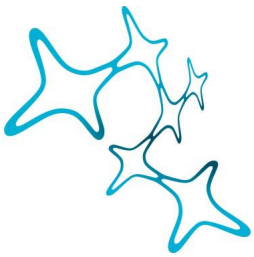
Knowledge Transfer in Cognitive Systems Theory: Models, Computation, and Explanation

Cameron Beebe



Graduate School of
Systemic Neurosciences
LMU Munich

Dissertation at the
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

January, 2020

Supervisor:
Prof. Dr. Stephan Hartmann
Chair of Philosophy of Science
Co-Director of Munich Center for Mathematical Philosophy
Ludwig-Maximilians-Universität München

First Reviewer: Prof. Dr. Stephan Hartmann
Second Reviewer: Prof. Dr. Christian Leibold
External Reviewer: Prof. Dr. Peter Asaro

Date of Submission: January 24, 2020

Date of Defense: January 22, 2021

"*It seems, therefore, that a general theory of systems would be a useful tool providing, on the one hand, models that can be used in, and transferred to, different fields, and safeguarding, on the other hand, from vague analogies which often have marred the progress in these fields.*"

Ludwig von Bertalanffy

Abstract

Cameron Beebe

Knowledge Transfer in Cognitive Systems Theory:
Models, Computation, and Explanation

Knowledge transfer in cognitive systems can be explicated in terms of structure mapping and control. The structure of an effective model enables adaptive control for the system's intended domain of application. Knowledge is transferred by a system when control of a new domain is enabled by mapping the structure of a previously effective model. I advocate for a model-based view of computation which recognizes effective structure mapping at a low level. Artificial neural network systems are furthermore viewed as model-based, where effective models are learned through feedback. Thus, many of the most popular artificial neural network systems are best understood in light of the cybernetic tradition as error-controlled regulators. Knowledge transfer with pre-trained networks (transfer learning) can, when automated like other machine learning methods, be seen as an advancement towards artificial general intelligence. I argue this is convincing because it is akin to automating a general systems methodology of knowledge transfer in scientific reasoning. Analogical reasoning is typical in such a methodology, and some accounts view analogical cognition as the core of cognition which provides adaptive benefits through efficient knowledge transfer. I then discuss one modern example of analogical reasoning in physics, and how an extended Bayesian view might model confirmation given a structural mapping between two systems. In light of my account of knowledge transfer, I finally assess the case of quantum-like models in cognition, and whether the transfer of quantum principles is appropriate. I conclude by throwing my support behind a general systems philosophy of science framework which emphasizes the importance of structure, and which rejects a controversial view of scientific explanation in favor of a view of explanation as enabling control.

Acknowledgements

I would like to express my sincere gratitude to numerous members of the Graduate School of Systemic Neuroscience, the Center for Neurophilosophy and Ethics of Neurosciences, and the Munich Center for Mathematical Philosophy. In particular my supervisor Prof. Dr. Stephan Hartmann has always been extremely patient with me, and supported my goals even if they might have seemed out of reach. I am thankful to Prof. Dr. Christian Leibold for agreeing to help supervise a neurophilosopher, it has pushed me to become more interdisciplinary in my work. Conversations with Prof. Dr. Ulrike Hahn and Prof. Dr. Gregory Wheeler have been extremely helpful in providing direction. I would also like to thank my colleagues Mario Günther, Gašper Štukelj, and Roland Poellinger for many spirited debates and discussions over the years. Of course, this dissertation wouldn't have been at all possible without the toleration and support of my wonderful wife, Anki.

Contents

Chapter 1

Introduction

> Without the criterion of 'control' we have no *objective* reason for rejecting witchcraft, or astrology, or dianetics, or any other system that claims to be 'real knowledge'. Ashby, 2008, p. 4303

Knowledge transfer in cognitive systems can be explicated through the effective mapping of structure from a model, which enables control in a target. I argue this from the point of view of a systems theorist, which is consistent with a mechanistic philosophy of science. However, I have one crucial objection to mechanists like Craver, 2007, who combines a systems theory perspective with what I think is an untenable notion of explanation. It is an objection in good faith, since I believe there to be much more agreement than disagreement overall. However, the notion of explanation which is found in the 'neo-mechanist' picture is unnecessary (or worse, mistaken). Additionally, if neo-mechanists wish to explain knowledge transfer in cognitive systems like I do, then I believe further amendments must be made towards the structural view of mechanisms I offer here.

Craver in particular advocates for an ontological approach to mechanistic explanations in science. This is at odds with what I take to be the received view. The received view of what an explanation is contends that it is some kind of communicable, linguistic-argumentative structure. This structure provides a meaningful connection between some parts (explicit or implicit) of a scientific representation, or multiple scientific representations, and ideally increases an agent's understanding when the explanation is communicated. That is, under the received view, an explanation is a function of some kind of representation for a purpose. If explanations in such a view are *objects*, they are formal or structural objects. They are certainly not the kinds of objects that scientists normally investigate. They are a kind of dependent object, formally related in some way to mechanisms in the world, which are the objects of study.

The causal-mechanical view of ontic explanations offered by Craver—primarily in neuroscientific explanations—seems to identify explanations (especially the good ones) with causal mechanisms in the world. Craver claims to get his ontic notion of explanation from Salmon, making the distinction between explanations in the world (identified with multi-level causal mechanisms)

and explanatory texts (some kind of logico-argumentative representations). For Craver, locating a causal mechanism in the world is identical to locating an explanation. The mechanism produces a behavior or phenomenon, and the corresponding explanation (identified *with* or *as* the mechanism itself) explains the behavior. In a slogan, one might say Craver's position advocates for *Explanation Without Representation*. This is Craver's *ontic* notion of explanation, and he maintains its distinctiveness (and the usefulness of the distinction).

> Salmon's most penetrating insight was to abandon the idea—explicit in the [covering law account of explanation] and Kitcher's [unificatory account]—that explanations are arguments. Instead, he defended an *ontic* view, according to which explanations are objective features of the world. This idea can be brought out by considering an ambiguity in the term, explanation. Sometimes explanations are texts—descriptions, models, or representations of any sort that are used to convey information from one person to another. Explanatory texts are the kinds of things that can be more or less complete and more or less accurate. They are representations. Other times, the term explanation refers to an objective portion of the causal structure of the world, to the set of factors that bring about or sustain a phenomenon (call them objective explanations). [...] Objective explanations are not texts; they are full-bodied things. They are facts, not representations. They are the kinds of things that are discovered and described. There is no question of objective explanations being "right" or "wrong," or "good" or "bad." They just are.
>
> Objective explanations, the causes and mechanisms in the world, are the correct starting point in thinking about the criteria for evaluating explanatory texts in neuroscience. Craver, 2007, p. 27

One might think that these two sorts of explanation—texts and ontic—can live harmoniously in a mechanistic philosophy of science. It is my impression that this is not Craver's intention, nor do I find the ontic notion at all plausible. Rather, his account of mechanisms goes hand-in-hand with the ontic notion of explanations. It is anti-representational, objective, and identified with mechanisms in the world. Unfortunately, I think this is unnecessary and a self-inflicted wound. I advocate for an alternative, which is just to drop the ontic notion and focus on 'explanatory texts' as enabling *control*, which is stated many times over by Craver as the purpose of explanation. I just don't think ontic explanations achieve this purpose if they aren't represented and communicable.

It is worth noting now that Craver's position has furthermore influenced a view that Piccinini, 2015 espouses regarding computational mechanisms and computational explanations, embodied in the slogan: *Computation Without*

Representation. More on this later in chapters 2 and 3. Continuing with Craver for now, he reconstructs a view of Cummins as follows:

> On his view, the *explanandum* is some capacity ψ of a system S. S's ψ-ing is explained by analyzing it into subcapacities $\{\phi_1, \phi_2, \ldots, \phi_n\}$ and showing that ψ is produced through the programmed exercise of the subcapacities. To show that ψ can be produced, in this sense, through the programmed exercise of the subcapacities, one specifies a box-and-arrow diagram showing how the subcomponents work together such that they ψ. For example, the capacity of the neuron (S) to generate action potentials (ψ) would presumably be explained by a box-and-arrow diagram that exhibits the programmed exercise of such capacities as rotating, changing conformation, and diffusing. Craver, 2007, p. 110

Craver continues with some notes on reductive explanation, and then comments that

> [...] the systems tradition rejects the idea that explanations are arguments. All that matters is that the phenomenon is realized by some underlying mechanism. Furthermore, systems explanations are not peripheral to the practice of neuroscience; they are much more accurate descriptions of neuroscientific explanations than the reduction model supplies. Craver, 2007, p. 110

I am not concerned here with whether Craver has accurately reconstructed Cummins's view of systems explanations. Rather, I wish to object to the assertion that the "systems tradition" generally rejects the idea that explanations are arguments. I show that there is an alternative systems theory framework in which explanations are arguments or representations of some kind. This framework is based on the general systems theory (GST) picture of Bertalanffy, 1969 and the closely related cybernetics developed by Ashby, 1958. The methodology these authors outline produces explanations that rely on formal structural properties common across representations in various domains of (complex) systems.

It may be that there are competing schools of thought within the systems tradition itself, but I want to emphasize here that at the very least there is support for an alternative perspective from systems theory—one in which explanations *are* arguments. Oddly, it seems plausible to take from Craver's own reconstruction of Cummins's view that there are systems tradition proponents who think explanations are arguments. The representation of a mechanism, and its analysis into components and a program of organized behavior, seems very much like an argumentative structure is being identified as an explanation.

Showing how something works just *is* an argument, however implicit it may be. We should maintain the distinction between how something works and *showing* how something works.

In general, it is my opinion that Craver conflates many times what are properties of complex systems (and multi-level mechanisms) with properties of explanations. This should be unsurprising given the focus on a notion of explanation that is ontological (and objective). This notion of explanation is incorrect at worst, and at best does not advance a *useful* new kind of explanation since we will always require the explanatory text anyways when trying to explain something to someone. Wright appears to have addressed similar worries (Wright (2015) and Wright (2012)), finding the ontic conception of explanation going back to Salmon to be at best a *misconception*. At the very least, we can state a pragmatic objection to ontic explanations. This consists simply of the observation that *no ontic explanation is an intersubjective explanation*. An ontic explanation is incommunicable, since it is not a representation of the world but identified with an object in the world itself. I cannot *use* an ontic explanation to explain.

While I do find Craver's discussion of normativity for systems explanations productive, I do not see the motivation for abandoning the logico-mathematical foundations that must be present to understand just what an explanation is. It simply cannot be *just* identified with a mechanism, nor just a superfluous reference to a physical mechanism in the world. There are practical functions of explanations, they need to be described or written down or represented in some way in order to achieve these functions. One function is, of course, to be communicable between agents such as to facilitate understanding of why a phenomenon occurs.

Explanations facilitate knowledge, and as I argue in chapter 2, knowledge can be seen in terms of prediction, control, and regulatory capacities. Control is not possible without a representative vehicle to transfer this knowledge from the explainer to the explainee. This of course does not mean that the mechanism must be transparent—and indeed one issue I discuss in chapter 4 in depth is the notion of epistemic opacity in artificial neural networks. It seems hardly productive (or even accurate) to say that the causal mechanisms in a network are an ontic explanation (...of the network).

I think a reasonable way forward is to save the progress Craver made with respect to multi-level causal-mechanistic explanation, and drop anything which does not fit in a general systems theory which also handles the object of this dissertation: knowledge transfer. Knowledge transfer in cognitive systems revolves around the transfer of control capacity in model-based systems. These systems can be understood as having an effective structural mapping between the model and a target operation. The structural view of *systems* as well as systems *theory*, and the methodology of GST and cybernetics, is opposed to both

Craver and Piccinini's slogans. Rejecting representations in systems theory explanations would be just as incompatible with the methodology as rejecting representations (however minimal) in the systems of study. Thus, my approach leaves behind the anti-representationalism evident in Craver's ontic notion of explanation and in Piccinini's notion of computation. I will show that there is sufficient motivation from within systems theory to maintain a notion of representation for explanations as well as a minimal notion for natural and artificial systems of interest.

In the following chapters, I argue that computational systems can be understood as model-based, where a pragmatic definition of physical computation depends on a structural mapping a user imparts to the system. We can likewise understand artificial neural networks as systems which learn an effective mapping (whether it strongly 'represents' or not) which establishes control for intended data distributions. Transfer learning in artificial neural networks means transferring a system's capacity to control to a new distribution. Finally, in high-level model-based cognition like that of analogical reasoning, knowledge is transferred from a source domain—the *model*—to a target domain. This is what we do when we apply models to target problems in scientific reasoning, in cognition, and in machine learning. This dissertation looks at this subject of transfer by examining examples in these contexts.

The examples I discuss are tied together by a loose philosophy of systems, or systems theory, which I refine based on the presented content of this work. I start with what I consider to be conceptually useful foundation found in Ludwig von Bertalanffy's *General System Theory* (Bertalanffy, 1969). Systems theories are theories concerned with systems broadly construed—physical systems, biological systems, neural systems, sociological systems, etc. Dialing into a more refined view, what is important is the formal representations involved in the analysis of systems, and so a systems theory can be understood as a formal methodology employed for study in systems sciences. I will draw attention throughout this dissertation to Ashby's cybernetics (Ashby, 1958), and I find it plausible to rebrand the associated systems framework as *structural* systems theory.

Typical systems we are interested in are those whose organization and relationship between parts are relevant for the behavior of the whole system. Complex systems theory studies systems where the behaviors and interactions of parts become unintuitive and convoluted. We might not be able to easily understand or predict how the system as a whole will behave under any given circumstances. On the other hand, so-called "heaps" are trivial systems whose behavior can be considered summative:

> We may define summativity by saying that a complex can be built up, step by step, by putting together the first separate elements;

> conversely, the characteristics of the complex can be analyzed completely into those of the separate elements. This is true for those complexes which we may call "heaps," such as a heap of bricks or odds and ends, or for mechanical forces acting according to the parallelogram of forces. It does not apply to those systems which were called *Gestalten* in German. Take the most simple example: three electrical conductors have a certain charge which can be measured in each conductor separately. But if they are connected by wires, the charge in each conductor depends on the total constellation, and is different from its charge when insulated. Bertalanffy, 1969, p. 67

Heaps are uninteresting systems, but they help contextualize the kinds of systems we are interested in. Perhaps systems *theory* can be better understood as a system *framework*. It is a framework of thought underlying a methodology concerned with identifying, describing, and understanding the properties and behaviors of *systems*. Particular systems have associated theories, but the method in which systems science is done is best thought of as a framework. Furthermore, general systems theory is a more general method aiming to tie together seemingly disparate systems for potential efficiency, explanation, and unification.

That is, we might define an ideal systems theorist as not limited to one domain of science like atomic physics or molecular biology. The agent might try to transfer models from one domain to another. There are arguably real benefits pedagogically for learning transferable complex systems principles in hands-on methods (like programming models, see Goldstone and Wilensky, 2008 for example). Promoting the learning of generally applicable principles is aided by the study of concrete (although idealized) model *systems*. Scientists and philosophers alike also find models (of various kinds) to be indispensable tools.

A general system theory (like that outlined by Bertalanffy) is about a methodology for managing inter-system models and relations. A particular system theory may have broad application, in which case it might also make sense to discuss it as a general system theory for the applicable classes of systems. Bertalanffy, 1969, p. 84 distinguishes three 'levels' of the description of systems and the phenomena they produce. The first is that of analogy, or what he calls "[. . .] superficial similarities of phenomena which correspond neither in their causal factors nor in their relevant laws." Considerably more important for scientific modeling, according to him, is a second level concerning homologies where "[. . .] the efficient factors [between systems] are different, but the respective laws are formally identical." Importantly, he considers the 'hydraulic' relationship between electrical current and fluid flow—a *classic* example in the philosophy of science—to be a homology, not an analogy. This is a formal distinction based on relations between structures in representations.

