliable results. As such, abductivism is often thought of as a step in the right direction. Yet, it is not an easy step to make. That's because how the notion of a best explanation should be operationalised is controversial. It is generally agreed that the best explanations exemplify virtues like simplicity, unification and fruitfulness.[6] However, these latter notions are themselves the subject of fierce debate. As already noted, some forms of inference/reasoning are claimed to be best construed as abductive, i.e. not inductive. Typically, this happens whenever the concluded statement of an inference introduces concepts that are not present in the premises, e.g. referring to causes or unobservables. There are also questions about whether Bayesian reasoning is best construed in inductive or abductive terms. Van Fraassen (1989) insists that it cannot, but others, e.g. Henderson (2014) and Romeijn (2013), beg to differ.

Nothing prevents one from putting all three logical approaches in the service of scientific discovery. The only complication is that rules must be put in place to avoid conflicting judgments from being issued.[7] Beyond attempts to mix and match inference rules, it's also worth dwelling on the fact that none of the above approaches can achieve much without a heavy reliance on heuristics. These are imperfect techniques, strategies, and rules of thumb, whose aim is to aid discovery. They allow the reduction of otherwise vast spaces of possible solutions (e.g. hypotheses) to something more manageable that, ideally, contains optimal or indeed correct solutions. In fact, heuristics can also be thought of as distinct approach to scientific discovery, not just as a tool for the benefit of the aforementioned approaches, but as a fully-fledged solution finder itself (Ippoliti 2018a). Contributions to the literature on heuristics have been made by scholars across the fields of philosophy, computer science and psychology, and cannot all be covered here, but it is worth trotting out a few ideas to give the reader a glimpse.

Langley et al. (1987) draw a distinction between two types of heuristics: general and special. The former are widely applicable but weak as constraints. The latter are narrowly applicable but strong. One example of a general heuristic they discuss, in the context of Max Planck's search for a function that describes blackbody radiation, is the postulation of an interpolating function to approximate two existing functions that fit different ranges of the data well. An even more general heuristic, and thus more widely applicable and weaker as a constraint, is offered by Laudan (1981). He insists that content preservation (from predecessor theories) should not be a requirement of new hypotheses, and that a notion of progress can still be maintained, if we understand progress as the ability of successor theories to "solve more problems than their predecessors" (145). Perhaps the most famous heuristic of all is analogical reasoning (Hofstader 1995). As the name suggests, the key here is to exploit analogies. Suppose there are two domains that share certain properties or relations. We may wonder whether an additional property or relation that the one possesses is also present in the other. Making that inference is precisely what's involved in analogical reasoning. Examples of it abound in science. Analogical reasoning has been successfully employed to infer the mathematical form of laws in various domains, including inverse square laws as applied to the domains of gravity and electrostatics, and conservation laws as applied to the domains of energy, linear momentum, angular momentum and electric charge. Yet another general heuristic is offered by Ippoliti (2018b). The inversion heuristic involves "*reframing the problem in terms of its opposite (or invert[ing] the focus of a problem), and making the contrary assumption*" (201) [original emphasis]. Lakatos (1978) discusses the importance of heuristics in cases where empirical results contradict the conjunction of a central hypothesis and some auxiliaries. According to his negative heuristic, which is quite general, the central hypothesis

[6] Operationalisations of the virtue of unification can, for example, be found in Votsis (2015; 2017).

[7] Urbas and Jamnik (2014), for example, offer a computational framework where various types of inferences can be accommodated, though it must be noted that their concern is not tied to scientific discovery per se.

must be saved at the expense of the auxiliaries.[8] According to his positive heuristic, the auxiliaries can be modified through context-specific strategies. These strategies qualify as special heuristics in the sense specified above. One such strategy employed in the history of science was the postulation of additional planets in the solar system, thereby modifying auxiliaries about the number of such planets, to explain the irregular orbit of known planets, instead of modifying the law of universal gravitation. Finally, Gigerenzer and Murray (1987) offer a heuristic that breaks with traditional routes to discovery as outlined above, i.e. largely bottom-up data-driven routes like inductivism and largely top-down theory-driven routes like deductivism. Their third route to discovery relies on the specific tools/metaphors scientists happen to employ. One example is the postulation of the hypothesis that the mind is an intuitive statistician on the basis of the development of tools like Neyman-Pearson statistics and Fisher's ANOVA.

The final aspect of scientific discovery worth mentioning concerns the burgeoning field of causal discovery. At the centre of this field is Reichenbach's principle of common cause, which holds that the correlation between two events can be explained either by causal connection between them or by a third event which acts as a common cause. The principle enables us to make some headway in attempting to discriminate between mere correlations and causal relations. It has been generalised and developed into what is now known as the 'Causal Markov Condition' (Pearl 2000; Spirtes, Glymour, and Scheines 2000). The condition applies to probability distributions P defined over variables, which are represented (via Bayes nets) as a set of vertices V in a directed acyclic graph G.[9] The resulting graphical representation is known as the 'Causal Bayes Nets' framework. Researchers have employed this framework to successfully build causal models on the back of experimental data that encodes the effects of variable interventions on probability distributions (Eberhardt and Scheines 2007).

3. Symbolic vs. Neural AI: The Case of Scientific Discovery

A great deal of progress towards the automation of scientific discovery has already been made. Assuming that what machines do in such cases is fully explainable, the rationalist model of scientific discovery is still very much in play. To understand how scientific discovery can be automated, however, we must first understand what kind of AI is available and what it does best.

Broadly speaking, AI can be split into two types: symbolic and neural. Symbolic AI, roughly, is the computational implementation of logical or quasi-logical inferences to process symbolic representations. That means the inputs and outputs of such systems are symbolic representations, e.g. expressions in a natural or formal language, and the processing itself is symbolic, i.e. a rule-like manipulation of the aforesaid expressions. That makes what they do human-readable and resembling trains of reasoning we might ourselves construct or endorse. Examples of symbolic AI include expert systems (Engelmore and Feigenbaum 1993), SAT/SMT solvers (Alouneh et al. 2019), automated theorem provers (Harrison 2009) and, more recently, computational argumentation (Čyras et al. 2021). Neural AI, roughly, is the computational implementation of statistical or quasi-statistical inferences to process vectorial representations. The inputs and outputs here are numbers, but the processing resembles the kind of distributed cognition we find in the human brain. As such, what such systems do is not typically human-readable. Examples of neural AI systems include shallow and deep neural nets of various architectures, e.g. CNN, GAN, RNN, that may employ supervised (Nasteski 2017), unsupervised (Celebi and Aydin 206), semi-supervised (Van Engelen and Hoos 2020), self-supervised (Jaiswal et al. 2020) or reinforcement (Arulkumaran et al. 2017) learning algorithms.

[8] Lakatos also specifies the conditions under which this heuristic breaks down, i.e. where it is preferable to modify/replace the central hypothesis rather than the auxiliaries.

[9] It holds that, conditional on all of its direct causes, a variable X in V is independent of all the other variables in V, other than X's effects.

The above characterisation of the two approaches is somewhat crude. Among other reasons, that's because some symbolic systems marshal statistical or quasi-statistical inferences, while some neural systems reproduce logical or quasi-logical inferences. Moreover, AI is continually evolving, making the boundaries between the two approaches increasingly harder to ascertain. Having said this, typical examples of each AI approach are sufficiently distinct to merit differential treatment. What matters for the purposes of this section is how such typical examples fare in the arena of scientific discovery. If we go by overall trends, we might be forgiven for thinking that the symbolic approach, which was popular from the 1950s to the 1980s, is no longer viable and that all effort should be redirected to the neural approach. That's because the neural approach has been behind some pretty noteworthy successes, e.g. the protein folding exploits of AlphaFold and the discovery of new antibiotics like halicin, in the last ten years. As we will see below, such a thought would, however, be premature. Each approach certainly has its own strengths and weaknesses. We can conceive of these strengths and weaknesses as, to some extent, succeeding or failing to replicate characteristics we value in human intelligence. These include (in alphabetical order): adaptiveness, compositionality, efficient use of data, interpretability, learning from data, reasoning, universality, and unstructured data handling. In what follows, and due to space limitations, we restrict detailed comments to just four of them: adaptiveness, compositionality, reasoning, and unstructured data handling.

Adaptiveness is an important feature for AI systems to possess as it empowers them to adjust their modelling in light of new data. This is perhaps more clearly associated with the neural approach. A big part of the story here has to do with the non-monotonic nature of the processing of neural nets. Non-monotonic inferences, the kinds of inferences we find in statistics, are those whose 'goodness' can be reappraised with incoming data. By contrast, monotonic inferences, the kinds of inferences we find in classical systems of logic, are those whose 'goodness' is unimpeachable even if the data changes. Although symbolic systems are capable of employing, and indeed often employ, non-monotonic inferences, neural systems are guaranteed to be adaptive because their very architecture is non-monotonic by default. Weights, which signify the strength of the connections between nodes in a network, are continually being updated, adjusting to each new set of input data.

As a quick illustration of how adaptiveness may be useful to scientific discovery, let us briefly consider Kepler's postulation of the first law of planetary motion. Kepler employed Tycho Brahe's meticulous observations to determine the orbit of Mars. At first, he employed common metaphysical ideas, namely that orbits are circular or made up of compounded circular motions. Realising that these ideas could not accommodate all the data, Kepler went on to try something different, thereby breaking with a two-thousand-year-old tradition. He posited that orbits are elliptical, a mathematical notion that is closely related to the concept of a circle. This idea successfully accounted for the data, and was enshrined in his first law of planetary motion. The adaptive response of human intelligence is clearly at work in this case of scientific discovery, as it allowed Kepler to drop a hitherto unquestioned metaphysical assumption to reconcile theory with data.

Compositionality is a revered principle in the semantics and syntax of natural and formal languages. Simply put, it enables language users to build increasingly complex meaningful expressions from simpler ones. The principle also has an ontological-cum-mathematical analogue, namely a mereological principle of compositionality, according to which increasingly complex objects or parts can be built out of simpler objects or parts. To the extent that languages and the world itself exhibit compositionality, it is therefore important for AI systems to be able to process and exploit it. Symbolic systems are explicitly compositional, as their *modus operandi* is modelled after logical or quasi-logical languages, where complex expressions are constructed out of simpler expressions in a recursive manner. Neural systems, by contrast, need not, and often are not, compositional, or, able to handle compositionality. This has proved to be a thorn in the neural approach's backside, though it must be noted that efforts are underway to try to overcome this deficit (Baroni 2020; Hudson and Zitnick 2021).

Forthcoming in E. Ippoliti, L. Magnani, and S. Arfini (eds.), Model-Based Reasoning, Abductive Cognition, Creativity. Inferences & Models in Science, Logic, Language, and Technology, Springer.

The principle of compositionality plays a key role in scientific discovery. We here restrict our comments to the lasting effects of ancient Greek atomism. According to this school of thought, all matter is made up of smaller and smaller constituents, culminating in the most fundamental and indivisible units, called the 'atoms'. The atomistic conception of the world was largely embraced by philosophers (and later physicists). Newton, for example, asserted: "… the smallest particles of matter… compose bigger particles… and many of these… compose bigger particles…, and so on for diverse successions, until the progression ends in the biggest particles" (Optiks: Query 31). That embrace only loosened in the early twentieth century, when the requirement that atoms are the most fundamental and indivisible units of matter was dropped, but the requirement that matter is compositional maintained. Atoms are now thought to be composed of nuclei and one or more electrons. The nuclei themselves are composed of protons and neutrons and these in turn are composed of quarks. Ernest Rutherford takes credit for discovering atomic nuclei in 1911, and did so by, among other things, extending the compositionality we find in nature to a smaller scale. Compositionality is at play in scientific discovery to this very day, with strings being the latest posited entity to vie for the role of the most fundamental unit out of which everything else is composed.

The ability to reason is arguably the most important feature of intelligence. By reasoning, we mean processing information in a way that is attuned to norms evidence and does not violate norms of rationality. The former include norms like the principle of total evidence (Kelly 2008) and attitude aggregation rules (Dietrich and List 2010). The latter include criteria like logical consistency and probabilistic coherence (Olsson 2005). In its classical form, logical consistency imposes a concurrent satisfiability constraint on sets of statements: they must be such that they can all hold (i.e. be true) together. Other forms of logical consistency exist, captured under the umbrella term 'paraconsistency', and are the subject of various non-classical logics. Probabilistic coherence refers to the requirement that probabilities must not violate the axioms of probability. The probability of mutually exclusive and exhaustive statements, for example, must add up to one.

The symbolic approach is most clearly associated with reasoning as it involves a more direct and explicit attempt to reproduce it. This takes the form of endowing systems with the syntax and inference rules of formal and informal reasoning frameworks. Expert systems, for example, employ a knowledge base, where domain-specific facts and generalisations are formulated in the syntax of the given language, as well as an inference engine, which contains logical and/or probabilistic rules (Gillies 1996). The knowledge base and inference engine empower the system to model rational decision-making in relation to the chosen domain of expertise. The neural approach, by contrast, is much harder to describe as being reasoning oriented. Although it routinely implements some types of inferences, e.g. statistical inferences like regression, its inner workings are often unintuitive and opaque, resembling nothing like human reasoning. Moreover, even when we set aside the question of what goes on inside a neural net, it has proven difficult for neural nets to reliably reproduce long chains of reasoning.

That this desideratum needs motivation may beggar belief for some. Given what we said above about the anti-rationalist stance towards scientific discovery, let us nonetheless furnish it. We hereby omit the obvious cases of how reasoning aids discovery and focus instead on less obvious ones, e.g. those that concern serendipitous discoveries. Even in such cases, e.g. the discovery of X-rays by Roentgen in 1895 or the discovery of the hexagonal structure of benzene molecules by Kekulé in 1865, scientists need to provide a rationale for how phenomena fit into the existing body of knowledge. Consider the case of X-rays in a little more detail. The theory of electromagnetism was put forth by James Clerk Maxwell three decades prior to their discovery. It soon become orthodoxy, and, in so doing, allowed the integration of all sorts of phenomena under the same banner. Visible light, for example, could now be thought of as a form of radiation that occupied a certain region of the electromagnetic spectrum, namely fairly energetic with short wavelengths. From the gaps in the electromagnetic spectrum, one

could reason that other types of radiation should exist, covering different ranges of wavelengths and energies. That's exactly what happened with X-rays. They were slotted right into this spectrum, having higher energies and shorter wavelengths than visible and even ultraviolet light. More generally, reasoning was employed to ensure internal and external consistency within and between well-confirmed theories, as well as to establish explanatory relationships between them. Such reasoning is now integral to good scientific practice, as illustrated by the numerous scientific textbooks and research papers that reproduce it.