The third level of description is that of *explanation*, "[...] the statement of specific conditions and laws that are valid for an individual object or for a class of objects." I am primarily concerned with *cybernetics* which is, following Bertalanffy, a general system theory concerned with certain classes of regulatory systems which effectively control inputs in apparent goal-directed or purpose-driven behavior. That we can describe these systems in such a way is at the very least convenient, if not actually suggestive of teleological mechanisms. This was an important stepping stone towards cybernetic systems theory as argued by Rosenblueth, Wiener, and Bigelow (1943). Cognitive systems like the human nervous system and artificial neural networks are classes of regulatory systems. Bertalanffy continues on the notion of explanation:

> Any scientific explanation necessitates the knowledge of these specific laws as, for example, the laws of chemical equilibrium, of growth of an organism, the development of a population, etc. It is possible that also specific laws present formal correspondence or homologies in the sense discussed; but the structure of individual laws may, of course, be different in the individual cases. Bertalanffy, 1969, p. 85

Just as a systems theorist might regulate the application of a model to a novel domain, a cognitive system might transfer knowledge in a similar way. So what is knowledge transfer? Transfer is generally taken to be a procedure for taking or moving some *thing* from one place to another. Knowledge transfer is the transfer of some known or believed facts from one place to another. Or, it is the transfer of information of some sort relevant for a cognitive system. At a low level, for example in a neural system, this might be the transfer (or application) of an adaptive mechanism to an input which is outside the typical distribution of environmental stimuli. In the case of a computational device which simulates or is a model of another system, the behavior or output of the device itself may represent information gained via an assumed mapping. Loosely, we might characterize the first sense being about controlling novel input, whereas the second is perhaps more about predicting. These might not be exclusive senses of transfer, in fact it seems reasonable to suppose that the transfer of control and predictive capacities go hand in hand.

At a high level, for example analogical reasoning, we might again notice two distinct ways in which model-based knowledge transfer takes place. There is the way in which a new property or behavior in a target domain is asserted to exist, because it exists in the source (model) domain. In contrast, there is reasoning which is done with a 'complete' analogical correspondence (mapping) between the domains. We could say that the first stage is concerned with establishing the relationship between the two domains, building a mapping. The second stage has already granted a mapping, and transfer proceeds along rather deductive rails. These two methods are characteristic of knowledge transfer at both high

and low levels, not just particular to higher level analogical transfer. The first stage is perhaps of greater interest primarily because it appears to involve some sort of *non*-deductive leap or discovery.

For the purposes of this dissertation, knowledge is best thought of in terms of *control*, as Wheeler (2016, p. 322) notes for a machine epistemology in a world of big data: "Knowledge is a means of control, not a special state of mind." Knowledge is demonstrated by the ability to control a system. This idea stems from Ashby and is thoroughly outlined in chapter 2. Knowledge transfer, then, is the transfer of the ability to control a system. Control is fundamental to the field of *cybernetics*, which was founded by Wiener (1948), and significantly developed by Ashby (1958). The Greek root κυβερνητική is also where the term *govern* derives from.

Again, the notion of control goes hand in hand with that of prediction, and so we might alternately think of knowledge as demonstrated by the ability to *predict*. In this case, knowledge transfer could be thought of as the transfer of predictive capacity from source to target. Ashby also considered control and prediction to be forms of *regulation*, and so knowledge is seen as the capacity to regulate. (Ashby, 2008, p. 4438-4439) In this case, we can wrap up the notions of prediction and control and consider knowledge transfer as the transfer of regulatory capacity. Later in chapter 4, I discuss the example of a thermostat as a cybernetic regulator, and how artificial neural networks can be thought of along similar lines.

Once we have a concrete notion of knowledge transfer, it becomes something which is the subject of regulation itself. Much of modern science has become primarily concerned with modeling—the identity of scientists becomes increasingly one of a *modeler*. Philosophers of science have spent plenty of effort on detailing examples of modeling in science, as well as characterizing the general procedure of model-based reasoning which transfers knowledge from source to target domains. One issue these modelers must face is when it is appropriate to apply one model over another, and determine what are 'good' and what are 'bad' models, 'good' analogies and 'bad' analogies. How might a methodology of science attack this problem? Can it prescribe some clear guidelines, or will the modeling process always remain messy?

Well, if we take the literature on analogical reasoning to be any indication, one might get the impression that there is no principled and coherent account that prevails—but our intuition is that *we know analogy when we see it*. Similar to Bartha (2010), we might proceed to try to build an overarching account of analogy by an analysis of seemingly *good* cases of analogical reasoning. We should look to examples of scientific reasoning to find these examples. Even though bad analogies have been made by scientists of the past, scientific reasoning on the whole is more rigorous and *if* we are to find examples of good analogical reasoning, it will probably be by scientists. But how do we evaluate

these examples, which ones are 'good'? Perhaps analogical reasoning is more about efficiency and effectiveness as a methodology than it is about agreeing in which cases an analogy (or the application of a model) is philosophically justified.

As Bartha (2013) notes, there is a difference between philosophical accounts of analogy (based on some normative principles), and computational approaches. I will follow the computational route to analyzing analogical and model-based reasoning, and see how far we can get. What might a computational system look like if it could 'reason' by transferring previously learned knowledge from one domain into another new (and perhaps previously unencountered) domain? I offer an account in chapter 5 based on techniques for transfer learning with artificial neural networks developed in the machine learning community. I argue it contributes to the philosophical discussion on analogical reasoning, providing a formally rigorous characterization of *good* analogical reasoning as *positive* transfer learning (and vice versa). This account is rigorous because there are clear quantities, like accuracy or time, which we can measure to say what we mean by good and bad. Efficiency is a value when survival is the goal, in stark contrast to traditional philosophical analyses.

That is, perhaps we should not look to judge good analogical reasoning by some a priori principle. Instead, we can look to the performance of a protocol that allows model transfer. Transfer is positive when a new task is aided (i.e. takes less time, or more accurately classifies or predicts data), and negative when the transferred structure hinders problem performance. Importantly, in the machine cases I am aware of, the relation between domains is asserted by external means. The data scientist augments the learned ANN and chooses a target problem based on a perceived similarity (by the scientist) or analogy with the previous information. For artificial general intelligence, I argue that these tasks must be automated. This automation constitutes a second feedback loop as discussed in the foundations of cybernetic systems theory. The important take away for philosophers is that artificial general intelligence using transfer learning will require some robust structure mapping engine and a way to manage model-based inferences among many models.

Some form of Bertalanffy's general systems theory (GST) is arguably a ready-made framework for discussing knowledge transfer between various domains. GST is intended to be applied to scientific reasoning, but it also provides a conceptual framework to discuss what an artificial inference environment should look like in which the relations (or informational links) that transfer knowledge (or learned parameters of a more abstract nature) are not externally determined. In loose terms, the idea is to use GST to provide a working framework to outline what must be achieved for an artificial intelligence to *automatically* transfer between appropriate domains—to reason analogically like

humans do. The systems theory perspective will thus tie together this dissertation in an interesting way, since GST was originally developed as a philosophy of science framework yet I argue it contributes, via cybernetics, to the theory of artificial general intelligence.

I take control theory, and cybernetics (at least Ashby's brand), to provide a philosophical underpinning for causal mechanistic explanations for computational and cognitive systems. In particular, I advocate a *structural* view of mechanistic explanation which I think Ashby has provided. My approach is also I think generally consistent with another modern cybernetic view by Seth (2015) as an underpinning of so-called Bayesian brain and predictive processing approaches. A structural approach is warranted for an analysis of knowledge transfer, and is consistent with the influential work of Gentner (1983) on a structural mapping account of analogies.

So, in addition to discussing knowledge transfer specifically, this dissertation in my view provides an imperative to suggest an alternative foundation for a neo-mechanist view in philosophy of science. The currently received neo-mechanist view put forth perhaps most prominently in the present context by Craver (2007) has been rightly criticized by Colombo, Hartmann, and Iersel (2014) as unecessarily presupposing a realist view, as well as holding a controversial view of explanation. This view of explanation identifies a mechanistic explanation with a causal mechanism in the world, distinguishing this from so-called explanatory texts. I find this view to be untenable not just because it is unnecessarily realist, but because explanations require representation. This is a pragmatic response motivated also in part by what I consider to be the spiritual successor to Craver's book by Piccinini (2015) concerned with the metaphysics of computational systems. While Colombo, Hartmann, and Iersel (2014) suggest an anti-realist view, I would like to try to split the difference by proposing a structural mechanist, or structural systems view.

1.1 Overview

Each chapter in this dissertation was written to be as self contained as possible, although I have tried to reduce any redundancies. The order in which they appear tries to weave together the most coherent story about knowledge transfer in cognitive systems theory. I have also added transitions, where appropriate, to chapters that were originally written as stand-alone papers.

Chapter 2 provides important background on systems theory, cybernetics, and an associated view of explanation. I relate these to prominent mechanistic views in recent philosophy of science, noting how their ontic notion of explanation differs. Most crucial to an analysis of cognitive systems is arguably the characterization of a cognitive system as computational.

Chapter 3 outlines a model-based view of computation. A brief analysis of analog computation is presented, taking into account both historical and more modern statements. I show that two very different concepts are tangled together in some of the literature—namely continuous valued computation and analogy machines. I argue that a more general concept, that of *model-based* computation, can help us untangle this misconception. A two-dimensional view of computational devices is offered, in which this model-based dimension is orthogonal to the dimension concerning the type of variables represented by components. The model-based dimension measures the structural relation a device has to a computational problem. I argue that this is a useful framework for assessing alternative computing devices and computational claims in an expanding landscape of computation. Under this view, the structure of a representation of the system is relevant for the use of a computational device by a user.

A model-based view is also natural for artificial neural networks (ANNs), discussed in chapter 4. ANNs can be understood as *cybernetic regulators*, where models are adaptively learned by these systems as they adjust their internal organization via feedback. Ashby's Law of Requisite Variety provides a simple game-theoretic foundation for understanding the power and potential of regulators. This regulatory game, augmented with the Good Regulator Theorem and the crucial concept of feedback, are sufficient to provide the concepts necessary to understand why ANNs are so effective for tasks like classification. Under this reconstruction, justification for belief in the effectiveness of ANNs is the same kind of justification we have for believing that a thermostat will regulate the temperature in a room. I argue this effectively reduces the epistemic opacity of ANN methods, at least for a wide range of interested non-experts. Even so, it isn't clear that epistemic opacity of ANNs is an in-principle worry for experts.

Following this, chapter 5 goes into detail on what is called transfer learning. Transfer learning with pre-trained artificial neural networks is a state-of-the-art technique which may aid in the quest for more general machine intelligence. Data scientists successfully apply neural network models to novel tasks considered to be outside the initial training domain, efficiently transferring their predictive power. It is straightforward to see how robust automation of transfer learning techniques could justify expectations of a notion of artificial general intelligence (AGI). I analyze this potential in an associationist framework, comparing transfer learning to a view of analogical reasoning in humans. This view characterizes analogical reasoning as an inductive *logic of discovery*, which is surprisingly akin to perception, and therefore central to general cognition among diverse domains. I argue this helps justify a philosophical approach for characterizing analogy as effective structure mapping. As a high-level conceptual tool, a general systems framework is suited for analyses of robust inter-domain transfer of models, and therefore is suited to characterize AGI.

In chapter 6, Roland Poellinger and I look at how one might characterize knowledge transfer under a Bayesian philosophy of science. We look at an example of analogical reasoning, and sketch what an extended Bayesian network looks like in order to communicate confirmatory information between model and target systems (specifically, the structural representation of these systems). Analog models can be used to investigate aspects of a target system we might not have easy empirical access to. Evidence from an analog model has been used to argue for the confirmation of a target theory. (Unruh, 2008, Dardashti, Thébault, and Winsberg, 2017) We investigate another example, a water-wave analog system of the quantum Casimir effect, and argue that analogical reasoning in this case cannot be sufficiently expressed by traditional Bayesian networks. Our formalization of the concept of analogy provides a novel reconstruction of Bayesian confirmation from analog models, which preserves the epistemic information between model and target system representations.

As a final case in motivating a structural systems theory for cognitive systems, chapter 7 discusses one of the most controversial trends in cognitive systems modeling. This trend considers the human cognitive system as quantum-like, utilizing the mathematical structure of quantum mechanics to model judgements and decision making. I consider how the mathematical structure is applied, and why it represents an interesting case of improper knowledge transfer. Decoherence in quantum theory explains why an agent should, when possible, have belief states which are classically compatible. Incompatible belief states, used by quantum-like modeling proponents to explain mistakes in human judgement and decision making, illustrate why these models cannot supplant well-justified computational level frameworks like Bayesianism. Such a framework provides a set of normative constraints which are built on classical logic, not quantum logic. The fact that the world is ultimately quantum has little-to-no bearing on the typical distribution of cognitive problems human agents have encountered in our evolutionary past, and are adapted to deal with. Just as a physicist should use classical physics to describe and solve typical macroscopic problems, a human agent should have classical computational goals for operating on beliefs. A regulatory system based on quantum control principles is not effective for survival. In particular, if quantum-like belief states do not decohere, prediction and control capacity of the regulator is compromised.

In the concluding chapter 8, I cement what takeaways I think are justified in the scope of this dissertation. I advocate for a structural systems theory view going forward, a view which I think Ashby made quite explicit. In particular, the emphasis on transformations or transitions of states is, I think, the correct level of structural analysis of systems. I also think the view of communicable explanations as enabling control is a compelling alternative to the ontic view proposed by neo-mechanists. These points were thoroughly brought to the surface by analysis of knowledge transfer in cognitive systems.

Chapter 2

Cybernetic Systems, Representation, and Explanation

> [There] is no such thing as an 'absolute system' in the 'objective' world, i.e. one devoid of an observer or experimenter: there is only a vast mass of *That Which Is*, behaving. [...]
>
> We can also see that the essence of absoluteness is not that it is something that I discovered in the world, as Curie discovered radium, but that it is a way by which information can be got out of a machine, or control exerted over it (the two mean the same). Ashby, 2008, p. 4043

The class of systems of most interest presently is one in which we assume some computational purpose determined by a user. In many cases, for example the embodied human nervous system, the user is presumed to be the system itself. Cybernetics offers us a mechanistic framework to explain the survival or effective use of a system as successful regulation of environmental disturbances. The apparent teleological behavior of these systems is not to be discarded lightly in our descriptions of them, if we follow the influential arguments made by Rosenblueth, Wiener, and Bigelow (1943). The history of cybernetics and adjacent fields in Cordeschi (2002, §4) shows how serious the issue has been, for useful descriptions of systems as well as for explanations of their behavior. This should not be confused with the false claim that evolution occurs in a purposeful direction.

These systems are also *computational* in the sense that we characterize many of the most important operations which occur in the system as computations. Computational operations are formally described by abstract machines, following transition rules from state to state. These computational operations will be a subset of the overall regulatory mechanisms, and are distinguished from non-computational mechanisms in that they are legitimately interpreted as having representations (however minimal) which stand in a modeling relationship to the target disturbances in the environment.