Successfully exploiting unstructured data is the last characteristic of intelligent behaviour to be considered here. This involves the ability to organise and categorise data as well as to recognise patterns in them. The neural approach comes closest to pulling this off, and perhaps even exceeds human abilities in some respects, as it leverages specialised methods. Unsupervised learning, for example, can be exploited to categorise data by creating new labels (categories) through methods like clustering. The symbolic approach, by contrast, appears to be unable to handle unstructured data. The very application of inferential rules presupposes that the data is already appropriately formatted in some symbolic form, e.g. expressed as formulae in a given logical language, implying that considerable structuring must already take place before processing.

This characteristic also plays an important role in scientific discoveries. A case in point is the discovery of Pluto's moon, Charon. Observational data, in the form of pictures, existed prior to Charon's discovery. They were recorded between the mid-1960s and the late 1970s to study Pluto's orbit. As such, there was no immediate reason to suspect that a natural satellite could be detected in those pictures. In 1978, while examining one of the pictures, James Christy noticed a fuzzy region next to Pluto. This spurred him to examine other pictures, leading to the realisation that the fuzzy region moved. Since the movement was periodic, it was justifiably inferred that Pluto had a moon, which came to be called 'Charon'. What is important, for our purposes, is that the data, though collected to determine Pluto's orbit, also contained sufficient information about the existence of Charon. That information, however, had to be discerned and treated as significant by the keen eye and sense of an astronomer. The ability to deal with unstructured data, or, to detect new structures inherent in data structured for different purposes, clearly needs to be replicated in machines.

We can provide a synopsis of the above assessment by means of Table 1, which also includes the other four desirable characteristics we excluded from detailed discussion. Very briefly, these are the characteristics of efficient use of data (i.e. how much data needs to be consumed before anything useful can be inferred), interpretability (i.e. how intuitive and transparent it is to understand the resulting models), learning from data (i.e. how data-driven is the method of constructing those models), and universality (i.e. how widely applicable are the methods employed).

AI system desiderata (relative satisfaction)	**Symbolic Approach**	**Neural Approach**
Adaptiveness		X
Compositionality	X	
Efficient use of data	X	
Interpretability	X	
Learning from data		X
Reasoning	X	
Universality (domain neutral)	X	
Unstructured data		X

Table 1: The Relative Satisfaction of Intelligent Behaviour Desiderata by Different Approaches to AI. This table provides a relative evaluation of the symbolic and neural approaches vis-à-vis eight desiderata that AI systems would need to satisfy if they are to exhibit human-level intelligent

behaviour, particularly behaviour that is beneficial to scientific discovery. The Xs represent significant progress and a current relative advantage of one approach over the other in meeting that desideratum.

A built-in assumption in the table is that exactly one of the two approaches has made significant progress in meeting the listed desiderata. This is an exaggeration, intended to highlight the differences between the two approaches. Moreover, the qualification 'current relative advantage' is there to leave open the possibility that the two approaches, when developed further, may be virtually indistinguishable with respect to the level of a given desideratum's satisfaction. In other words, the table roughly captures the current capabilities of symbolic and neural approaches, not their future or in-principle capabilities. For example, though their reasoning abilities may be dissimilar at present, they may converge at the limit. Overall, the table's purpose is to encode a snapshot of what many AI theorists and practitioners deem to be the current state of affairs in AI. As Schockaert and Gutierrez-Basulto put it: "Symbolic reasoning [symbolic systems] and deep learning [neural systems] are two fundamentally different approaches to building AI systems, with *complementary strengths and weaknesses*" (2021: 59) [added emphasis].

4. Hybrid Approaches

The main motivation for the proposed marriage between the neural and symbolic approaches to AI is their largely complementary strengths and weaknesses. A hybrid, also known as 'neuro-symbolic', approach thus raises the prospects of a powerful tool in the quest to both understand scientific discovery but also automate it. An early proponent is Valiant, who singled out the problem of characterising a semantics for cognitive computation as essential to computer science, and suggested that it can be solved via a hybrid system: "The aim here is to [integrate] the two most fundamental aspects of intelligent cognitive behavior: the ability to learn from experience [i.e. the neural net approach], and the ability to reason from what has been learned [i.e. the symbolic approach]" (Valiant 2003: 97). A growing number of AI theorists and practitioners are coming to the same conclusion. Marcus and Davis (2019), for example, claim that "When push comes to shove and the best AI researchers want to solve complex problems, they often use hybrid systems, and we expect this to become more and more the case" (p. 125). Kautz (2022: 118) seems to agree and argues that "the next big scientific advance in AI" will involve hybrid systems.

Asserting that a hybrid approach may stand a better chance at creating the conditions for truly rational, creative, and adaptive intelligent behaviour is only a starting point. What is needed next is a blueprint for how to integrate neural and symbolic systems. Following Kautz (2022), we consider six different integrations that are already practiced:[10]

1. Symbolic Neuro symbolic
2. Symbolic [Neuro]
3. Neuro | Symbolic
4. Neuro: Symbolic → Neuro
5. Neuro_{Symbolic}
6. Neuro [Symbolic]

Below, we provide a rough outline and examples of each of these. We then turn our attention to an assessment of their plausibility as carriers of the hybrid flag.

The first approach, Symbolic Neuro symbolic, consists of neural nets, whose inputs and outputs are symbols. Kautz describes this approach as "the deep learning SOP (standard operating procedure) for natural language processing" and gives as examples word2vec and GloVe (Pennington, Socher, and Manning 2014). These take sequences of words and convert them into vectors of numbers that can

[10] The names are due to Kautz.

then be processed by a neural net. The output of the network is then converted back into symbolic form. The only processing taking place is thus neural. As suggested in the quotation, the whole area of natural language processing qualifies as hybrid under Kautz's characterisation of this approach.

The second approach, Symbolic [Neuro], involves systems that at their core employ symbolic processing. What's special about these systems is that they control neural sub-systems that perform tasks like pattern recognition. One example is AlphaGo. It consists of a Monte-Carlo Tree Search algorithm, tantamount to a symbolic system, but also a neural sub-system that relies on deep reinforcement learning to carry out heuristic state estimations. Other examples abound, especially those underpinning "[m]ost robots and autonomous vehicles" (Kautz 2022: 118). Of key interest to us here is that in such hybrids, the neural system is subservient to the symbolic one.

The third approach, Neuro | Symbolic, involves neural and symbolic systems as equal partners, i.e. without a subservience relation between them. In Kautz's preferred phraseology, neither is a 'subroutine', but are rather 'co-routines'. The Neuro Symbolic Concept Learner (Mao et al. 2019) is one such example. This is a model, which learns (visual) concepts, words and even sentence parsing from images and image-related question-answer pairs. It does so without explicit supervision and employs both a neural module to extract object-level representations from the images, but also a symbolic reasoner module to process the question-answer pairs.[11] Another example is the NS-VQA model (Yi et al. 2018), which performs visual question answering by employing a neural system for visual recognition and a symbolic system for reasoning.

The fourth approach, Neuro: Symbolic $\rightarrow$ Neuro, centres around a neural system. Symbols come into play only to the extent that the system is trained on high-level symbolic representations like rules. A cited example is Charton and Lample (2020), where mathematical problems are used to train a neural net with a transformer architecture. The symbolic representations are gotten by converting mathematical expressions into trees and the trees into sequences. The latter are then fed into a seq2seq model for natural language processing. This seems to be the approach followed also in CORGI, a conversational neuro-symbolic system that mimics commonsense reasoning by identifying unstated assumptions. The system involves a neural theorem prover that is trained on traces of successful proofs (Arabshahi et al. 2021).

The fifth approach, Neuro_{Symbolic}, involves the conversion of symbolic structures into constraints that mould the shape (e.g. connectivity) of a neural net. An example is the logic tensor networks found in Serafini, Donadello, and Garcez (2017). They involve embedding first-order fuzzy logic formulae into neural nets as real-valued tensors – a process known as 'tensorisation'. When properly implemented, the networks behave like symbolic systems, leading to the approximation of reasoning under various kinds of logical calculi. Other examples along the same lines include rule-based knowledge representations like KBANN (Towell and Shavlik 1994) and formula-based knowledge representations, typically via restricted Boltzmann machines (Tran and Garcez 2016).

The sixth, and final, approach is Neuro [Symbolic]. At its core, this involves a neural system that employs and controls a symbolic sub-system to perform reasoning-related tasks. The operative analogy here is with dual process theories in psychology (Kahneman 2011; Rossi 2022), according to which the brain consists of two systems. System 1 is fast and intuitive, presumably like neural AI systems. System 2 is slow and logical, presumably like symbolic AI systems. It may also be claimed, as Kautz does, that the former is 'ultimately in charge', though that's a claim that dual process theorists need not endorse. It's worth noting that Kautz is partial towards this approach, asserting that it "has the greatest potential to combine the strengths of logic-based and neural-based AI" (2022: 119).

[11] It uses natural supervision, which is a form of self-supervision.

Kautz's classification of neuro-symbolic approaches, though helpful, is ultimately flawed. That's because the characterisation of some approaches as 'hybrid' is ill-fitting, as they do not involve symbolic processing in any meaningful sense, i.e. the logical or quasi-logical manipulation of symbols. To be precise, this is true of the first and fourth approaches. The only symbolic ingredient in these approaches is (some of) the 'ultimate' input and output. In both the Symbolic Neuro symbolic approach and the Neuro: Symbolic $\rightarrow$ Neuro approach, the processing is done exclusively by a neural engine. Another way to bring the point (i.e. that these approaches should not be classified as hybrid) home is that if we interpret numbers (strictly speaking, numerals) as symbols, as is commonly done in the philosophy of mathematics, then all neural nets would turn out to be hybrid. Clearly, this would reduce the distinction between hybrid and non-hybrid systems to absurdity. Still, we may strike a conciliatory tone here, conceding that the first and fourth approaches are hybrid but only in some weak sense of the term.

The fifth approach, Neuro_{Symbolic}, is potentially another exception. Instead of sticking merely to symbolic input and output, as the first and fourth approaches do, the fifth approach utilises symbolically formulated knowledge to constrain the shape of neural nets. As such, neural nets have the capacity to imitate symbolic processing, but do so by functional approximation. It's rather hard to say whether such systems should be classified as hybrid in a strong sense, like those we find in approaches two, three and six, or in a weak sense like those we find in approaches one and four. For that reason, we take the cautious route of classifying them as something in between.

Besides the strength of their hybridism, a simple way to describe how each of the above six approaches differs from the others is via the kind of processing they perform and, potentially, the relation of subservience between them. In what follows, we present a tentative classification of these approaches together with a few others that fill out the gaps. For expedience, we use the acronyms SS and NS for symbolic and neural systems respectively, and whenever applicable cite the corresponding name given to them by Kautz:

Strongly Hybrid (processing: neural and symbolic)
s1. SS is subservient to NS (Neuro [Symbolic]).
s2. NS is subservient to SS (Symbolic [Neuro]).
s3. Neither system is subservient to the other but working together (Neuro | Symbolic).

Moderately Hybrid (processing: neural imitating symbolic or else symbolic imitating neural)
s4. NS is constrained by symbolic rules (Neuro_{Symbolic}).
s5. SS is constrained by neural rules.

Weakly Hybrid (processing: neural or else symbolic)
s6. NS processing vectorised symbolic input into symbolic output (Symbolic Neuro symbolic).
s7. NS processing vectorised symbolic input into symbolic output with symbolic rule training (Neuro: Symbolic $\rightarrow$ Neuro).
s8. SS processing formalised neural input into neural output.
s9. SS processing formalised neural input into neural output with neural rule training.

Some guidance on how to read this classification is needed. First, no suggestion is made that the additional approaches (s5, s8 and s9), i.e. those not included in Kautz's original list, are in use and/or desirable. Second, note that even if brains are organised in a manner akin to s1, as suggested by Kautz, this need not mean that the future for AI is paved along this way. For all we know, any of the above approaches may be an advisable way forward. Third, note that even more approaches can be generated by nesting the existing ones, e.g. a system exemplifying s6 may be nested within a broader system that exemplifies s2. Fourth, note that there is nothing about the natures of either the neural

or the symbolic architectures that prohibits them from mimicking each other to a tee. In fact, abstract or concrete demonstrations of these kinds of imitations, both broad scoped and narrow, are fairly common in the literature. Trivially, (nearly) all neural nets are run on digital computers whose basic operations are symbolic in nature, since digital circuits are made up of logic gates. Conversely, the domain of universal function approximation contains theorems whose purpose is to establish the corresponding capabilities and limits in different types of neural nets (e.g. Maiorov and Pinkus 1999).

5. A Proposed Hybrid System

In this section, we propose a system best characterised as falling under s2, though with something like s6 or s7 nested within. The system proceeds by primarily exploring the space of (interesting) derivations. It is: (i) neural in that it can extract *some* symbolic representations via neural nets and (ii) symbolic in that it employs derivations to produce novel symbolic representations. It should be emphasised that the system is neither (fully) descriptive of actual scientific discovery nor complete with respect to all the problems of discovery. It is described in broad strokes, leaving many difficult issues untouched, though hopefully throwing sufficient light on others.

Scientific discovery through derivation is the idea that, once a scientist selects the right assumptions, they can derive a hypothesis merely through the application of logical inferences. Newton's law of universal gravitation, for example, can be derived from several assumptions including his second law of motion and Kepler's third law of planetary motion.[12] Selecting the right assumptions and even carrying out the derivation are both challenging parts of the task. That task becomes even more challenging, however, when some of the assumptions are missing. Imagine, for the sake of the argument, that Kepler's third law was unavailable to Newton. Would he have still discovered the law of universal gravitation? Though we cannot answer that question with any degree of certainty, we can nonetheless say that he would not have been able to derive it, due to the missing assumption. These kinds of cases, where assumptions are missing, are prevalent in science. Moreover, derivations of hypotheses are often constructed, though often only in retrospect, i.e. after the hypotheses have been discovered in a non-derivational manner. Still, so long as there is a derivational path from the assumptions to the hypotheses, we can rationally reproduce the said discoveries.