For the kinds of systems I am discussing, I will utilize the general systems framework of *cybernetics*. Cybernetics is a particularly wide framework that

can be classified under the notion of a GST. That is, it will exhibit the sorts of reasoning and methodology of a GST. I take my cue from Bertalanffy:

> Cybernetics, as the theory of control mechanisms in technology and nature and founded on the concepts of information and feedback, is but a part of a general theory of systems; cybernetic systems are a special case, however important, of systems showing self-regulation. (Bertalanffy, 1969, p. 17)

A cybernetic regulator aims at mitigating disturbances from the environment, with respect to a regulatory goal. Ashby, 1958

2.1 Structure Mapping

Structural systems theory (SST) is a minor refinement on the ideas already present in Bertalanffy's GST and Ashby's cybernetics. At the risk of oversimplification, I can offer a slogan and hasty summary of the premises I take to be central to regulatory systems in SST: *A Machine is a Mapping.* There are perhaps historically three ways to cash out this slogan, or three filters to interpret it through, which are unlikely to be exclusive of each other. First, a filter of Shannon's communication theory, where states are put into and come out of a communication channel. A description of the system at this level is a mapping of probabilities and information (entropy). (Shannon, 1948) Second, there is the filter of Turing's abstract notion of computation. The mapping here is characterizing state transitions among abstract computational states. Finally, the filter of dynamical systems analysis, particularly high level descriptions of systems. A mapping would here correspond to transitions between states in phase space. Which filters to use would likely depend on the individual scientist's goals, and the class of systems of interest. Ashby, for example, seemed to prefer the first and third methods of description. (See e.g. Asaro (2011), Pickering, 2010, p. 148-149, and Ashby (1962).) Regardless, I take this to be the first order notion of *structure* involved in SST. Already there is an obvious trouble for neo-mechanists (Piccinini in particular) to bolt onto SST without dropping the anti-representational stance.

As mentioned, it will be important for structural systems theory to distinguish between superficial analogies in the function of a system, and mechanisms which are structurally similar or *homologous*. This aspect of the methodology operates at a second order level: mapping a structure (which is itself perhaps best characterized as a mapping). An influential account of analogical reasoning in cognitive science is that of Gentner (1983), where it is characterized as structural mapping. The structure of a representation (of a model or a problem) in a source domain is mapped onto a target domain. Hesse (1966) distinguishes

further between lateral relations in analogical reasoning, based on perceived similarities between representations, and vertical relations that are based on causal relationships between the entities represented. These analyses provide an account of the of scientific reasoning that is relevant for example in the case study of chapter 6. Scientists perceive some similarity in tasks, and apply the structure that they think is a good analysis of a source task and map that structure onto the new target problem. If successful, there will be a structural relationship between the models used in the different domains.

Homologies are like analogies, but where perceived similarities are actually structurally similar and not just superficially similar. The distinction between an analogy and homology may not always be clear cut, but we should be aware of the spectrum of similarity relations, and how we might characterize the relationships between classes of systems. A popular example is the case of wings in both bats and birds. Is this an analogy or a homology? Bat wings and bird wings are functionally similar, in that they enable the flight of each animal through the air, but the similarity between them is arguably superficial. If we tried to learn about the individual mechanisms of one organism (feathers creating a surface for lift) by 'applying' the analogy or what we know of the other (skin membrane forming a patagium), perhaps we would consider the endeavor relatively unproductive aside from characterizing the basic aerodynamics. The individual parts and mechanisms (feathers, skin) do not share causal factors or relevant laws to achieve flight. Or, if they do, they might be 'superficial' aerodynamic laws of lift.

In that case, what we learn might be applicable to create airplanes, whose wings are even more superficially similar to the wings of bats and birds than bat wings are compared to bird wings. If a scientist, engineer, or philosopher is trying to gain knowledge, control, or practical insight about aerodynamics, then perhaps the general aspects of surface area, lift, and weight which bat wings and bird wings both satisfy could be construed as homologous. This would only be because, for the particular use case, the level of description found in the representations of the systems abstracts away from the irrelevant (for this use case) aspects of the lower level mechanisms (cells, bones, etc.). But this seems to me to be what Bertalanffy is trying to define away: a homology should present relevant similarities at a deeper structural level. Otherwise it is just an analogy if the similarity of representations between systems abstracts away from what would be relevant causal factors or laws on a deeper analysis.

If the objective is to learn about the components and mechanisms of the wings and the winged organism, the fact that both organisms can achieve aerodynamic lift with sufficient surface area is a functional similarity which is superficial. It is just a background assumption which determines a base category for filtering the investigator's attention. The kinds of mechanisms which are present in the feathers or skin membrane are relatively irrelevant to the

aerospace engineer's representation, but might be essential to the comparative biologist. In this case, there are higher standards for the establishment of a homology among mechanisms which enable a functional wing in the organism. The organization, development, or behavior of parts in the respective organisms would need to follow formally identical laws. Even if the parts themselves are different in important material respects, the point is that there are deeper structural similarities in a homology compared to an analogy.

A GST methodology should guard against superficial similarities between domains, while also finding meaningful common models applicable to a wide class of systems because of some shared structural properties. It makes sense that the methodology is formulated in structural terms, along the structuralist lines discussed by Ashby (1958, §6) for cybernetic systems theory. The distinction between analogy and homology, for Bertalanffy, stems from discussions in biology and zoology:

> Analogies are scientifically worthless. Homologies, in contrast, often present valuable models, and therefore are widely applied in physics. Similarly, general system theory can serve as a regulatory device to distinguish analogies and homologies, meaningless similarities and meaningful transfer of models. [...]
>
> The homology of system characteristics does not imply reduction of one realm to another and lower one. But neither is it mere metaphor or analogy; rather, it is a formal correspondence founded in reality inasmuch as it can be considered as constituted of "systems" of whatever kind. (Bertalanffy, 1969, p. 85)

We could say that the difference between analogy and homology is understood along structural lines. An analogous organ in an organism has a similar function to another organ (in another organism) but is unrelated in a structural or causal way. In this case, a shared structure could be a shared evolutionary path. One might follow the lines of (Jardine, 1967) in identifying the structure with representations of mechanisms among parts in complex systems. An evolutionary homologous organ has more structural similarities to another organism's organ due to shared evolutionary ancestors. Mechanistically, a homology between systems is when the relevant parts are causally effective for a similar function.

2.2 From Cause to Control

As a recap of the neo-mechanist picture, Craver (2007) puts forward a causal-mechanistic account of multi-level causal explanation in neuroscience. Multi-level causal mechanisms are constrained by parts and behaviors in each level,

and our explanations are mosaics of these mechanisms. Importantly, explanations for Craver are *ontic* in the sense that they are identified with the causal mechanism in the world and are *not* represented in any way formally or linguistically. Explanations represented so are deemed by Craver to be explanatory *texts* and not of primary concern to explanatory analyses for systems in neuroscience. Similarly, Piccinini (2015) argues for a notion of *computation without representation*. I believe these anti-representational accounts significantly clash with the historical foundations of systems theory, computation, and cybernetics.

That said, one reason I do not want to completely disregard Craver and Piccinini is I believe they have made some significant and detailed philosophical progress in the direction of systems theory explanations. I do not want to throw out the baby with the bathwater. I would rather part out the pieces of their accounts I do not believe fit to the overall picture of GST, and replace them with more appropriate notions. There are however significant areas of overlap to be found in the foundational literature, in my opinion, which Piccinini, Craver and others may not be aware of. Particularly important is the account of causation relied upon by Craver (and subsequently Piccinini). This is the account of causation referred to as *interventionist* or *manipulationist*. See Woodward (2016) for an overview.

Arguably, Ashby (1958, p. 55) has already provided an interventionist picture couched in cybernetic and systems theoretic terms:

> Suppose we are testing whether part or variable i has an immediate effect on part or variable j. Roughly, we let the system show its behaviour, and we notice whether the behaviour of part j is changed when part i's value is changed. If part j's behaviour is just the same, whatever i's value, then we say, in general, that i has no effect on j.
>
> To be more precise, we pick on some one state S (of the whole system) first. With i at some value we notice the transition that occurs in part j (ignoring those of other variables). We compare this transition with those that occur when states S_1, S_2, etc.—other than S—are used, in which S_1, S_2, etc. differ from S *only in the value of the i-th component.* If S_1, S_2, etc., give the same transition in part j as S, then we say that i *has no immediate effect on* j, and vice versa. Ashby, 1958, p. 55

Ashby develops and utilizes this interventionist or manipulability notion without much reference to causes at all, but does continue to use the notion of a change in variable yielding an immediate effect. Causal arrow diagrams are instead "diagrams of immediate effects". What is important is again the *control* of the system.

This is not to say that this is historically the first such use of interventionist ideas. Rather, I want to draw the attention of the neo-mechanists to such

statements since cybernetic systems theory is (at least according to Bertalanffy) a paradigm example of systems theory. It is therefore significant in the history of systems thinking, and in particular for computational and cognitive systems theory.

The boundaries of what constitute a particular system must be determined by some User and a representation of the relevant factors for control. Absent this representation, I do not see how the methodology of systems science can operate. Thus, at least for mechanists, it can provide no *explanations* without identifying those parts of a mechanism that can be controlled to produce the phenomena in question.

> In the concepts of cybernetics, a system's "largeness" must refer to the number of *distinctions* made: either to the number of states available or, if its states are defined by a vector, to the number of components in the vector (i.e. to the number of its variables or of its degrees of freedom). The two measures are correlated, for if other things are equal, the addition of extra variables will make possible extra states. A system may also be made larger from our functional point of view if, the number of variables being fixed, each is measured more precisely, so as to make it show more distinguishable states. We shall not, however, be much interested in any exact measure of largeness on some particular definition; rather we shall refer to a relation between the system and some definite, given, observer who is going to try to study or control it. Ashby, 1958, p. 61

I think this is still in good enough agreement with the neo-mechanist picture of Craver, except for the notion that representation is not a part of explanation. Representation is also important just to get a grip on just what the system consists of (i.e. what a point in phase space represents). Furthermore, it will have a bearing on what is considered *typical* environmental input to the system. For example, if we are studying a system, and want to know whether the system will be *stable*, we need to define the class of what we expect the system to be capable of regulating.

> A system can be said to be in stable equilibrium only if some sufficiently definite set of displacements D is specified. If the specification is explicit, then D is fully defined. Often D is not given explicitly but is understood; thus if a radio circuit is said to be "stable", one understands that D means any of the commonly occurring voltage fluctuations, but it would usually be understood to exclude the stroke of lightning. Often the system is understood to be stable provided the disturbance lies within a certain range. What is important here is that in unusual cases, in biological systems for instance, precise specification of the disturbances D, and of the state

> of equilibrium under discussion a, may be necessary if the discussion is to have exactness. Ashby, 1958, p. 79

I argue that this is correct, but clearly at odds with the recent attempts to do without representation in mechanistic explanations. Throwing out representations is a fatal error: we need them to specify the phase space of part behaviors in the causal mechanism. The phase space changes depending on what we represent. Foregoing a representation in a mechanistic explanation is not something that can just be done casually on metaphysical grounds, since the explanation depends on the representation of what is causally relevant (and therefore relevant for control) in the mechanism. It cannot be any other way, representation is *necessary* for explanation. Or, at least it is necessary for *useful* explanations.

Explanation is done with *purpose*, just like computation. This purpose, to be achieved, must be formally *specified*.

2.2.1 Computation with Representation

A spiritual successor to Craver's project is Piccinini's book *Physical Computation*. (Piccinini, 2015) Comparing slogans again: if Craver says explanation does not require representation, then Piccinini says computation does not require representation. The idea is that there is some way to metaphysically distinguish computation (and computational systems), by pointing to causal mechanisms. An earlier paper by Trenholme (1994) also attempts to dispense with representations, in what seems to me to be a similar metaphysical attempt to define "naturalistic" analog simulation. I disagree with Piccinini and Trenholme for similar reasons to why I have dismissed Craver's ontic notion of explanation without representation.

The kinds of systems we are concerned with have to represent, at least in a minimal pragmatic sense. Otherwise we run into difficulty just trying to convince ourselves that a computational system is computational. We have to presume that the system has an aim, falling in line with Rosenblueth, Wiener, and Bigelow (1943), and that the mechanisms which it uses to achieve this aim have a non-random relationship with the challenges and stimuli it faces while trying to achieve this aim. Presuming the aim itself already presupposes some way of specifying this aim, for if we could not do this then we cannot communicate amongst ourselves that the system is acting so as to fulfill the aim. In other words, the system might be computing but if we cannot specify *what* it is trying to do and *why* it is trying to do it, then the fact that it is computing is irrelevant and uninformative. If this is the metaphysics of computation, then it is wholly uninteresting. Similarly, if there exists some explanation for a novel phenomenon but nobody can *explain* it, then for all practical purposes *there is no explanation existing*. The important point is that there is a user-relative

functional role to computation and explanation. The pragmatic view gets you everything you need, unless you want to do metaphysics in the realm of ideas.

We need representations, otherwise we would look at a computational device like the slide rule and cannot explain why it implements multiplication. It is designed to represent or encode a specific set of transitions corresponding to the mathematical relationship of log scales which, when used properly, can be used by an agent to compute multiplication. The distance slid corresponds to 'addition', and the log scales allow the user to read off the functional output of 'multiplication'. I slide it one unit to the right and it multiplies a number by two. The truth of this statement is vacuous, however, if there is no representation or model of what multiplication *is* (in this case it is related via logarithmic rules). I cannot determine whether the slide rule actually computes multiplication if I do not specify what multiplication is and how (a description of) the slide rule can realize or implement the given model.

The network of knowledge and inferences enabled by a non-representational account is not productive. We, as agents, are not able to "plug into" the network. It is running parallel to the scientific method and the practice of scientists. If no information (literal bits) can be extracted from the non-represented computation or explanation, we cannot construct a model or know where to intervene in a causal mechanism. If we cannot intervene to demonstrate control, it makes no practical difference whether the physical computation exists. I find similar issues with the case of Trenholme (1994), who rejects representations and asserts that there are natural or physical isomorphisms that hold between causal structures of an analog simulation and the target. I do not see the scientific use provided by positing metaphysical isomorphisms between systems, if they are not somehow grounded in a correspondence with at least some practically realized isomorphism among representations. It seems to me to be an unnecessary journey into Platonism. I will return to Trenholme in the next chapter, but want to note here that, like Piccinini, I still do find a lot to agree with.

Rather than dispensing with representations at all, we might just try to have a watered down or less loaded notion of representation. This is obviously a very controversial area of study, and an area which no doubt my account of SST would benefit from further work on. I offer only a brief comment at this point. Perhaps for SST, we can get by with using the word representation (in the kinds of systems we are concerned with) to refer to something like *an effective encoding of an effective mapping.*

2.3 Knowledge as Control

The conception of knowledge most relevant to the topic of this dissertation is that of control. Knowledge is nothing more than the ability to control a system.

While there might be some other sense of knowledge which doesn't need to exhibit control, it may be of little use for scientists. Demonstrating control is a part of experimentation, explanation, and a scientific methodology. I outline this view by citing specific passages in W. Ross Ashby's journal compilation, where he makes it explicitly clear. As I find these passages to be not only informative on the subject of the dissertation as well as historically interesting, I quote these passages at length here, and occasionally throughout.[1]

> 'Knowing' a system means, ultimately, being able to control it. This means that the 'knower' $[K]$ has within his brains the organization that will convert an actual state S_i (of the system), given to K via his sensory receptors, into that set of parameter values α as will lead to system S going to an assigned state S_j. Ashby, 2008, p. 4292

Ashby then discusses roughly three aspects of the system to focus on solving for control. We need to know the state desired, the current state of the system, as well as what to set any internal parameters to. Ashby continues:

> K thus becomes a transducer with two inputs, one of which, the 'goal', can be taken for granted. Then 'state desired' being given, $[K]$ codes correctly, if he 'knows', all the S_i's into corresponding α's. Ashby, 2008, p. 4292

To verify or test knowledge of a system, we can disturb or manipulate the system by 'kicking' it into some state S_k (instead of S_i) and observe subsequent responses. Assuming the overall system S has some means of adjusting itself to achieve the goal state S_j, it will be able to demonstrate knowledge by demonstrating control or mitigation of the disturbing kick.

> K will promptly re-code $[S_k]$ to a new value of α, which brings about the transition $S_k \rightarrow S_j$. And if K knows *all* about the system S, the whole [system $S + K$] will bring S_k to S_j whatever the kicks do. Ashby, 2008, p. 4293

Control can be thought of in a concrete sense as managing effective state transitions.[2] Knowledge is then understood as a mapping of state transitions which enable control.

[1] Ashby's digital journal collection is extensively cross-referenced by keywords. It could be that some of these quotes are found (their meaning or word-for-word) in Ashby's other writings, but as the digital journal collection is extensively linked it is arguably easier to work with.

[2] As an aside, this is relevant also for subsequent discussions in this dissertation regarding the nature of computational devices. The theory of abstract machines, by Turing and others, is precisely about specifying a transition of a system under any situation.

> If there is a parameter P that can inform K of which state of S is to be the resting state, and if K, given S_i and S_j, can convert S_i to that α as will make S_i pass over to S_j, then K can be said to know S completely. Ashby, 2008, p. 4293

An essential point for systems of the kind we are concerned with here, and central to cybernetics, is that the whole system ($S + K = \Sigma$) is equipped with feedback. Information from the output is allowed to flow back in. The idea is that such a system can learn to 'know' a state transition mapping which is effective for control. An ineffective mapping displaying 'ignorance' can, through feedback, become effective.

> K's 'getting to know' will then correspond to 'changing K's organization until all the fields [of Σ] have the desired property'.
>
> This implies that under the drive of the feedback, K cannot stop until it 'knows' S. (And it implies that 'difficulty of getting to know' is not merely equal to but identical with 'difficulty of getting stable'.)
>
> This seems to settle the 'epistemological' question pretty thoroughly.
>
> Notice that this method regards 'control' as the basic form, or test, of knowledge. Ashby, 2008, p. 4293

Ashby continues in a later journal entry, outlining what I argue is a clear account of knowledge transfer:

> For if knowledge is control, and if K knows how to control S, K has the 'correct' code for turning information about S's state S_i into the appropriate action α, the 'goal' being given. If K is to pass this knowledge on to another scientist K' he can pass on nothing but this coding. He must [therefore] pass on a substitution or transformation; thus, the goal being given he passes on [a] transformation.
>
> This seems to me to be more realistic and fundamental than Eddington's 'all communicable knowledge is knowledge of group structure'. Clearly, groups will soon enter, but they do not come in primarily. Ashby, 2008, p. 4311

Ashby characterizes such transformations here from states to actions, $S_i \rightarrow \alpha_j$. In a footnote, he also assumes "[...] that 'scientific' knowledge is communicable knowledge." That scientific knowledge is communicable means, for Ashby, that the principles of communication theory apply. He summarizes the entry by stating (emphasis mine): *Scientific knowledge is knowledge of a transformation.* This means it is knowledge of a relation between states of a system in the world, which can be used to predict and control. Scientists make predictions, and test them by demonstrating control.

Whether this notion of knowledge as control provides us with a definitive philosophical account is, of course, up for debate. It is definitely interesting and useful for the purposes of this present work, as well as historically. In this author's opinion, it is a significant account that should be explored further. It could be, for example, that a control theory account of knowledge is necessary, but not sufficient, for a philosophically satisfying account of scientific knowledge or understanding. I proceed under the assumption that what I have shown so far warrants further exploration and the comparison I have made with other relevant philosophical accounts.

2.3.1 The Foundation for Structural Mechanists

> We start by finding a curious object, a so-called 'brain', and we apply scientific methods to find out how it works. When we have succeeded in laying bare the mechanism and the principles of its working we find, as we are bound to find, that these 'objectively discovered' principles are intimately related to the scientific methods that *we* have used for its study. Ashby, 2008, p. 4306

Ashby was influenced by Eddington's philosophy of science, including his essay on *The Concept of Structure*, and the idea of scientific knowledge as structural knowledge (e.g. group structures). See for example the dated sections of his journal in Ashby (2008, p. 0345-0351,0371-0377). Another crucial development for Ashby in laying out a structural picture of systems was the mathematical writings of the Bourbaki group, as referenced in his chapter *The Black Box* in Ashby (1958, §6). Likewise, Claude Shannon's theory of communication (Shannon, 1948) is frequently cited throughout Ashby's *An Introduction to Cybernetics*. The role of Shannon's communication (and information) theory is not just used to outline the interactions between a regulatory system and the environment it aims to control for survival, but also in characterizing the organization and structure of the system itself. For example, the design of an effective artificial system will encode some structure which, in information theoretic terms, will be non-random. This is naturally consistent with a notion of model-based computation, or of a system which learns a model through processing ordered data.

It is also worth noting at this point that Ashby had also read Erwin Schrödinger's *What is Life?* (Ashby, 2008, p. 1910-1916) which includes the physical conception of biological systems in terms of open systems which locally reduce entropy. Ashby interprets the work as explaining "how an organism lives, metabolically, by extracting orderliness from the environment, putting back the disorder, and keeping some of the orderliness for itself.". It is not an exaggeration to say that this view interprets living systems as effectively 'metabolizing'

information, as absorbing structure. If the system is not successfully digesting order, its survival is compromised.

Under this view, the study of such systems will thus need to express how this structure is absorbed. This will be in terms of the study and discovery of mechanisms, and corresponding mechanistic explanations. Like Eddington suggested, the study and discovery of mechanisms will be the discovery of some robust structural relations that hold in experiments among entities in the world. Explanations need to be communicated to other scientists, and this explanation consists of descriptions and expressions of the structural relations which appear to hold. Furthermore, and this is the most crucial point, these explanations and the structural relations they encode must enable control over the studied systems. Such a view has the potential to save what progress has been made by the neo-mechanists, at least for cognitive and computational systems of interest.

To be clear, I am not arguing about the particulars of structuralism in the philosophy of mathematics. See for example Reck and Schiemer (2019) for an overview of debates on the topic. My view, in agreement with what I will present from Ashby on the topic, is related to structuralism in mathematics in the sense that structure in systems must be represented somehow. What is more important than sets or groups is the organization and information—the *structure*—of a system. Whether computational, biological, or physical, *how* we represent the system is less important than *that* the system must be represented.

It is my view that at least some structuralists of one variety or another can find agreement with the picture suggested here. I also want to be clear from the outset that I do not claim to be the originator of the structuralist ideas I think underpin the ideas discussed. Rather, the ideas were already present in the so-called systems tradition. If anything, I am just providing a synthesis of ideas which I think are harmonious, and noticing that a structural systems theory (a structural theory of mechanisms) seems to be warranted for study of knowledge transfer in cognitive systems.

2.4 The Taxonomy Problem

The taxonomy problem for computational systems is the problem of distinguishing true computational systems from non-computational ones. We are pretty sure that calculators are systems which perform computations, and not rocks. How do we do this? Piccinini (2015) has a very good discussion on the history of this problem in the foundations of computation, the issues with previous attempts at solving the problem. In his take he ultimately appeals to a mechanistic account of computation (similar of course to Craver (2007)) for individuating computational systems by their functional properties. The details of this account are plenty, and cannot be gone into here.

In my opinion the details largely distract from a much simpler pragmatic solution of the taxonomy problem—which we are liable to accept if we do not wish to follow Piccinini's slogan of computation without representation. That is, if one subscribes to the slogan then *of course* one has a lot of explaining to do on how exactly some physical system computes and another doesn't. He is, in a sense, looking for a way to *metaphysically* distinguish between kinds of systems. On the other hand, if we are allowed representations, we can simply rule out rocks and other trivial systems as non-computational because they are not *used* as computational systems.

Any practical taxonomy relies on our respective representation of the system at hand, its boundaries, the model of computation supposedly being implemented, and the kinds of values capable of being usefully represented in the system. Our taxonomy of computational systems will be influenced by our solution to the boundary conditions. That is, we must first represent what parts are relevant parts of the system, and what parts are not (or are superfluous). For example, a calculator with and without dirt on it represents the same computational system. It is composed of the plastic and silicon elements, as well as a liquid crystal display, etc.

My argument is that it is *also* relevant how we represent the boundaries of the system and what the system is attempting to do. We cannot just point to the physical system, for obvious reasons that it is ambiguous what is being pointed to without some assumptions. A basic level of description, like a phase space where each point represents the state of each component of the system, and the space represents all possible states of all components, crucially depends on where the system stops and the environment starts. The 'act' of defining the boundaries of the system is an act of representation. It changes what a point in phase space represents or indexes. Determining the boundaries of a computational system is an act of representation in the sense that the state transitions in phase space look different for different boundaries. Once we accept that representation is unavoidable in defining what are (and are not) computational systems, we might look to what systems theory could say towards these issues.

A similar taxonomy problem arises in the foundational literature on biological systems theory, where we wish to distinguish genuine living organisms from non-living organisms. Bertalanffy (1969) identifies the core difference with living biological organisms being characterized thermodynamically as open systems. This is in opposition to closed systems as the traditional object of thermodynamics. Closed systems are required by the second law to tend towards equilibrium—there is no import or export from the system that can produce a steady state at a distance from equilibrium. An open system can allow (or import) material from the environment, such that the system nets negative entropy or a positive amount of free energy. The open system can 'spend' this excess energy to do work, and maintain a state at a distance from equilibrium.

We can see this particular taxonomy problem as being solved by a *representational choice*. This representation is not about the particulars of the internal mechanisms and components of the system, but about where we draw the boundaries between the system and the rest of the world.

Computational systems are cybernetic systems in the sense that they are regulators (which may be static or unlearning). At a fundamental level, they are controlling the flow of information from inputs and squeezing the possible transitions into a well-defined output subset of states of the world which would otherwise be physically possible. However, not every cybernetic system is a computational system. I think this is one of the critical missing pieces to previous analyses which encounter taxonomy problems in the foundations of computation, e.g. Piccinini (2015). For example, there has been a traditional worry about accounts of computation trivializing the behavior of some systems like the brain since under some accounts it seems like any system can be considered computational. These pan-computational worries can be avoided by focusing on the use of a given system.

Additionally, now, we can consider cybernetics as adding further refinement to the taxonomy of natural systems. Rocks are not only non-computational due to lack of computational use, they are poor regulators. The fact that they are poor at regulating disturbances from the environment, for example temperature disturbances, means that the rock is not good at maintaining itself at a steady state at a distance from equilibrium. Temperature information flowing from the environment into the rock will not be controlled, just delayed by a material factor. For it to be a cybernetic system, it must be able to regulate typical environmental disturbances well, meaning it reduces the entropy (information) one would gain by observing the rock's state after a random disturbance. We might say that any *typical* disturbance in the rock's existence knocks atoms and molecules out of position, and they will never return. The rock as a system will only monotonically tend towards equilibrium, towards sand. It will also not heat up internally of its own accord, it will not "flip its own bits". At best, the system would be a rather useless model of a rock, or a thermal delayer in an expanded system with the rock as a boundary which slows down the transfer of heat information into another system. A magical thermostat rock system, which had the ability to maintain a constant temperature of 20°C no matter the thermal disturbances, would be a regulator precisely because we would know with certainty after a random disturbance that it would be the same temperature. The useful information gained about the state of the system is zero, compared to a random coin toss (where we would learn 1 bit). More on the information theoretic properties of regulators later.

Additionally, the ordinary rock system does not have an internal model (or an effective encoding of an effective mapping, which as I mentioned earlier

we might just call a representation) of the environment. Or, we are unjustified in concluding that if it *does*, that it enables any successful regulation of environmental disturbances since the rock system monotonically approaches equilibrium. The rock clearly fails the criteria given in what is called the *good regulator theorem*. (Conant and Ashby, 1970) While stomachs are regulatory systems[3], we still might wonder why they do not qualify as good regulators under the theorem, and why we think they are not computational systems. More on the good regulator theorem in chapter 4.

Computational systems are also open systems. They *must* consume, or import, or be provided with, energy from an external source. Not only this, but as noted in Landauer (1961) there will be a non-zero thermodynamic transport of energy (heat) imparted to the environment. This is important to distinguish a computational system from, for example, crystals, plants, and other systems which are in a state at a distance from equilibrium yet we do not think they are computational systems. Crystals have organization, but they do not consume energy by themselves and they do not have a thermodynamic cost associated with certain tasks.

This brings us to an important problem discussed in the philosophical literature for decades: do systems such as crystals or waterfalls (or even rocks) compute a class of functions (or perhaps any arbitrary function)? Well, is the rock (by itself) an open system? No. When it is supposedly implementing some function, does it heat up? No. Then, it is not a computational system. But, lets say that we observe a rock heating up. Is this evidence of computations taking place? Perhaps.

Then we can check our other condition. Is the rock an open system? It is not. If we put the rock in a controlled environment, it will not by itself heat up. In fact, it will remain or approach equilibrium with its environment. The other way around, if the system under consideration is not *just* a rock, but involves transport of energy in an open system, then perhaps it is performing computations. If, when we suppose computations might be taking place, the system also heated up—then we would be in a position to potentially classify the system as a computational system.

We see, though, that this crucially depends on aspects of representation. An open system might be closed under a different description, under different boundary conditions. This is arguably a positive outcome of the two conditions account of computational systems offered here. It should be easy to rule out certain systems, and if these conditions are fulfilled then further study of the system is warranted. These are necessary conditions, not sufficient. It shouldn't be easy to decide, given the fulfillment of these conditions, whether computation is taking place. We only have that the system *might* be a computational system, given a suitable interpretation.

[3]See also the discussion in Shagrir (2010).

Chapter 3

Model-Based Computation

> Thus *the act of "designing" or "making" a machine is essentially an act of communication from Maker to Made,* and the principles of communication theory apply to it. Ashby, 1958, p. 253

This chapter attempts to outline a remedy to what I view as a confusion in the conceptual framework used to characterize computational devices.[1] This confusion is present in some influential literature concerning analog computation, where analog devices are taken to be synonymous with devices which compute using continuous valued variables. Through an analysis of this confusion and the wider computational landscape, I hope to contribute to our understanding of some recent claims by introducing what I call *model-based computation*. Model-based computation can be seen as a distinct 'dimension' with which to evaluate devices in a wider computational landscape, and it allows us to see the flaws in the confused argumentation present in the literature cited. This dimension can be considered orthogonal to the variable types (e.g. binary valued or continuous) represented by components in a device. Furthermore, I argue that this two-dimensional view is a natural extension for current notions of computation, and is well-motivated from the analysis provided.

A first step in this project is to provide evidence that there is a conceptual confusion present in discussions of analog computation. This helps establish what analog computation is *not*, and motivate the discussion in subsequent sections of what it *is*—and how a more general two-dimensional notion of computation accommodates it. We begin with two statements from Nielsen and Chuang's textbook of quantum information theory, which I quote at length for the unfamiliar reader:

> In the years since Turing, many different teams of researchers have noticed that certain types of analog computers can efficiently solve problems believed to have no efficient solution on a Turing machine.

[1] An earlier version of this paper originally appeared in Beebe, 2016, and was extended in *Natural Computing* as Beebe, 2018. The present chapter is reproduced with permission, and is extended and improved by incorporating several new references, updating definitions and clarifying the argument structure, and by providing additional discussion of important points.

> At first glance these analog computers appear to violate the strong form of the Church-Turing thesis. Unfortunately for analog computation, it turns out that when realistic assumptions about the presence of noise in analog computers are made, their power disappears in all known instances; they cannot efficiently solve problems which are not efficiently solvable on a Turing machine. This lesson — that the effects of realistic noise must be taken into account in evaluating the efficiency of a computational model — was one of the great early challenges of quantum computation and quantum information, a challenge successfully met by the development of a theory of quantum error-correcting codes and fault-tolerant quantum computation. Thus, unlike analog computation, quantum computation can in principle tolerate a finite amount of noise and still retain its computational advantages. (Nielsen and Chuang, 2010, p. 5)

> One might suspect that quantum computers are just analog computers, because of the use of continuous parameters in describing qubit states; however, it turns out that the effects of noise on a quantum computer can effectively be digitized. (Nielsen and Chuang, 2010, p. 164)

There seems to be an assumed notion of analog computation as a delicate and noise-intolerant business. Even discussions like those from Turing (1950, §5) concerning the sensitivity of physics to variations in initial conditions might be taken as supporting such a notion.[2] A careful reading reveals that this statement can only be understood in support of the noise tolerance of discrete state machines, and not a statement claiming that any analog computation is noise intolerant. It might be possible to infer something about noise intolerance in some continuous variable computational devices, but that is a separate question (and there might be practical ways of avoiding the problem we do not want to close ourselves off to).