In what follows, we focus on cases such as the above and show how derivational methods may still be valuable. We thus describe scientific discovery through derivation as a two-step process:

> **Step 1** (optional): Postulate potentially relevant missing statement(s), assumptions A_n, …, A_{n+m} or a hypothesis H_i, via neural net or symbolic methods to supplement those that are available and relevant, assumptions A_1, …, A_{n-1}, in relation to a domain of inquiry D_i.[13]
>
> **Step 2**: Employ combinations of those available statements (including the ones that were added in step 1) to derive additional missing statement(s), assumptions A_n, …, A_{n+m} or hypothesis H_i, via symbolic methods (e.g. automated theorem provers) thereby putatively making a discovery.

Note that since the output of this process may be the derivation of either a novel hypothesis or a novel assumption, the description employs the neutral term 'statements' instead. Moreover, note that the phrase 'thereby putatively making a discovery' conveys the qualification that not every derivation of a statement is tantamount to a discovery. Many derived hypotheses, for instance, will be unilluminating,

[12] Many derivations between successor and predecessor theories involve approximate or de-idealised versions of older assumptions. We set this very important issue aside as it has been explored in several other works. For details about the continuities and discontinuities between Kepler's and Newton's assumptions, the reader may consult Rott (1990).

[13] In most cases, there will be some available assumptions and some missing. Starting from zero available assumptions is also possible but it makes the task considerably harder.

uninteresting or even plain false. It's also worth noting that, by 'available and relevant', we mean statements that have fairly substantial evidence accumulated in their favour.[14] Finally, the two-step process is a but a rough blueprint for discovery. One lacuna is the under-description of step 1. If we replace the term 'postulate' with the term 'discover' in step 1, then the two-step process seems to be suggesting that we may need to make some discoveries before we can make others. Although this is not strictly speaking circular, as different discoveries are involved in steps 1 and 2, we still need to specify some methods for postulating statements in step 1. We give some initial guidance in this section and the next.

Let us begin with step 1. As noted in section 2, some of the key benefits of employing neural nets include their adaptiveness and ability to handle unstructured data. Overall, neural nets involve data-driven methods that allow us to more easily detect patterns, construct concepts and build models of target systems. As such, they are invaluable in our efforts to understand the world around us, especially when we have little to go on, as is often the case with step 1. So, although both symbolic and neural net methods may be employed in this step of the process to postulate missing statements, we are more likely to fall back on the later when guidance is at its lowest. One particularly powerful tool in the neural net toolbox is the ability to extract novel features (read: variables) from data.[15] As Kautz (2022) notes, one major motivation and insight of deep neural nets "was to eliminate the need for manually engineered features" (112). Extracting novel features brings all sorts of benefits to the table, not least of which is the ability to shed fresh light on data, leading to the detection of patterns, the construction of concepts and the building of models that would otherwise remain unconceived.

Besides the challenge of looking for structure in the data with neural net methods, there is also the challenge of extracting symbolic representations from them. There are generally two ways of doing so. One is fairly straightforward and involves their outputs, the other is fraught with complications and involves the models constructed with such methods. Let us take the straightforward way first. As already noted in the previous section, weakly hybrid systems like s6 and s7 can be employed to process vectorised symbolic input into symbolic output. Most deep learning natural language processing systems work this way. One way to exploit such systems is by training a model in a specific domain of science via peer-reviewed papers, and then tasking the system to produce a summary of all the results and conjectures in that domain. Any statement produced that is not already included in our original list of available assumptions can then be added to it. In step 2, one can employ combinations of those statements to make further postulations and potentially discoveries.

If postulations/discoveries can already be made in step 1, as suggested above, why bother with step 2? Two reasons can be given in reply. First, trying to discover more is surely not a bad thing. Second, the goodness of a postulation in step 1 needs to be validated. A great way to do so is by deriving additional statements from those postulates plus available statements, and checking whether these additional statements yield interesting predictions. If the predictions are confirmed, then we have added reason to accept the postulates made in step 1. If, on the contrary, they are disconfirmed, then we have reason to reject them. That's why step 2 is needed. Another way of saying roughly the same thing is that our knowledge ultimately needs to be systematically organised, and the most systematic means of organisation is deduction.

We can now turn our attention to the second way in which we can extract symbolic representations from neural nets, namely models they produce. As already noted, this way is fraught with

[14] This condition may be loosened, though it is worth bearing in mind that this would lead to an increase of false positives.

[15] Though it should be noted that feature extraction is not exclusively tied to neural nets. One intriguing example is the Causal Feature Learning framework (Chalupka, Eberhardt and Perona 2017), which constructs macrovariable causal hypotheses out of microvariable inputs.

complications. Two complications stand out. Models produced with neural nets are often uninterpretable, or at least not easy to interpret. Moreover, they are unnecessarily complex. In what follows, we consider each complication in turn.

The uninterpretability or 'black box' nature of many neural net models is well-known. Note that this is a problem that needs to be solved, but it is not a problem that is unique to the system proposed in this section. Even so, it is worth touching upon here. Some promising efforts in addressing this problem are currently underway – for an overview, see Zhang et al. (2021) – and they involve either constraining the design of neural nets or post-hocly explicating or modifying them to increase their interpretability. One popular method relies on the post-hoc extraction of rules expressed in logical form. For example, we may make a model (more) interpretable by extracting "a decision rule set" consisting of the following individual rules "If ($x_2 < \alpha$) $\wedge$ ($x_3 > \beta$) $\wedge$ . . . , then $y = 1$" and "If ($x_1 > \gamma$) $\wedge$ ($x_5 < \delta$) $\wedge$ . . . , then $y = 2$" and so on, concluding with "If (x_4 . . .) $\wedge$ (x_7 . . .) $\wedge$ . . . , then $y = M$" (Zhang et al. 2021: p. 6), where the xs denote the input features, y denotes the output feature, and the Greek letters denote ranges of the xs for which y yields the same value. Other methods involve uncovering/imposing semantics on the neural nets but also assigning credit/blame to input features on the basis of their importance in accurately predicting the output. Although it's not clear how much of the black box problem can be mitigated through these efforts, they at least give us hope that it is not a completely intractable one.[16]

The other complication, unnecessary complexity, is also a problem that is not unique to the system proposed in this section, and it is, in fact, closely related to the issue of interpretability. That's because part of the difficulty of interpreting such models concerns their complexity. By complexity here we mean things like the number of learnable parameters (i.e. weights), connections and nodes. For example, some of the most successful models have billions of weights. Taking such models and directly converting them into a symbolic representation would be utter madness, as no symbolic reasoner would be able to do anything useful with them in a sensible amount of time. Thankfully, there are ways that attempt to deal with this problem too. We can do so either before we build a model (through careful design) or after. One fruitful method that takes the latter path is pruning. It involves the selective removal of parameters or nodes from models, which are then trained/retrained faster using the original datasets (Le Cun, Denker and Solla 1989). The upshot is the production of models that are "less than 10-20% of the [original model's] size" and that enjoy comparable accuracies (Frankle and Carbin 2018: 1). More impressively, evidence is surfacing that pruning can result in models that exceed the accuracy of their pre-pruned rivals. Other methods involve weight clustering (i.e. reducing the number of unique values that weights can take) and quantisation (i.e. reducing the number of bits that are needed to express the parameters) – see Freire et al. (2023) for an overview. Once again, there is no clear sense of how much of the complexity problem can be mitigated through these methods, but there is definitely hope that future iterations of models may be significantly simpler.

Let us now turn to step 2. As noted in section 3, some of the key benefits of employing symbolic methods is their ability to automate reasoning as well as make it explicit and human-readable. Powerful symbolic methods like automated theorem provers have been around since the 1950s, way before anything sophisticated was achieved via neural nets. The Logic Theorist, for example, was a program developed by Allen Newell, Herbert Simon and Cliff Shaw in 1956 and was used to successfully prove dozens of theorems present in Whitehead and Russell's *Principia Mathematica*. More powerful proving methods and systems have since been implemented for various purposes, including scientific discovery, most notably in expert systems. In what follows, we consider two cases where automated theorem proving methods can be utilised to implement step 2 of the above process. The first of the two cases concerns the derivation of a missing assumption with the help of a hypothesis, while the second concerns the derivation of a missing hypothesis from some assumptions. Since we do not generally know whether we have enough assumptions, or, indeed, whether we have assumptions that

[16] For a brief discussion and critique of methods like these, the reader may consult Sullivan (2022).

are true enough to derive a statement that is illuminating, interesting or has some non-trivial truth content, the proposed procedure can merely provide candidates for such missing assumptions or hypotheses. To use an earlier expression, the result of the procedure is merely a 'putative discovery'.

We start with the former case, as it is a little simpler to motivate. Suppose, for the sake of the argument, that an illuminating, interesting or truth-containing hypothesis H_i has been posited and can be partly grounded in available assumptions $A_1 \wedge \ldots \wedge A_{n-1}$.[17] Suppose, moreover, that no derivation from those assumptions to that hypothesis can be performed without first positing some extra missing assumption(s). There is a way to find those assumption(s) by exploiting the reversibility/inter-derivability capacity of deduction. In the case at issue, this means that hypothesis H_i must be derivable from the conjunction of some combination of assumptions $A_1, \ldots, A_n$, where A_n is missing (i.e. currently unknown), but also that the same conjunction must be derivable from H_i. We may also add a 'material' requirement here (Schurz 1991) to avoid repeating content that is redundant or irrelevant, e.g. adding a tautology to our assumptions brings nothing of value to the derivation. In what follows, we assume that redundant and irrelevant content is prohibited.

To achieve the derivation of the missing assumption, the following conditions must be met:

1. $H_i \vdash A_1 \wedge \ldots \wedge A_{n-1}$
2. $A_1 \wedge \ldots \wedge A_{n-1} \nvdash H_i$

For expedience, we here assume that the combination of conjuncts at issue involves all the available assumptions. From the satisfaction of these two conditions, we can then infer that there is at least one missing assumption, which we can encode as a singular statement A_n, which when conjoined to $A_1 \wedge \ldots \wedge A_{n-1}$ results in the inter-derivability between H_i and $A_1 \wedge \ldots \wedge A_n$. More formally:

3. $H_i \dashv\vdash A_1 \wedge \ldots \wedge A_n$

Intuitively, we can understand A_n as that part of the content of H_i, which, when added to the available assumptions $A_1 \wedge \ldots \wedge A_{n-1}$, makes up for the gap between the content of those assumptions and the full content of H_i. More formally, A_n is the relative complement of $A_1 \wedge \ldots \wedge A_{n-1}$ in H_i and can be expressed thus:

4. $A_n: H_i \cap (A_1 \wedge \ldots \wedge A_{n-1})^C$

The derivation of A_n can be mechanised by instructing an automated theorem prover to look for a maximally general (i.e. logically strongest) statement: (i) that is derivable from H_i and (ii) none of whose 'material' parts (i.e. non-redundant and relevant logical consequences), including the maximally general statement itself, is derivable from $A_1 \wedge \ldots \wedge A_{n-1}$. In simpler terms, the automated theorem prover will be tasked to find that statement following from H_i whose content is greatest and does not overlap with any of the content present in $A_1 \wedge \ldots \wedge A_{n-1}$.

Let us now consider the second case. This concerns the derivation of a missing hypothesis from some assumptions. We can employ the same reversibility capacity of deductive inferences to postulate the hypothesis. As with the first case, hypothesis H_i and the conjunction of assumptions $A_1 \wedge \ldots \wedge A_n$ must be inter-derivable: $H_i \dashv\vdash A_1 \wedge \ldots \wedge A_n$. Unlike the first case, $A_1, \ldots, A_n$ are all available (since A_n has already been postulated) but H_i is missing. To achieve the derivation of the missing hypothesis, some adjustments to the assumptions would already need to be in place. That's because successor theories

[17] By truth-containing hypothesis, we mean a hypothesis that has some truth content. Such hypothesis may be true, approximately true or partially true. For a discussion of the differences between these notions, the reader may consult Niiniluoto (1999).

typically introduce new concepts that are absent in their predecessors but also apply some corrections to the conjectured relations between concepts. Newton's law of universal gravitation introduces the concepts of (an accelerating) force, mass, and gravitational constant, the former of which is implicit in Galileo and Kepler but the latter two absent. Moreover, Newtonian physics corrects Galilean and Keplerian physics by adjusting some of the relations between their concepts. For example, planetary orbits are no longer perfect ellipses, as Kepler dictated, for the simple reason that they are perturbed by gravitational accelerations generated by all sorts of bodies with mass. How we get to these concepts and corrections is a non-trivial problem that we set aside for the purposes of this paper.[18] Instead, we bypass it by focusing on the already adjusted assumptions available to Newton. More generally, we focus on that stage in the process of scientific discovery where the available assumptions utilise the new concepts and the requisite corrections.

Given the assumed inter-derivability between H_i and $A_1 \wedge \ldots \wedge A_n$, we can express the content of our target hypothesis as follows:

1. H_i: $A_1 \wedge \ldots \wedge A_n$

Articulated thus, H_i is not edifying. To become edifying, we need to restate it in a manner that provides conceptual economy and unity. Once again, our example serves as a trusted compass. Newton's law of universal gravitation (but also his laws of motion) does not just repeat the corrected versions of Galileo's terrestrial laws and Kepler's celestial laws. Rather, it consolidates those versions of the laws into a conceptually simple and integrated package. To replicate this manoeuvre, we can instruct our symbolic reasoner to look for that maximally general (i.e. logically strongest) statement which is inter-derivable with $A_1 \wedge \ldots \wedge A_n$ but which also employs the smallest number of concepts already present in those assumptions. The result should be Newton's law of universal gravitation, as without the notions of force, mass, distance and the gravitational constant, one presumably cannot derive all and only the consequences related to that law, without deriving the law itself.

The aim of this section was to put forth a neuro-symbolic system, best characterised as falling under system s2 (with some nesting from s6 or s7). The proposed 2-step system, it was claimed, holds promise in automating at least some portions of scientific discovery. It exploits symbolic and neural methods, by playing to each tradition's strengths, to derive potentially interesting, illuminating or truth-containing statements. Some complications (interpretability and complexity) arising from the attempt to combine neural net methods with symbolic ones were discussed. It was argued that there are potentially successful technical countermeasures, offering hope that the complications are not insurmountable. At any rate, extracting symbolic representations from *models* produced by neural nets is not necessary, as the *output* of NPL models may also suffice for the purpose of postulating additional statements (step 1). Finally, by way of demonstration, we considered two cases where such methods can be utilised to produce missing statements (step 2): the derivation of a missing assumption with the help of a hypothesis, and the derivation of a missing hypothesis from some assumptions.