Do these statements about sensitivity to noise actually have relevance for evaluating analog computation generally? I argue that an affirmative answer stems from the core misconception that continuous valued 'organs' (that is, components performing specialized functions) are not only essential to analog computation (they are not) but that a device which has such organs is *synonymous* with analog computation.[3] This is simply not the case. Continuous values are neither necessary nor sufficient to characterize analog computation.

[2] Indeed, Ashby and Turing diverged in the way they sought to build models of the brain or intelligence. Ashby preferred cybernetic feedback models, whereas Turing preferred simulations on general purpose symbolic machines. See for example the discussion in Asaro (2011).

[3] We will see shortly that von Neumann, among others, used the term 'organ' for computational components.

The reader may also wish to see argument 6 from Scott Aaronson's page on skeptics of quantum computation, where he claims "We know that analog computers are not that reliable, and can go haywire because of small errors." Aaronson proceeds to respond to the question of "why a quantum computer should be any different, since you have these amplitudes which are continuously varying quantities." In his response, he makes the very conflation at question here, namely that analog computation is synonymous with continuous value computation.[4]

Additionally, there is not just *one* kind of physical system that can constitute an analog device, meaning that such statements of an analog computer "going haywire" could only be safely interpreted on this level as referring to a particular architecture—but clearly it is meant as a general statement about analog computation. We could take these statements to refer to some formal generalization of the concept of analog computation, but we will see that such conclusions about the weakness (or effectiveness) of an analog computer must be evaluated with respect to a particular *model*.

These claims against analog computation are not supported by the analysis of analog computation offered in section 3.1, where I introduce what I think is a much clearer conception of analog computation. I then proceed in section 3.2 to outline some thoughts on what I call 'model-based computation' respecting this conception. Afterwards, I evaluate two discussions in light of this notion of model-based computation. The first includes computational claims about the brain and the current notion of hierarchical generative models in cognitive science. We will see in section 3.3 that hierarchical generative models seem to describe model-based computation as outlined in this present work. The second discussion in section 3.4 will focus on analog models in physics, in particular the notion of analog simulation recently put forth in Dardashti, Thébault, and Winsberg (2017).

Relevant aspects of the notion of model-based reasoning will be discussed in section 3.5, where we see the relationship with ideas already present in cognitive science. This relationship is arguably unsurprising, given that our notion of computation has historically been informed by the inference capabilities of an intelligent agent. A computer was, after all, initially a term used to refer to a particular desk job for a human. In the concluding section 3.6 I will draw attention to what I think is the importance of this discussion for the developing market place of alternative computing, and summarize the view of a two-dimensional landscape of computation.

[4] http://www.scottaaronson.com/democritus/lec14.html

3.1 What is Analog Computation?

Rather than being defined by the continuity of parameters, an analog computer can be defined through a literal treatment of the word *analogy*—and in fact Neumann (1963, p. 293) among others even refers to two classes of computing machines, analogy and digital machines. The term *analogy machines* sounds much different to our modern ears than analog computer, but I argue it more accurately represents this area in the landscape of computation. Particularly in modern computer science where alternative or specialized computing devices have become more common, it is arguably worthwhile to have a clearer conceptual overview of this landscape.

I argue that by clearing up the issues mentioned above, we can see two dimensions characterizing the landscape. The first dimension has to do with the types of variables processed by computational components (or organs). The other is the extent to which the structure of a device *models* a target problem. The first dimension is arguably uncontroversial, and I do not focus on justifying its incorporation in the two-dimensional view. The second dimension about models warrants significant discussion, and forms the bulk of the rest of this present work.

Before going into model-based computation generally, it is helpful to first analyze deeper the notion of analog computation. Bernd Ulmann, distilling contributions from many previous authors on the subject, has provided us with a clear assessment of what analog computation *is*. I quote at length:

> First of all it should be noted that the common misconception that the difference between *digital computers* on one side and *analog computers* on the other is the fact that the former use discrete values for computations while the latter work in the regime of continuous values is wrong! In fact there were and still are analog computers that are based on purely digital elements. In addition to that even electronic analog computers are not working on continuous values — eventually everything like the integration of a current boils down to storing (i.e., counting) quantized electrons in a capacitor.
>
> If the type of values used in a computation — discrete versus continuous — is not the distinguishing feature, what else could be used to differentiate between *digital* and *analog* computers? It turns out that the difference is to be found in the structure of these two classes of machines: A digital computer in our modern sense of the word has a fixed structure concerning its constituent elements and solves problems by executing a sequence (or sequences) of instructions that implement an algorithm. These instructions are

> read from some kind of memory, thus a better term for this kind of computing machine would be *stored-program digital computer* since this describes both features of such a machine: Its ability to execute instructions fetched from a memory subsystem and working with numbers that are represented as streams of digits.
>
> An analog computer on the other hand is based on a completely different paradigm: Its internal structure is not fixed — in fact, a problem is solved on such a machine by changing its structure in a suitable way to generate a *model*, a so-called *analog* of the problem. This analog is then used to *analyze* or *simulate* the problem to be solved. Thus the structure of an analog computer that has been set up to tackle a specific problem represents the problem itself while a stored-program digital computer keeps its structure and only its controlling program changes. (Ulmann, 2013, p. 2)

It should be noted that the analysis here, attempting to clarify the confusion of what analog computation is by disentangling continuous variables and analogy machines, departs from a related attempt by Trenholme (1994). Trenholme, in order to characterize 'naturalistic' analog simulation (which I will return to shortly) wants to resolve confusion over analog vs. digital in another direction: to analog vs. symbolic. The discussion of this approach in Asaro (2011) is worthwhile, but I do not find it as convincing or as clarifying as my present approach, as there are too many commonalities with what I find disagreeable in Piccinini (2015). It also does not appear to disentangle the variable value dimension and the model-based dimension, which I think needs to happen in any case. Though, as I note elsewhere, I do find Trenholme's analysis of analog simulation as defined by isomorphisms between causal structures to be in the right direction.

Going back to von Neumann, we find the beginning of the next most essential aspect of analog computation (and computation in general)—that computation depends on the *use* of a system. Even though the components in a device might be ultimately continuous, it depends on how we intend to use the system. We also see evidence that the confusion concerning analog computers has been around for quite some time:

> The electromechanical relay, or the vacuum tube, when properly used, are undoubtedly all-or-none organs. Indeed, they are the prototypes of such organs. Yet both of them are in reality complicated analogy mechanisms, which upon appropriately adjusted stimulation respond continuously, linearly or non-linearly, and exhibit the phenomena of "breakdown" or "all-or-none" response only under

> very particular conditions of operation. (Neumann, 1963, p. 297-298)

We should be careful in parsing this particular quote, since von Neumann uses 'analogy' and 'continuously' in the same sentence. It seems that here he has also conflated analogy with continuity, although it is unclear whether he means something more than continuous.[5] In other places he seems to maintain a clearer distinction, but in any case the issue has not gotten clearer in the more modern statements quoted earlier. However, what we see is that 'proper use' is essential to defining computation. The close relationship historically between analog computation and modeling in science is also discussed by Care (2010) in depth. These ideas will be discussed throughout the paper, but for now we can state more accurately what we mean by an analog computer.

Definition 1 (Analog Computer)**.** *An analog computer is a device whose internal structure is malleable and can be set up to have similarities to aspects of the class of problems it is used to solve. Additionally, these similarities by themselves should be sufficient to form a* ***model*** *of the relevant class of problems. In our proper use of the device, the organs involved are interpreted by the model to function in a way that is consistent with our understanding of what would be required to solve the target problems.*

While some analog computers under this definition can indeed be considered as (ideally) implementing differential equations or having continuous organs relevant to our purposes, this must be recognized as only a subset of potential uses of such a computer. In other words, the definition does not explicitly endorse smoothness or rule out digital systems. What is more important for the notion of analog computation, and for developing a richer conception of computation, is that the user and the architecture both play important roles in their relationship to a *model*.[6] The user has to develop a model, or recognize similarities, or utilize analogical reasoning to set up the system in such a way that it can solve the problems at hand.

Our view of the architecture reflects this modeling procedure, meaning that as von Neumann notes an 'all-or-nothing' organ might be liable to be characterized under other usages as a more or less continuous valued organ. What should be clear at this stage is that analog computation utilizes a *model* to frame the use of the device, not unlike how models are used in other areas of science to represent sets of properties and relations relevant to a given inquiry. For analog computers, this seems to have historically been models incorporating similarity

[5] A similar conflation also appears in Borko (1962), where the author shifts from a model-based notion that analog devices calculate by "using analogies", but then subsequently identifies an analog device as a "continuous function computer".

[6] Thus, in the remainder of this article, the reader should note that when I use the term 'analog', even as an adjective, it does not refer to continuity.

and analogy. However, I argue this is just one type of a more general category of what can be called *model-based computation*. At this point I diverge slightly from Ulmann's statements, although the central premise is, I think, present in his work already quoted and thus I am offering more of a naturally implied extension than a meaningful divergence.

Before moving on, however, we can state that there may be particular objections to construing analog computation in the manner done here. One might say that 'analog' has taken on a new meaning, and that for all intents and purposes analog computation is just defined nowadays as computation with continuous variables. I must insist on disagreeing with this approach to redefining terminology, since it leads to unnecessary confusion. Piccinini and Bahar (2013) throw out the very idea of analog-model based computation that Ulmann has brought to our attention (and which I am agreeing with here):

> In another sense, "analog" refers to representations that bear some positive analogy to what they represent. There is evidence that nervous systems contain and manipulate analog models of their environment [...] But analog *models* need not be represented and manipulated by analog *computers*. (Piccinini and Bahar, 2013, p. 466)

The authors are correct if what they mean by analog computer is just a computer with continuous variables. However, I find the arguments and examples advocated for here (and extensively in Ulmann's work) to convincingly establish that analog computation is not (and has not even historically been) identical to computation with continuous variables. With this in mind, the authors might be sympathetic to the solution advocated for here: disentangling continuity and analogy into two separate dimensions of the general notion of computation. One dimension is that of the types of variables manipulated in components. The other dimension concerns the extent to which a computational device can *model* a target.

Later, Piccinini (2015, §12) indeed seems to recognize the confusion in the literature with continuous variables, although his discussion in the chapter largely focuses on electronic analog computers and presumes real variables. He also claims (Piccinini, 2015, p. 199) that the "notion of an analog model [...] is orthogonal to the notion of analog computation." He seems to acknowledge that this is not historically true, but he thinks it is conceptually accurate. If he means by this statement that the model is orthogonal to the types of variables manipulated, then we are in agreement. If he instead means that the model is orthogonal to what it means to compute, my account is in disagreement since I think it is clear models are representations.[7] Taking into account his view that

[7] Also, Care (2010) advocates a deep relationship between analog computing and modeling in science supported by an in depth historical analysis.

computation does not require representation seems to support the latter interpretation, and goes against the model-based notion of computation presented here.

3.2 Model-Based Computation

The notion of model-based computation is easily inferred from those definitions already provided for analog computation by Ulmann. The notion is just slightly more general, in that the model used may or may not incorporate similarities or analogies to the extent that analog models do.[8] That is, even if there are similarities in the device, these similarities may not be sufficient by themselves to form a *useful* model of the target problem class. Some other parameters might, for example, be invoked to make the model work even though it is unknown whether such a parameter corresponds to the target. It could even be that the parameter is *known* to be dissimilar to the target, but the model is sufficient nonetheless for our uses. Analog computation is then a special case more accurately thought of as shorthand for analog-model computation. It might be that good examples of model-based computation are in fact using analog models, but it is arguably a small class within the computational landscape compared to any potential model-based computation—if only for the reason that analog models are a restricted class of models in general.

But what is a model? This question has been widely addressed in the philosophy of science community, and a few brief notes might be helpful before moving forward. There are a variety of different kinds of models which are used in science. There are toy models, idealized models, scale models, mathematical models, and many more kinds of models. Each of these kinds may have overlap with other kinds, they are not exclusive of each other. All models, it seems, need a target object or set of data which is to be represented or accounted for in some way in the model. See Frigg and Hartmann (2012) for more on models in science. Model-based computation may involve many of the same aspects as other models in science.

Definition 2 (Model-based Computer)**.** *A model-based computer is a device which may have a malleable internal structure, and which can represent aspects of the class of problems it is used to solve. The representations should be sufficient to form a model of the target problem class. Under proper use, the organs in the device can be interpreted by the model to function in a manner that we take to solve the target problems. This may or may not be consistent with our understanding of the target problem class.*

[8] Although it may be an open question whether all models are in fact rooted in analogy or similarity, I do not focus on such an argument here.

It is useful to go one step further in this section, to discuss model-based simulation. This is helpful before encountering analog simulation in later sections. Model-based simulation is a refined form of model-based computation, in which the dynamics of the device are relevant for the user or target problem (as opposed to just a functional relation). Generally, the dynamics of a model-based computer may or may not model what we know about the dynamics of the target system. Simulation operates on a richer model that deems relational aspects (such as temporal or dynamical relations) of the device relevant. One can have static features represented in a model which, after use, has an output which functionally represents a useful computation concerning a target problem. However, the dynamics of using the model may be irrelevant to the kind of dynamics present in the target system. In this case we would not say that there is model-based simulation present.

Just reading the output of a slide rule, for example, does not seem to involve simulation but just accomplishes a computation with the model. Take two equal length sticks with logarithmic scales on them lined up side by side. Multiplication can be calculated by sliding one of the sticks relative to the other by a factor. That is, 2 times 4 could slide one stick by 2 on the logarithmic scale (representing a multiplication of 2). Then, one looks up the other factor and reads off the corresponding value on the other stick. In this case, 4 would be lined up with 8, the result of the calculation. It is also interesting to note that a slide rule is typically called an analog computer. Under the misconception of analog computer as necessarily involving continuous variables, what role does continuity play in the use of a slide rule? It is arguable that the continuously adjustable aspect of the device is incidental to the actual use and function of the computer since outputs are also not real numbers.

The dynamics of sliding the stick does not seem to model an algorithm for multiplying some integers. The model in this case involves not only the physical ruler, but also the reasoning and mathematics involved to create the scales encoded in the ruler. The preparation of the computing device has utilized precomputed knowledge (i.e. $\log(xy) = \log(x) + \log(y)$) to functionally output values consistent with an algorithm for multiplication, but the sliding dynamics are not particularly relevant for the computation. One could just introduce notches at intervals, but surely a notched slide rule does not suddenly become a digital computer. It is just now implementing a discrete computational model of logarithms. I think it is quite reasonable to expect that model-based computation, in general, does not necessarily include relevant dynamics with the target system. When it does, a stronger notion of model-based simulation may be applicable. We will see later an example of analog simulation in which the dynamics are relevant *and* similar.

As this is a relatively simple example, we can say a bit more about what a model consists here. For the slide ruler, our target problem is modeled by

a set of input operations which obtain, upon a functional relation, a desired output. That is, our model is the set of representations and operations upon such representations (a mapping) that are used. The end step after a specific procedure can be interpreted as our desired result, that which is to be computed. If our problem is to multiply two integers, a slide rule encodes two logarithmic scales. We operate upon such scales by sliding (adding distance) one factor, and then reading out the number mapped from the factor on the first stick to the number on the other stick.