6. The Structural Correspondence Heuristic

The methods for generating new statements in step 1 are clearly quite liberal. Similarly, the combination of available statements in step 2 to derive new statements can result in too many possible derivations, especially if we take that liberal stance in step 1. Both of these problems can be mitigated with aptly selected heuristics.[19] These are routinely employed in the field of computational scientific discovery to reduce the search space of possible solutions – in our case, hypotheses or auxiliaries. We

[18] Gardenfors (2004) offers a putative way forward in constructing concepts that may then be employed in conjunction with the current proposal.

[19] For an overview of heuristics in the field of computational scientific discovery, the reader may consult Sozou et al. (2017).

already saw several of these in section 2 of this paper. In this section, we will discuss the structural correspondence heuristic, a general heuristic that, to the best of our knowledge, is original to this paper. This heuristic is directed towards symbolic methods and may help, together with other heuristics, both steps of the proposed discovery process.

The structural correspondence heuristic begun life as an observation that successor theories are, in part, structurally continuous to their predecessors. Several successor-predecessor theory pairs in the history of science seem to exhibit this continuity, including the following pairs:

> the theory of electromagnetism – Frensel's wave theory of light (Poincaré 1905)
> the oxygen theory of combustion – the phlogiston theory of combustion (Schurz and Votsis 2014)
> the kinetic theory of heat – the caloric theory of heat (Votsis and Schurz 2012).

As has been argued elsewhere, the reason for these continuities is that successor theories need to be at least as successful as their predecessors, if the latter are indeed successful, and, thus, must preserve those aspects of their structure that are responsible for that success.[20] This reason appears to have played a role in the construction and development of various structurally continuous theories by scientists such as Newton, Young, Bohr, Clausius, Dirac, Heisenberg, Newton and Young (Fadner 1985). It is thus clearly a heuristic worth trying out and can be applied in conjunction with other heuristics, affecting both steps of the discovery process.

To shed light on how this heuristic works, and thus how it can be automated, we need to decompose its process into three stages. The first stage involves the determination whether a predecessor theory is empirically successful. By empirical success, we mean that the theory makes accurate predictions about some domain of phenomena. If it is indeed successful, then one must try to track down those of its structures that are responsible for that success. That's the second stage. By structures, we mean any implicit or explicit mathematical relations like equations, laws or principles. The determination of where to lay the credit for success requires the piecemeal differential removal of content from a theory to determine which parts do what predictively. Oftentimes, scientists themselves arrive at a good idea of what's behind the success, through a series of experiments that probe the various parts of a theory, or, at the very least, through the disagreement encountered over the success-producing abilities of some theory parts. The case of the luminiferous ether is instructive here, as it turned out to be an unnecessary posit that can be jettisoned without losing any of the empirical success enjoyed by Fresnel's wave theory of light or Maxwell's theory of electromagnetism. The third stage concerns what we do we with those structures once they are identified. Simply put, we try to embed them in a successor theory, at least in some limit form. The reason for the 'limit form' qualification is that newer structures often correct older ones and may thus be only approximately continuous with them in certain ranges of phenomena. One example of such a limiting relation between theories can be found in the statistical relation between classical frequency and quantum frequency in the limit of large quantum numbers (Bokulich 2008).

All of the above stages can be automated, though we need only consider the second and third stages here, for we can simply suppose that we already have a store of various scientific statements, which includes empirically successful theories. The second stage involves the determination of where to lay the credit for that success. This can be automated by first breaking down a theory into its basic content parts and then checking which parts are responsible for that success. As we saw earlier, any hypothesis or theory is logically equivalent to a conjunction of 'material' assumptions. These can be considered

[20] Note that this heuristic is in direct conflict with Laudan's own heuristic. Moreover, note that in the philosophy of science, the structural correspondence between successor-predecessor pairs has been used as a reason to support either realist or anti-realist views. We avoid this issue here but simply assume that some sort of structural correspondence may be useful in theory construction and development.

its content parts. To break down a theory into its basic content parts, we may simply turn to our store of scientific statements to dig out any assumptions relevant to the given domain.[21] We can then test which combination of those assumptions are individually necessary and jointly sufficient for the success enjoyed by that theory. This can be accomplished by deriving all and only the successful consequences of that theory via the given combination of assumptions. Once again, an automated theorem prover can be made use of here to assemble the requisite proofs. The third stage involves the use of the selected assumptions as 'limit form' prerequisites for the construction of the missing statements. This means that we should be able to derive 'limit form' versions of those assumptions from any candidate missing statement. Once in place, the structural correspondence heuristic may thus help whittle down the search space.

7. Conclusion

In this paper, we explored the prospects of automating some aspects of scientific discovery through AI. We did so by first examining the merits and demerits of symbolic vs. neural net approaches in AI. It was argued that, at least on the face of it, symbolic approaches excel at certain tasks (e.g. reasoning), while neural net approaches excel at others (e.g. pattern detection). We then turned our attention to hybrid approaches, which seek to blend the two types of approaches together in order to get the best of both worlds. Several such neuro-symbolic blends were considered, but not all of them were hybrid in the strong sense of involving both symbolic and neural processing. Of those that were deemed strongly hybrid, one was singled out as holding the most promise, namely s2. On this approach, a symbolic system is in control while a neural system plays a subservient role. A proposal was sketched along these lines, where computational scientific discovery is conceived of as a two-step process. The first step seeks the postulation of missing statements, either through symbolic or neural net means. In the second step, all available and relevant statements are fed into a symbolic reasoner to derive additional statements, typically hypotheses, which make for putative scientific discoveries. As the space of possible statements that can be generated by this method is vast, some heuristic constraints will need to play a confining role in this process. The paper concluded with a proposal for a new heuristic. Our hope is that, when jointly deployed with other heuristics, structural correspondence may help provide an efficient and effective way to automate some aspects of scientific discovery.

Acknowledgements

I gratefully acknowledge the input of several scholars in the form of direct or indirect feedback and discussion. They include Pat Langley, Jan-Willem Romeijn, two anonymous referees, and the audiences at the 2023 AAAI Spring Symposium on Computational Scientific Discovery (San Francisco), the Epistemology of Data-Intensive Science Seminar (Seoul National University), and the Mini-Symposium on Medical Misinformation at the Centre for Reasoning (University of Kent), where versions of this paper were presented. Last but not least, I gratefully acknowledge the editorial team, Emiliano Ippoliti, Lorenzo Magnani and Selene Arfini, for their stellar work and patience.

References:

Alchourrón, C. E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(2), 510-530.

Alouneh, S., Abed, S. E., Al Shayeji, M. H., & Mesleh, R. (2019). A comprehensive study and analysis on SAT-solvers: advances, usages and achievements. *Artificial Intelligence Review*, 52, 2575-2601.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6), 26-38.

Bokulich, A. (2008). *Reexamining the quantum-classical relation: Beyond reductionism and pluralism*. Cambridge: Cambridge University Press.

[21] Seeing as more fine-grained content parts may lead to a more fine-grained determination of where to lay the credit, we may also attempt to break down assumptions into more basic assumptions (i.e. assumptions with less content).

Celebi, M. E., & Aydin, K. (Eds.). (2016). *Unsupervised learning algorithms* (Vol. 9, p. 103). Cham: Springer.

Cellucci, C. (2013). *Rethinking logic*. Dordrecht: Springer.

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: an overview. *Behaviormetrika*, 44, 137-164.

Čyras, K., Rago, A., Albini, E., Baroni, P., & Toni, F. (2021). Argumentative XAI: a survey. *arXiv preprint arXiv*:2105.11266.