Consider that a plastic model of a molecule consists of all aspects standing in a representational relation to an *actual* molecule (whether similar or dissimilar), and the operations we can do with the plastic pieces (which may or may not reflect 'operations' possible in actual molecules). At a minimum, a model for model-based computation likewise consists of a set of representations and a set of operations upon these representations. The representations may be linguistic or symbolic, for example, while the operations may be thought of as functional relations between them.

3.2.1 Benefits?

In the present work I am remaining relatively qualitative in my analysis of the 'computational landscape', however some general comments may be of interest concerning any formal results associated with analogy machines or model-based computation. If there is any genuine 'speed-up' to be found compared to classical computation, I think it is primarily the result of two sources. The first potential source is simply due to the architecture and type of values processed in the given system.

The second, and likely more important, source of potential speed-up in any particular model-based computer is that it may front certain information in the 'premises' of the set-up.[9] In other words, some computational work might have already been done in the design of the system. This includes pruning off certain forks in reasoning or avoiding certain lengthy calculations that do not need to be investigated or reported by the program. As a simple example, just consider Deutsch's problem and finding out whether a black box implements a balanced or constant function of four possible functions $f : \{0, 1\} \to \{0, 1\}$.

A classical computer requires two evaluations of the black box, sending both a 0 and 1 through. We learn not only whether it is constant or balanced, but also *which* of the four functions is performed. A quantum computer can, by throwing out the irrelevant information of the specific function and encoding the global property of the function cleverly into the phase, tell us in one go whether the box implements a constant or balanced function. Our model of the

[9] The benefits or drawbacks of any specific device depends on the model involved, and thus it makes no sense to talk of 'general' speed-up (or slow-down) results for model-based computing.

problem works with the architecture to cleverly set up the computation such that it (ideally) tells us only what we need to know and nothing more.

Any complexity claims should always be aware of these 'fronted' or indirectly utilized resources. If we haven't recognized these resources adequately, we might be mislead by certain claims of speed-up or complexity. In statements such as the following from Rubel, for example, we can see that these resources are alluded to by mentioning that the scientist has a 'feel' for the computing device:

> It is fashionable nowadays to downgrade analog computers, largely because of their unreliability and lack of high accuracy (roughly one-tenth of one percent at best). But analog computers, besides their versatility, are extremely fast at what they do, which is solving differential equations. In principle, they act instantaneously and in real time. Further, in contrast to the situation in digital computing, the operator of an analog computer has an extremely good "feel" for what the computer is doing. Analog computers are still unrivaled when a large number of closely related differential equations must be solved. (Rubel, 1985, p. 78-79)

While Rubel is specifically referring to analog computers, I think the statement is generally applicable to model-based computation. It is this 'feel' that I think imparts some of the benefits to model-based computation, since one has already done some work in constructing the model and in understanding how to work with the particular architecture. Many models provide the user with a 'feel' for the target problem or system, even with the acknowledgement that in reality there are certain features of the model which are non-representative or known to be false. See e.g. Frigg and Hartmann (2012, §4.2). By utilizing a model in computation, the features of the model (such as idealization, etc.) have restricted the computational possibilities to things which fit the use—thus streamlining any process to just those which are relevant for the User. For this reason, a model-based computer (or analog computer) is not necessarily a general purpose computer.[10]

3.3 Computational Claims about the Brain

I would like to now draw attention to a few important areas of debate, and offer some preliminary thoughts on how they might be viewed differently under the reframing of the computational landscape offered here. The first area

[10] Piccinini (2015, p. 203) also notes that we should distinguish between 'general purpose' as referencing computational universality (i.e. a Universal Turing machine for digital computation), and 'general purpose' in the sense that we can do many things with the same device. A model-based computer in the account offered here might be neither universal nor useful for many different things.

concerns computational claims about the brain. In one influential take, Searle (1990) equates the question *Is the Brain a Digital Computer?* with *Are Brain Processes Computational?*. After the preceding discussion, this seems like a mistake.[11] Digital computers are of course computational, but something that is computational is *not* necessarily *digital*. A digital computer may be repurposed in another context (with another user) to implement another form of computation, as noted by von Neumann:

> By an all-or-none organ we should rather mean one which fulfills the following two conditions. First, it functions in the all-or-none manner under certain suitable operating conditions. Second, these operating conditions are the ones under which it is normally used; they represent the functionally normal state of affairs within the large organism, of which it forms a part. Thus the important fact is not whether an organ has necessarily and under all conditions the all-or-none character—this is probably never the case—but rather whether in its proper context it functions primarily, and appears to be intended to function primarily, as an all-or-none organ. (Neumann, 1963, p. 298)

To be fair, Searle's discussion does touch upon some very legitimate issues with these questions. However, it is not clear that his discussion translates easily for the notion of model-based computation advocated for here. I want to agree with Searle's (and von Neumann's) comments on *use* being fundamental to computation, but avoid the framing of computation as equivalent to *digital* computation. Digital computation is but one *subset* of potential user-dependent contexts which may constitute a computational device. Not only does a model-based notion of computation help clear this issue up, but it importantly emphasizes at its core the user-dependent context which is so central to the notion of computation generally. This helps us grasp better what alternative models of computation are doing for us, namely that they can be used to subjectively prune away irrelevance or to emphasize certain relevancies for particular uses.

Then, what would it mean if we asked *Is the Brain a Model-based Computer?*, and is this different still from Searle's second question *Are Brain Processes Computational?* In the scope of this present paper I cannot answer all of the interesting questions brought up in this topic, but I can discuss one recent approach in cognitive science that arguably fits the notion of model-based computer.

[11] Even more so than Searle himself might have admitted.

3.3.1 Bayesian Brain and Generative Modeling

There is a growing use of Bayesian probabilistic methods and hierarchical generative models (HGM) in cognitive science. See e.g. Friston (2010) and Clark (2013). Some of this literature can be taken as arguing that the brain be considered a model-based computer as we have defined it here: that it is a device whose evolution in time effectively performs computations based on a *model*. Take Friston's description as an example:

> The Bayesian brain hypothesis uses Bayesian probability theory to formulate perception as a constructive process based on internal or generative models. The underlying idea is that the brain has a model of the world that it tries to optimize using sensory inputs. [...] In this view, the brain is an inference machine that actively predicts and explains its sensations. (Friston, 2010, p. 129)

This approach is argued by the authors to have the capacity of unifying several areas of cognitive science. Whether the specifically Bayesian approach is the final unifier may still be at question. Nonetheless, the approach not only fits the model-based account of computation I have advocated here, but even seems to fit the more restricted sense of analog computation since the modeling that a Bayesian brain is doing is related via similarity to the external world. The brain, under this view, is constantly simulating the world and adjusting its model according to the errors experienced. The HGM is amplifying relevant or similar features of a model via feedback with the environment, while dissimilar features fall out of focus (and, under the Bayesian approach, obtain lower probabilities).

Under this framework, we would answer 'yes' to the question of whether the brain is a model-based computer, and also 'yes' to the question of whether brain processes are computational. However, this may be a bit premature since we have noticed that computation is dependent on a user—and what would be *using* this model-based computer? This is no trivial problem, and in fact relates to longstanding mind/brain problems and what is called the "homunculus fallacy" (HF). See e.g. Searle (1990, §V). Can the model-based conception of computation add anything new to this problem?

Without being overly conclusive, I suggest that the hierarchical generative model of cognition may be a good step towards addressing the user problem. The reason is that it simply accepts a finite regress and offers a more general notion of model-based computation.[12] While this isn't solving the problem (or avoiding the fallacy) in the traditional sense, it is simply not so unreasonable

[12] Attempting to mitigate or explicitly accepting the HF is a required step, since as Searle notes, "... The homunculus fallacy is endemic to computational models of cognition and cannot be removed by the standard recursive decomposition arguments." Searle (1990, p. 36) What can be done, I argue, is to put a new spin on the issue.

to suppose that the brain—as a computational system—involves complex hierarchical modeling of the external world. The representations in this model no doubt still succumb to HF objections, but not in a naive way.

The slightly more sophisticated view does not succumb to an infinite regress traditionally associated with the homunculus fallacy, since there are finite levels in the hierarchy. The homunculus could just be the topmost level in the hierarchical generative model, and it 'uses' the computations from lower levels in the hierarchy. For a discussion of this approach in the foundations of cybernetics, which HGMs are arguably appropriating in a new way, see e.g. Ashby (1958, §4/7, 4/11, 13/14). Now, a reader familiar with the HF would likely object and say that the topmost level of the hierarchy is still problematic, since it is not 'used' by a higher level user. I do not know a way out of this objection, nor whether it is useful to reconcile. I can only say that the entire integrated 'body + HGM' system definitely seems to use the HGM, for all of the reasons why people think HGM is a good model of cognition in the first place.

In any case, it seems that a sophisticated model-based notion of computation does not do worse for computational claims about the brain than what has been accomplished previously. The brain doesn't need to be a digital computer, or a general purpose analog computer.[13] However, it is clear that a brain-like system which utilizes a model (i.e. does model-based computation) of the external environment to generate minimum error or minimum surprise is much different than Searle's formulation of these issues.

A potential benefit of this view of cognition was hinted at by Kenneth Craik in a 1943 publication:

> [I]n the particular case of our own nervous systems, the reason why I regard them as modelling the real process is that they permit trial of alternatives, in, e.g. bridge design, to proceed on a cheaper and smaller scale than if each bridge in turn were built and tried by sending a train over it, to see whether it was sufficiently strong. [...]
>
> It is likely then that the nervous system is in a fortunate position, as far as modelling physical processes is concerned, in that it has only to produce combinations of excited arcs, not physical objects; its 'answer' need only be a combination of consistent patterns of excitation—not a new object that is physically and chemically

[13] A similar point is made by Piccinini and Bahar (2013), however their re-structuring of the cognitive computationalism debate ends with a hybrid notion of computation for cognition that combines digital computation with the conflated notion of analog computation discussed earlier. Besides the obvious disagreement we have in defining analog computation, I think the notion of model-based computation discussed here supports in some ways their assertion that neural computation is a kind of its own (not particularly digital, not particularly continuous).

> stable. [...]
>
> My hypothesis then is that thought models, or parallels, reality—that its essential feature is not 'the mind', 'the self', 'sense-data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. (Craik, 1943, p. 52-57)

In cognitive science, this is now known as 'mental modelling', and it seems quite consistent with model-based computation as I have construed it here. Thagard characterizes mental models as "psychological representations that have the same relational structure as what they represent." Thagard (2010, p. 447) I argue that the benefits which Thagard and Craik attribute to the brain's ability to model are just the same as what I have outlined earlier. The low-level processes which contribute to model-formation end up trimming irrelevant representations off, and of course this is combined with parallelism in the neural architecture. Together with a system-defined *use* (i.e. the survival of a human or the optimization of some task), we can legitimately characterize 'mental models' in cognition as an instance of model-based computation. We can also find model-based computation in external scientific models, and in the next section I discuss a special case of MBC.

3.4 Analog Simulation in Physics

For the second kind of computational claim to be discussed, I move to modeling in physics. A few recent publications in the physical sciences (along with some philosophy of physics) have drawn attention to the use of analog models in scientific reasoning. One notable example is that of fluid systems displaying analogous phenomena to Hawking radiation (the phenomena of photons escaping the event horizon of a black hole). See Unruh (2008). These models have been argued, under strict conditions, to be performing analog *simulation* by Dardashti, Thébault, and Winsberg (2017). Importantly, these systems seem to allow us more access to black hole phenomena than would otherwise be possible.

The reader should already be anticipating the main point of this section: these sort of systems are *analog computers* in the clearest sense—they are based on strict similarity conditions, and as alluded to earlier are prime examples of model-based computation (specifically analogy-based). They are simulating while also displaying formal and physical similarities with the target computational problem. The type of simulation these systems do is arguably providing even stronger results than traditional simulation in which the architecture of the computing device is irrelevant to the simulated problem. However, because

of the background knowledge involved in constructing a table-top system, we might be less surprised at the outputs because we have a good 'feel' for what the system can do.

The strict models used in analog simulation are based on formal similarities (such as isomorphisms) between the systems of equations describing both the computing device (i.e. a table top fluid system) and the target system (i.e. a black hole). We mentioned earlier that for model-based simulation, the dynamics of the computation are relevant (but may not be similar). For analog simulation, the dynamics of the device must preserve relevant similarities with the dynamics of the target system. An earlier paper, by Trenholme (1994), also defined analog simulation through an isomorphism between causal structures between systems (although again, this was combined with a non-representational and metaphysical notion of isomorphism). This brings us to the last important step in this short paper, namely re-connecting our discussion with previous work concerning model-based *reasoning*.

3.5 Model-Based Reasoning in Science

The literature concerning model-based reasoning is divided among a few contexts. In philosophy of science, model-based reasoning is discussed in the sense of using a scientific model to make inferences about a target system (which the model represents). See e.g. Frigg and Hartmann (2012, §3). The systems discussed in the previous section are good examples. The scientist may, for example, use the model to justify a theory of Hawking radiation or to suggest new experimental questions. Inferences are carried from the domain of a model to assert information or knowledge about the target system. The model is assumed to have some relevant similarity to the target system.

There are, however, several other senses in which we might further distinguish flavors of model-based reasoning—particularly when talking about approaches to characterize human cognition. Consider a logical model, and the notion that a model-based reasoner chooses among sets of truth values of variables in a given premise set. That is, an agent's model is comprised of logico-semantic content, and model-based reasoning in this sense is some more or less straightforward deductive process.

This seems fairly distinct from the above notion, in that inferring from the domain of a model to a target domain seems not to be deductive, but rather abductive. The former notion of model-based reasoning has something to do with the seemingly non-deductive inferences that individuals and scientists make from a model. For example, inferring some property of a real system from an idealized model in physics.

Another sense of model-based reasoning is in diagnostics, or in artificial intelligence systems which have a model of the environment. See Davis and

Hamscher (1988). Now, this might sound very close to the notion of HGM and 'mental models' I discussed previously. However I think it is important to distinguish between model-based reasoning as somehow providing rules or guidelines in an argument or in an artificial inference system, with model-based computation as I have construed it. Model-based computation can be a part of model-based reasoning, but it isn't clear that model-based computation *is* model-based reasoning.

Reasoning is an active process (one might even say conscious), whereas computation—aside from the User's set up of a problem or the interpretation of an output—seems to be passively implemented. Model-based reasoning may likely be involved in constructing a particular computing device, but it isn't clear that what the device is doing should also be considered model-based reasoning. Or, it isn't clear that our *use* of the device as a model-based computer constitutes model-based reasoning as understood by previous work on the subject. Nonetheless, it seems to be that a more in-depth analysis of these two notions may be fruitful.

One could perhaps say that the most fundamental distinction to be made in this discussion is an *intra-inter* distinction. That is, there is the reasoning or processes involved in exploring a model—these inferences are *intra*-model. On the other hand, we have the seeming abduction or analogical argument from the model to the target system. This is *inter*-model, it exceeds the domain of the model and applies to our representation of the target system.

For a model-based computational device, it is important that our intra-model operations can produce a state which achieves inter-model significance. That is, at the end of the physical operations of the device, our *use* indicates a functional relationship of the outcome with some property of the target system (more precisely, a formal representation of the target system) or a solution of a target problem. In other words, the outcome corresponds under a 'use function' to a desired calculation. Under certain circumstances, there may be a formal similarity between the intra-model operations and the operations which could in principle be applied to the target system (or target problem). This would be identified with the special case of analog model computation and analog simulation.