Hudson, D. A., & Zitnick, L. (2021). Compositional transformers for scene generation. *Advances in Neural Information Processing Systems*, *34*, 9506-9520.

Dietrich, F., & List, C. (2010). The aggregation of propositional attitudes: towards a general theory. *Oxford studies in epistemology*, 3, 215-234.

Eberhardt, F., & Scheines, R. (2007). Interventions and causal inference. *Philosophy of science*, 74(5), 981-995.

Engelmore, R. S., & Feigenbaum, E. (1993). Expert systems and artificial intelligence. *Expert Systems*, 100(2).

Frankle, J., & Carbin, M. (2018). 'The lottery ticket hypothesis: Finding sparse, trainable neural networks'. *arXiv preprint arXiv*:1803.03635.

Freire, P. J., Napoli, A., Ron, D. A., Spinnler, B., Anderson, M., Schairer, W., ... & Prilepsky, J. E. (2023). Reducing computational complexity of neural networks in optical channel equalization: From concepts to implementation. *Journal of Lightwave Technology*.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Gillies, D. (1996). *Artificial intelligence and scientific method*. Oxford University Press.

Harrison, J. (2009). *Handbook of practical logic and automated reasoning*. Cambridge University Press.

Henderson, L. (2014). Bayesianism and Inference to the Best Explanation. *The British Journal for the Philosophy of Science*, 687-715.

Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York: Basic books.

Ippoliti, E. (2018a). Building Theories: The Heuristic Way. In D. Danks and E. Ippoliti (Eds.), *Building theories: heuristics and hypotheses in sciences* (pp. 3-20). Springer.

Ippoliti, E. (2018b). Heuristic logic. A kernel. In D. Danks and E. Ippoliti (Eds.), *Building theories: heuristics and hypotheses in sciences* (pp. 191-211). Springer.

Ippoliti, E., & Nickles, T. (2020). Introduction: Scientific Discovery and Inference. *Topoi*, 39, 835-839.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2.

Kautz, H. (2022). The third ai summer: Aaai robert s. engelmore memorial lecture. *AI Magazine*, 43(1), 105-125.

Kelly, T. (2008). Evidence: Fundamental concepts and the phenomenal conception. *Philosophy Compass*, 3(5), 933-955.

Lakatos, I. (1978). The methodology of scientific research programmes. Cambridge: Cambridge University Press.

Lample, G., & Charton, F. (2019). Deep learning for symbolic mathematics. *arXiv preprint arXiv*:1912.01412.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery*. MIT Press.

Langley, P. and Arvay, A. (2019) 'Scientific Discovery, Process Models, and the Social Sciences', In *Scientific discovery in the social sciences* (pp. 173-190). Springer, Cham.

Laudan, L. (1981). A problem-solving approach to scientific progress. In I. Hacking (Ed.), *Scientific revolutions* (pp. 144–155). Oxford: Oxford University Press.

LeCun, Y., Denker, J., & Solla, S. (1989). Optimal brain damage. *Advances in neural information processing systems*, vol. 2.

Lipton, P. (1991). *Inference to the Best Explanation*. London: Routledge.

Magnani, L. (2011). *Abduction, reason and science: Processes of discovery and explanation*. Springer Science & Business Media.

Maiorov, V., & Pinkus, A. (1999). Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1-3), 81-91.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv*:1904.12584.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons*. b, 4, 51-62.

Newton, I. ([1687] 1999). The Principia: mathematical principles of natural philosophy. Univ of California Press.

Nickles, T. (1980). *Scientific discovery: Logic and rationality*. Boston: Springer.

Nickles, T. (2013). Scientific discovery. In *The Routledge companion to philosophy of science* (pp. 497-506). Routledge.

Niiniluoto, I. (1999). *Critical scientific realism*. OUP Oxford.

Norton, J. D. (2005). A little survey of induction. In Peter Achinstein (Ed.), *Scientific Evidence: Philosophical Theories and Applications* (pp. 9-34). John Hopkins University Press.

Olsson, E. J. (2005) *Against Coherence: Truth, Probability, and Justification*, Oxford: Clarendon Press.

Poincaré, H. ([1905] 1952). *Science and hypothesis*. New York: Dover.

Popper, K. ([1935/1959] 2002). *The Logic of Scientific Discovery*. New York: Basic Books.

Romeijn, J. W. (2013). Abducted by Bayesians?. *Journal of Applied Logic*, 11(4), 430-439.

Rose, D., & Langley, P. (1986). Chemical discovery as belief revision. *Machine Learning*, 1, 423-452.

Rott, H. (1990). Approximation Versus Idealization: the Kepler-Newton Case. *Poznán studies in the philosophy of the sciences and the humanities*, 17, 101-124.

Russell, B. (1927). *The Analysis of Matter*. London: George Allen & Unwin.

Schockaert, S., & Gutiérrez-Basulto, V. (2021). Modelling symbolic knowledge using neural representations. In *Reasoning Web International Summer School* (pp. 59-75). Cham: Springer International Publishing.

Schurz, G. (1991). Relevant deduction. Erkenntnis, 35, 391–437.

Schurz, G., & Votsis, I. (2014). Reconstructing scientific theory change by means of frames. in T. Gamerschlag et al. (eds.), Frames and Concept Types, *Studies in Linguistics and Philosophy*, vol. 94, 93-109.

Schurz, G. (2019). *Hume's problem solved: The optimality of meta-induction*. Mit Press.

Sozou, P. D., Lane, P. C., Addis, M., & Gobet, F. (2017). Computational scientific discovery. *Springer handbook of model-based science*, 719-734.

Sullivan, E. (2022). Inductive Risk, Understanding, and Opaque Machine Learning Models. *Philosophy of Science*, 89(5), 1065-1074.

Towell, G. G., & Shavlik, J. W. (1994). Knowledge-based artificial neural networks. *Artificial intelligence*, 70(1-2), 119-165.

Tran, S. N., & Garcez, A. S. D. A. (2016). Deep logic networks: Inserting and extracting knowledge from deep belief networks. *IEEE transactions on neural networks and learning systems*, 29(2), 246-258.

Urbas, M., & Jamnik, M. (2014). A framework for heterogeneous reasoning in formal and informal domains. In *Diagrammatic Representation and Inference: 8th International Conference, Diagrams 2014*, Melbourne, VIC, Australia, July 28–August 1, 2014. Proceedings 8 (pp. 277-292). Springer Berlin Heidelberg.

Valiant, L. G. (2003). Three problems in computer science. *Journal of the ACM (JACM)*, 50(1), 96-99.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, 109(2), 373-440.

Van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford, UK: Clarendon Press.

Votsis, I., & Schurz, G. (2012). A frame-theoretic analysis of two rival conceptions of heat. *Studies in History and Philosophy of Science Part A*, 43(1), 105-114.

Votsis, I. (2015) 'Unification: Not Just a Thing of Beauty', *Theoria: An International Journal for Theory, History and Foundations of Science*, vol. 30(1): 97-114.

Votsis, I. (2017) 'Unification through Confirmation', EPSA15 Selected Papers, *European Studies in Philosophy of Science*, vol. 5, Berlin: Springer, pp. 83-93.

Votsis, I. (forthcoming) 'Theory Change through a Logical Lens', in M. Martinez (ed.), *From Contradiction to Defectiveness to Pluralism*, Synthese Library.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. (2018). Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5), 726-742.