3.6 Conclusion

At this stage, I offer a preliminary framework in Fig. 3.1 of computation to reflect model-based computation as discussed in this present work. I suggest that, at the very least, we should think of computation as consisting of two dimensions. One dimension represents the characteristics of physical devices and computational organs—or, our characterization of the dynamics or behavior of

such components. The other dimension has to do with the status of representation in a model—ranging from a model which is totally dissimilar to a target problem, to very similar (like in analog simulation). Importantly, these dimensions can be thought of as more or less orthogonal, and we can see why not being conscious of these dimensions might lead to the kinds of objections discussed earlier concerning analog computation. Ulmann has drawn our attention to the fact that there are two fundamental dimensions of computation that need to be distinguished to really understand what certain devices are doing.

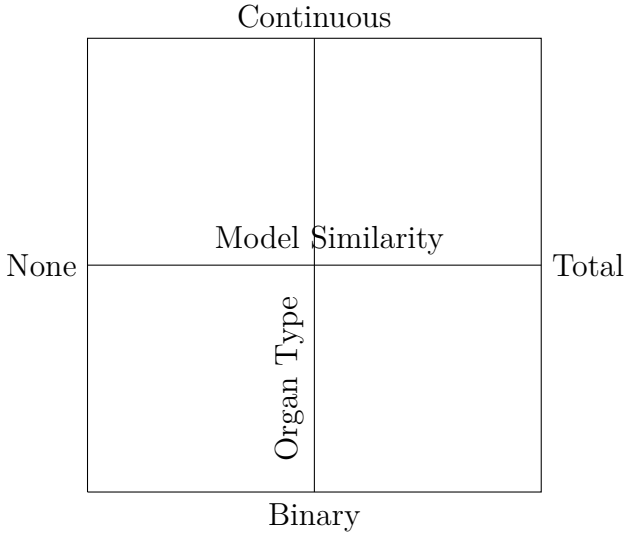


FIGURE 3.1: A proposed picture of the computational landscape in the context of model-based computation discussed here.

Briefly, lets take a look back at our examples and where they might fit into this framework. Absolute locations on the diagram for particular devices are unlikely to be uncontroversial. First, on the model similarity dimension. Take the slide rule example discussed earlier. The representations involved in the model do not particularly seem relevant or similar to multiplication. Even though our description of how the ruler implements multiplication is a step-by-step procedure, the manner in which the computation is achieved seems fairly dissimilar to a *particular* way of multiplying two integers. A slide rule is not the only means of computing multiplication, and so unless our target problem is explicitly to compute multiplication by a logarithmic means we would probably place a slide rule somewhere to the left of *total* similarity. However, that might just be the actual implicit purpose of a slide rule, in which case we would indeed mark it somewhere further to the right on the model similarity dimension.

The example concerning analog simulation, on the other hand, involves a table top model which is characterized by certain well-defined similarities (i.e. isomorphisms of causal structure) to the target problem. This is very near to total similarity. The mathematical representations of the fluid system are similar to those of the black hole, as well as the dynamics ('operations') which

occur when the system is allowed to evolve in time. Again, we can see that the more specific our model and device is, the generality of what can be computed is limited. It is unclear how multiplication could be accomplished by such a device, or why we would want to try. Quantum simulation, which is the simulation of one quantum system by a controlled quantum system, would also likely fall far to the right on the similarity axis.

Then there is the vertical dimension in the above framework. This concerns the types of values used by organs in the device. If only binary values are used, then this corresponds to a traditional digital computer. Somewhere in-between are any number of discrete-state devices. A notched slide rule would probably be in the lower right quadrant, whereas the traditional idea of a 'smooth' slide rule would be in the upper right corner, near other analogy machines and analog simulators. This fact does illustrate just how practical this two dimensional view is. In the ideal case, we could imagine some device whose organs operated with continuous variables. There is always the question of whether such values are recoverable (i.e. quantum amplitudes), but nonetheless I think the extremities on this axis are reasonably clear. General purpose digital computers would be along the bottom, perhaps in the lower left corner. A more specialized digital device like a GPU might be a bit closer to the middle, but still on the bottom. More general purpose analog devices would probably be somewhere in the upper middle. And so on.

This is all a preliminary outlook on what has so far been discussed. I have argued that this framework is beneficial for theoretical computer scientists and philosophers alike. Future work is justified, particularly in discussing more in depth case studies and attempting to plot them in such a landscape to get an overall picture of the landscape of computation. Even if the precise plotting of computational devices is not the primary goal, it seems like the exercise of discussing computation along these dimensions helps to get a grip on what one is working with in any given alternative computing device.

A model-based notion of computation helps us understand *why* certain architectures or models might perform better on, for example, optimization problems. Take D-Wave's supercooled annealing chip, for example. Its usefulness derives from a combination of architectural features and model-based considerations in the set-up of the device, and these determine the types of problems that it can be useful for. It is worth investing in because it exploits a combination of pre-computed modeling considerations with an architecture that also reflects these considerations (whether it truly exhibits a "quantum" advantage or not).

There is an interplay between our understanding and descriptions of physical systems and their dynamics, and what we consider to be our model for computation. For a general purpose digital computer, it might suffice to map a representational model (i.e. digital computation) onto our description of the

physics of the device. On the other hand, it might be useful to first describe what a system *does*, and to see what could be computed if we set up and controlled such a system in certain ways. That is, let the description of the physics of the system define our model of computation. As an example, consider Carver Mead's outline of neuromorphic computing hardware:

> The fact that we can build devices that implement the same basic operations as those the nervous system uses leads to the inevitable conclusion that we should be able to build entire systems based on the organizing principles used by the nervous system. I will refer to these systems generically as *neuromorphic systems*. We start by letting the device physics define our elementary operations. Mead, 1990, p. 1631

Specialized computational devices like IBM's TrueNorth, one of many recent attempts at neuromorphic hardware, illustrate the model-based notion of computation for neuromorphic computing. One can imagine that pairing such a device with a specialized artificial neural network (ANN) algorithm can result in model-based computational efficiencies. For example, matrix multiplication is a typical operation performed in an ANN one might want to design for in a computational device meant to implement ANN algorithms.

A convolutional neural network (CNN), for example, is an ANN whose convolutional layers learn through back propagation appropriate kernels to convolve with the input. One could think of it as a model-based neural network (specifically for image processing), whereas the neuromorphic hardware might be model-based to process typical neural network operations. The operation of a CNN is inspired by the receptive field of biological neuron vision processing. A convolution in the continuous case is defined for two functions (e.g. of time) f and g, sliding a kernel over a signal:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \tag{3.1}$$

Even though TrueNorth was a neural-inspired architecture, it was not developed explicitly for deep CNNs. Still, the result of matching efficient neuromorphic architecture with neuro-comutational models is impressive. See Esser et al. (2016). We see that as the model-based considerations and hardware architecture align, even if imperfectly, the performance on certain classes of problems increases—for example, in machine learning and image recognition tasks. The tool (or combination of tools) has a better fit to the problem, encoding certain assumptions about what should be computed, how and why.

We can understand the justification to find an architecture that reflects our model as closely as possible, such as a neuromorphic hardware specially built

for convolutional neural networks. In the special case, we have a device that performs analog simulation as outlined earlier. However, this likely comes at a cost of computational generality. We may make some pragmatic compromises in our model for ease of use, for finding an appropriate architecture, and to increase the domain of application of the device, etc. Optimizing hardware for matrix multiplication is one area which seems to be a fruitful compromise, since it is a typical operation in many popular artificial neural networks.

In conclusion, model-based computation seems to be a worthwhile notion to entertain when discussing alternative computing. If a User wants to compute certain things that are modeled better by neuromorphic or quantum computers, then using these devices might provide some computational advantage. The notion of model-based computation does not entail any kind of dramatic proposal to re-draw complexity classes or endorse any view on hyper-computation. In fact, it is clear from the discussion that complexity claims should be wary of fronted resources by the modeling process.

As a conceptual tool, model-based computation helps us get a better grasp on the landscape of computation. This account also gives an intuitive picture of what does and doesn't compute (since a computational device computes relative to a User). I have argued that this tool is useful for analyzing and understanding various kinds of computational claims. Perhaps it can also help us keep track of the emerging market for specialized computing devices.

Chapter 4

Neural Networks as Cybernetic Regulators

> The same point of view may be applied to the brain, and we can see how one part of a brain can show towards another part the objective behavioral relationship of designer to machine. We can begin to see how one part—a basal structure perhaps—can act as "designer" towards a part it dominates, towards a neural network, say. Ashby, 1958, p. 255

Cybernetics provides a useful conceptual framework for characterizing artificial neural networks (ANNs). This should be unsurprising, given the complex historical interplay of the central ideas and the interdisciplinary influences among biology, neuroscience, computer science, connectionism, and all flavors of AI. (Cordeschi, 2002) The cybernetic interpretation of ANNs, consistent with decades of subsequent development and use by control theorists (see e.g. Leigh (2012, §16)), provides an accessible level of analysis. Indeed, one might consider control theory to be an engineering discipline which matured out of some of the central ideas of cybernetics. It would only be natural for philosophers to make use of this rich conceptual history to understand what ANNs are and what they are capable of. This chapter attempts to bolster philosophical analyses in the area, in particular against the worry that ANNs are epistemically opaque, and outline how general cybernetic concepts usefully connect to modern machine learning. This is enough to justify a stand-alone project, but for the purposes of this dissertation this chapter also serves as a build up for chapter 5 and a transition from the previous chapter 3.

The apparent epistemic opacity of computer simulation and machine learning techniques means we are faced with the problem of justifying decisions based on these techniques. As noted for example by Humphreys (2009), such opacity may arise for human epistemic agents who are unable to follow all of the computational steps—and even if they could, there may be a fundamental limit to what we can understand. One example which seems to fall into the class of opaque techniques is classification using ANNs. Trained classifiers are used widely by data scientists in industry, and may be used to justify and automate

a variety of decisions. Such widespread usage will inevitably encounter objections: why should we trust in these techniques when they are opaque? It will be useful to have at least some sort of explanation ready at hand, even though it may not make the token neural network model transparent. The account offered here will hopefully reduce some opacity worries, but it might also redirect these worries to more salient issues concerning the use of ANNs. For example, *how* an ANN fits a task might not be as worrying as the fact *that* it fits.

There are a wide variety of audiences we may wish to explain (or justify) the behavior of ANNs to. There are expert explanations, and layman explanations. Expert explanations are about refining understanding for highly-trained experts in machine learning or in a related technical field such as linguistics, neuroscience, mathematics, statistics, physics, etc. The field of data science is surprisingly diverse with respect to the backgrounds that practitioners hail from. High-level explanatory analyses will assume and build upon the linear algebra, calculus, and statistics used in ANNs for machine learning purposes. They aim to "illuminate" the so-called black box by explicating mechanisms, doing statistical analyses, and trying to find the essential computational aspects and parameters of the ANN models that are relevant for giving the expert a detailed understanding of what is going on inside.

Layman explanations, on the other hand, I take to be about the conceptual foundations about what kind of objects ANNs are. Such an explanation will give a non-expert a sense of the types of behaviors and results these objects are capable of. It will help develop intuitions, which could be refined by further study and experimentation, but which should suffice for a wide range of justification inquiries. It is this sort of understanding I am aiming to characterize in this present chapter. If I am successful, I would hope that the concepts outlined here give entry level data scientists, laymen, and other interested non-experts (such as philosophers) a useful way of thinking about ANNs that is accessible and not too formal. There are, after all, more people who would like a basic conceptual understanding than there will ever be experts.

Expert attempts to illuminate so-called "black box" models will be of little use for such an audience. Individuals who are not formally inclined, or do not have expert familiarity with linear algebra and statistics will not be illuminated. The high cost of specialization required for experts to understand the details means that further insights are diminishing returns. Any additional explanatory power or justification they provide are likewise not worth the trouble. Experts can and should keep investigating for more esoteric insights. For widespread explanatory purposes, however, any worthwhile insights which change the overall justification picture for the effectiveness of ANNs will have to be translated into an accessible and intuitive picture. At present, I think the cybernetic account of ANNs is such a picture.

If analyzing an ANN is anything like analyzing a brain, it seems reasonable

to suppose we can use a common methodology or framework to understand both kinds of systems. If cybernetics (and cognitive systems theory in particular) is such a framework, which I find uncontroversial, then the epistemic situation we have with respect to an ANN falls primarily under the same jurisdiction as other similar systems. I think non-experts may find the notion of a *cybernetic regulator* a useful starting point for understanding the kinds of objects that ANNs are. I argue they are cybernetic regulators with large amounts of parameters. Furthermore, their regulation is error-controlled—that is, the parameters which achieve regulation are adjusted based on a measure of the current performance. These concepts are expanded upon in the next few sections.

4.1 Some Basics

Perhaps the most well known examples of cybernetic regulators are mechanical governors, for example centrifugal governors (see e.g. Maxwell, 1868). The mechanical feedback in a governor controls, for example, the amount of fluid propellant for an engine. A crucial machine in the subsequent history of cybernetics was Ashby's table top device called a homeostat. The purpose of the homeostat was to be demonstrative of core principles: it showed how a simple system could control, through feedback, environmental disturbances and 'adapt' to a stable state. See Ashby (1958, §5/14), and also Asaro (2008) and Cordeschi (2002, §4, §5) for some context on the homeostat. Another simple and familiar example is a thermostat: a system which controls the air temperature in a given volume (e.g. a room). It has a sensor for the current temperature, and a method of determining the distance (and direction) between the sensed temperature and the current *regulatory goal*. This goal is set by the user. Furthermore, the regulator either contains (or is coupled to) a mechanism whose actions ideally produce a state of the world aligning with the regulatory goal. That is, under the control of a thermostat, the temperature of a room should equal the set temperature.

Cybernetics as a framework gives us the concepts and intuitions we need to understand the behavior of objects like thermostats. In this chapter I claim it is also a very useful framework for developing intuitions about ANNs. I argue that the bulk of understanding we wish to obtain about ANNs is already present in the cybernetic picture, specifically that provided by Ashby (1958). ANNs are cybernetic regulators very similar to a thermostat. Once the basic concepts are introduced, and this similarity is understood, we can see that some aspects in modern usage of ANNs (such as regularization) are relatively easy to grasp. This perspective is useful for a large demographic of individuals who may be in search of an explanatory framework to develop intuitions about the kinds of objects ANNs are.

First, a very brief introduction to ANNs. For more details on the fundamentals of ANNs, see for example Basheer and Hajmeer (2000). A basic example of an ANN is a perceptron which provides basic predictions or classifcations of a given input stimulus or data example $\mathbf{x}$ (a vector of features). It is composed of nodes (neurons) with activation functions, weights $\mathbf{w}$, and biases b. A simple perceptron has no intermediate nodes, and no layers in-between the input and output layers. It can separate linearly separable data, and its activation function is a Heaviside function:

$$H(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

We apply the Heaviside to the dot product of weights and data, $H(\mathbf{w} \cdot \mathbf{x} + b)$. This has the effect of producing an artificial 'neuron' with a *threshold*. The neuron "fires" when the argument of the function (which is a number) exceeds the threshold. The output of the function is a prediction about the data example. A multi-layer perceptron (MLP) with non-linear activation functions (e.g. ReLu, sigmoid) is capable of approximating arbitrary non-linear functions. Hence, one way to understand ANNs is as function approximators. An MLP effectively has non-linear perceptrons *hidden* inside of itself which activate on input from another layer in the network, and not directly on a stimulus (data). The weights $\mathbf{w}$ in an MLP are a matrix, as each node is connected between the layers (they are *dense*). Large numbers of hidden layers in an ANN is referred to as a "deep" neural network.

ANNs *learn* in a systematic and automated manner, where input data is transformed on a forward pass resulting in an output. This output could for example be a class prediction in a classifier, or an action prescription in reinforcement learning. Weights are updated on a backward pass according to a form of gradient descent on the error surface, where a weight's contribution to the model's performance is calculated and adjusted for future forward passes. An algorithm which propagates errors back through the network is called a back propagation algorithm. The end goal is to have a trained network which can predict or classify the data (e.g. images), as well as *new* data that the network has not trained on.

I will discuss primarily supervised learning, which occurs when we feed data examples (such as images of dogs and cats) into the network, knowing which images are cats and which are dogs and assigning them labels (0 and 1, for example). There are other types of learning, but my discussion does not significantly hinge on the differences between them. For example, the reader may have heard of unsupervised learning, which unfortunately sounds like it is fundamentally different from supervised learning. However, the basics are still the same. Indeed, Yann LeCun has stated that he prefers to use the term "self-supervised"

learning, since the algorithm itself is coming up with labels (and therefore the error gradient). A mixture of machine and human annotated data is also possible, for example in so-called active learning. See for example Holzinger (2016) for more details.

The important point is that there is a *feedback loop* which communicates error to a model, which is then used by an update mechanism to improve the model. I take this to be the central conceptual point for understanding the kind of objects ANNs are. If ANNs and the associated methodologies were considered to be opaque, such a conceptual understanding offers significant transparency. Essentially, I am claiming that ANNs are best understood conceptually as cybernetic regulators. This will for some appear to be trivially uninteresting, bringing up old history. However, as I have noted, my motivation is in large part to find an explanatory framework suitable for rendering ANNs transparent to a growing demographic of interested parties. Some parts of this article may go beyond what can be considered accessible to the entire demographic, but I include them in an attempt to bolster what I have already found to be an effective way to reduce the opacity of ANNs for laymen.

4.2 What Exactly is a Cybernetic Regulator?

A cybernetic regulator is part of, or interacts with, a complex system. It controls environmental disturbances by appropriate actions, resulting in a state of the world aligning with a regulatory goal. We call the system to be regulated the *reguland*. Ashby (1958, §11) outlines the idea with a very simple decision game between two players. Take the very simple case of a thermostat above. This example is widely used historically in the literature in and around the development of cybernetic ideas, and Cordeschi (2002) details many influential arguments using thermostats. We can represent the 'thermostat game' in table form, at first just considering a broken thermostat which only regulates into two classes of states. That is, the potential goal temperatures are not on a continuum or discrete multi-partite scale, but simply partitioned into two distinguishable classes of states. Lets say above and below 20°C.

	R_1	R_2
E_1	<20	≥ 20
E_2	<20	≥ 20

Both players E and R ('Environment', 'Regulator') have access to their actions (picking rows and columns, respectively), and can see all possible outcomes represented in the table. E goes first, choosing a row, and R plays a column for some outcome at the intersection of E_i and R_j. In the case of the

thermostat, this is the set temperature (greater or less than 20°C). In the special case above, *the number of potential plays R can make equals the number of possible outcomes of the world, and the number of possible disturbances from the environment.* Lets say that E_1 is cold air less than 20°C entering from an open window, and E_2 is hot air above 20°C entering. This case is uninteresting except to note that R has a very simple strategy to achieve perfect regulation (always 'winning') for either goal. What the R_1 and R_2 plays look like are left to the reader as an exercise.[1]

Consider now a better thermostat which can target integer degrees between 15°C and 25°C. There might be environmental disturbances for which R has no appropriate play to obtain the desired state of the world (set temperature of the room).

	R_j		
	15	21	12
	29	22	21
E_i	21	16	19
	21	24	17
	23	20	21

Lets say the temperature is set to exactly 21°C. In the set up of this game—i.e. the characterization of the regulator and its environment—there is always a winning play R can do to achieve the desired state of the world, no matter what the environment does. If, however, the temperature is set to 18°C, there is no play R can make which results in this state of the world *for any possible* E_i. We also see that there are states of the world which cannot be the goal of our thermostat (by stipulation of the example). On the whole, this thermostat is still pretty bad. We can imagine increasing its regulatory capacity (and performance) by increasing the number of regulatory goals it can be 'programmed' to. In this case, there is no additional *plays* which the R needs to be capable of. Extending the setting range from 10°C up to 30°C means at least *if* our goal is on the table, R has a response under some plays from E.

Ideally, a perfect regulator would be able to control any E_i and yield the regulatory goal as the state of the world. In the above game, the only possibility for perfect regulation occurs when the goal is 21°C. The *variation* of regulatory responses by R is, in this sense, insufficient to control the variation in E. We can try augmenting the game with more moves for R, which perhaps result in more comprehensive regulatory capacities. Depending on the mechanisms of control in the overall system, it may be possible to improve regulation for a large class of goals just by increasing the number of R's potential moves until regulation meets some threshold of performance. If we just consider in this

[1] In the case of E_1 and E_2 what does R need to *do*?

toy example that columns (plays) for R are made up of random integer entries between 10°C to 30°C, eventually R will have a move for each E_i which results in any desired temperature (assuming R has the ability to see the 'game' and choose the appropriate actions). It can become a perfect regulator for all goals it can aim at.

This example provides the basic intuition for a general law of the behaviors of regulators construed in this way, first provided by Ashby, 1958. Broadly, Ashby's *Law of Requisite Variety* (LRV) says that "only variety in R can force down the variety due to $[E]$". A good regulator R has a sufficiently variable strategy profile for responding to moves by E, enabling it to accurately and effectively control the outcomes of the game. Ashby, 1991b provides a discussion of the LRV in information theoretic terms, influenced by the work of Shannon's communication theory.[2] A regulator, in effect, reduces the 'flow' of information from environmental disturbances passing into the system. If the temperature of a room should be regulated by a well-designed thermostat, the fluctuations of cold air entering the room from an open window should not result in equivalent fluctuations of the room's temperature.

4.2.1 Error-Controlled Regulator and Feedback

One way in which a regulator can improve its performance is by *learning*. Imagine that the plays available to R are chosen from a uniformly random distribution in response to the input. Its regulatory performance would not be good, but by equipping R with a mechanism which reports the errors back into the regulator we might be able to take advantage of that information. We could use it to make improvements for future attempts to control similar environmental disturbances. We could adjust or update the distribution from which plays are chosen. A simple example of reinforcement learning treats the distribution as a bag of colored marbles. We choose plays by picking marbles at random from the bag.

Lets say there is a uniform distribution over 5 regulatory plays, the probability for each play is $\frac{1}{5}$. In marbles, lets say there are 5 marbles of 5 different colors for a total of 25 in the bag. The environment throws a situation at me (itself assumed part of its *own* distribution of situations), and I draw randomly from the bag for my response in an attempt to control the situation to a desired outcome. If it is a positive outcome aligning with my goal, I note the color of the marble and add another of the same color to the bag. Lets say it was a blue marble. Now there are 26 marbles, 6 of them blue. The reader can check that the probability of choosing a blue marble at random from the bag is now greater, and since probabilities sum to one, the probability of each other color

[2] In fact, it seems to be Ashby's intention to think of the concept of variety (of distinguishable physical states of a system) as underpinning Shannon's notion of information. In a two state system, these states are coded into bits.

has gone down. We continue in such a way, adding marbles to reinforce plays which achieve an outcome aligning with the regulatory goal.

Without going into too much detail, it is clear that over time we can adjust the random distribution in a deterministic way which *adapts* or *learns* a distribution of responses that increases regulatory performance. While there are limitations to the kind of learning in this toy example, we can imagine that a regulatory system like a thermostat could begin with poor regulatory capacity and random responses, but slowly adapt in a determined way. Over time, with the right reinforcements, such adaptation could result in much more effective control of typical temperature disturbances. Trainable weights in an ANN are adjusted in a similar way, called back-propagation. In other words, information of the error (distance) between the play (prediction attempt) and the actual goal is fed back into the regulatory mechanism. This information can then be used to adjust parts of the mechanism, which will change future control actions. If the best actions are represented as minima in an abstract landscape, there are algorithms which systematically descend that landscape based on the error information. Gradient descent (and stochastic gradient descent) are standard algorithms which do this. Thus, it makes sense to treat ANNs as error-controlled cybernetic regulators.

4.3 Shattering and the VC Dimension

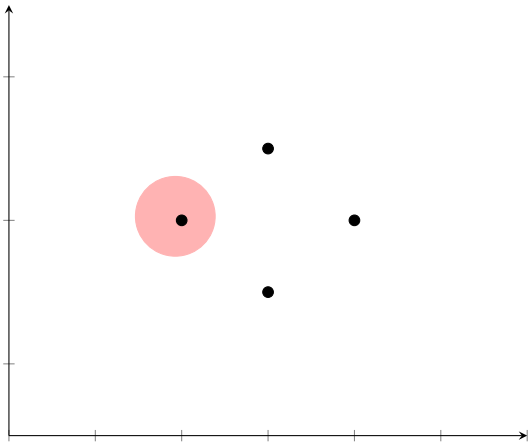
We are on our way to characterize artificial neural networks as cybernetic regulators. First, we make a pit stop to discuss an important mathematical concept called *shattering* which underlies formal analyses of machine learning techniques. If we conceptualize shattering in game theoretic terms as above, then shattering is just another regulatory game. In set theoretic terms, class R shatters set E if we can construct the power set like so:

$$\mathcal{P}(E) = \{r \cap E \mid r \in R\} \tag{4.2}$$

R can be thought of as a set of plays r (themselves sets), E a set of environmental disturbances, and $\mathcal{P}(E)$ is all combinations of disturbances in E. When each member of the power set $\mathcal{P}(E)$ can be captured or controlled by an appropriate play r, we say that R shatters E. A regulator built with the capacity to shatter all potential disturbances would be a universal regulator. Practically, we suppose there are finite disturbances generated in a distribution D, where our typical set of disturbances are a subset of this distribution: $E \subset D$.

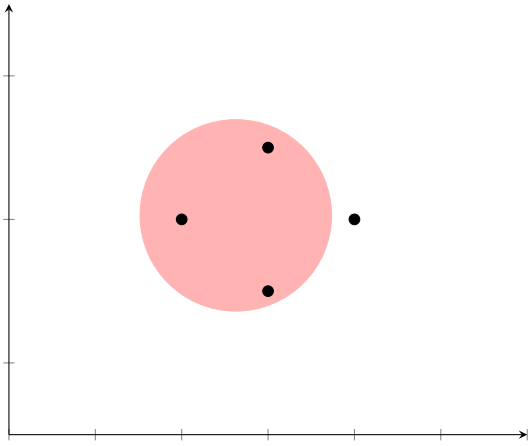
As a visual example, we can think of R as a set of curves or shapes which attempt to separate E as a set of data points. Consider the simple example of four data points, each equally distanced along the circumference of a unit circle. We try to shatter these points by using circles which can be placed anywhere in

the plane and scaled smaller or bigger. Shattering the set of these points using the set of circles of any size and location in the plane means we want to be able to capture any set of points within a circle. Circles are our regulatory *plays*, and we have infinitely many of different radius and location at our disposal.



$$x^2 + y^2 = r^2$$

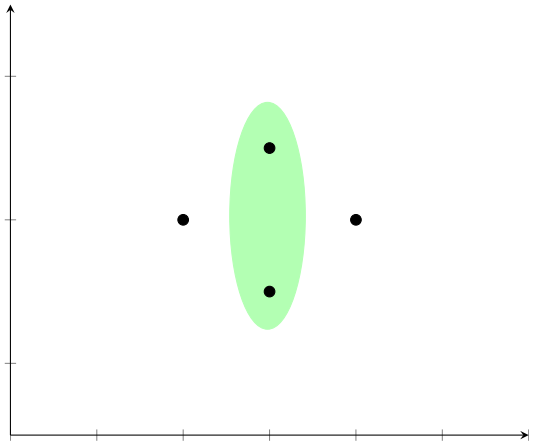
Above, we can see that it is trivial to capture each singular point. However, we also want to capture all combinations of points. Shattering means being able to construct the power set of the set of points using only circles.



$$x^2 + y^2 = r^2$$

We can also capture in a circle the four sets of two adjacent points, as well as the four sets of three adjacent points. All four points is also trivial. However, with a circle of *any* radius or center, we cannot capture *just* the two non-adjacent points across from each other without also including one of the other points. Therefore, we cannot construct the power set containing all combinations of these points. To do so, we have to increase the regulatory capacity of our set of plays (i.e. set of shapes). Whereas in the simple game Ashby provides we

needed another "move" in the strategy profile, in this game we will need to use another shape which can capture the two sets of opposing points. By separating the radius into two parameters, we can use ellipses to squeeze between the other points. Ellipses *shatter* this simple set of points.



$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

This further demonstrates the intuition that if we continue to increase the strategy profile of our regulator, the parameters determining the fitting capacity, we will be better able to respond accurately to disturbances. In an ANN, this translates to an intuition that if we increase the number of *trainable* parameters, after training for the same amount of time (or steps) we will have a more closely fit model than with fewer parameters. The model more closely fits the training distribution, however, and may not indicate generalizability. After increasing parameters, we may notice diminishing returns in the capacity of our model. Thus, another measure is actually more important than the number of parameters. This is called the Vapnik-Chervonenkis (VC) dimension, and it measures the cardinality of the largest data set shatterable by the set of plays. In the toy example above, the set of circles has a VC dimension of 3, whereas ellipses along either axis have a VC dimension of 4. While increasing the number of parameters may increase the VC dimension, depending on the problem the VC dimension may exceed the number of parameters.

> The use of the VC-dimension in statistical learning theory runs against the idea that the generalizability of a theory goes along with its simplicity, calculated in terms of the number of its parameters. Standard examples exist of hypothesis classes with low VC-dimension and a very large number of parameters (e.g., support vector machines), and vice versa. Indeed, the set of classifiers $\{\text{sign}(\sin bx) \mid b > 0\}$ shatters any finite set of distinct points distributed along the x-axis, although they are governed by just one parameter. Corfield, Schölkopf, and Vapnik, 2009, p. 55

An interesting upshot to the game-theoretic conceptualization offered here is that we can relate the VC dimension in machine learning (and set shattering) to the notion of regulation (or control) of a cybernetic system. The VC dimension can then be thought of as a formal measure of the ability of available regulatory mechanisms to squeeze the input of a system into some defined goal state, i.e. a measure of control capacity. Importantly, the VC dimension does not always coincide with the number of parameters (Vapnik, 2000, p. 82-83). However, for two systems characterized by the same VC dimension (regulatory capacity), one or the other may be much simpler or more efficient. This brings me to the next important point about how we understand the kinds of objects that ANNs are. They are not just cybernetic regulators with large amounts of parameters, but the *mechanisms* they utilize are non-trivial. They relate in some relevant way to the problem (data) at hand—in other words they can be considered *models*. We may then, using for example the VC dimension as a guiding measure, distinguish between good and bad models.

4.4 Restating the Good Regulator Theorem

In getting a grip on the kinds of objects that ANNs are, I have discussed so far how regulatory capacity can be thought of in game-theoretic terms, and introduced the VC dimension as a better measure than the number of parameters. It is furthermore pragmatically useful to think of (trained) ANNs as regulatory *models*.[3] For present purposes, it is important to outline how we can distinguish the important features of these models, and how we measure their effectiveness. If we can get a sense of what a *good* ANN model looks like, we can reduce our sense of epistemic opacity about how they function. This brings me now to what is known as the Good Regulator Theorem (GRT) from Conant and Ashby (1970):

Definition 3. *The simplest optimal regulator R of a reguland E produces actions $r \in R$ which are related to the events $e \in E$ by a mapping $h : E \mapsto R$.*

If we are to relate the theorem to modern ANN models, this formulation is rather unclear. We can imagine that such a theorem aims to state a complexity criterion for *good* machine learning models. However, I think it needs to be reformulated. For context, the authors seem to claim that unnecessarily complex (but still optimal) regulators are not models, which I think is an oversight:

> [The] best regulator of a system is one which is a model of that system in the sense that the regulator's actions are merely the system's actions as seen through a mapping h. [...]

[3] ANNs are already commonly referred to as models, I am just clarifying the conceptual sense in which I see them as models.

> [The Theorem] leaves open the possibility that there are regulators which are just as successful (just as 'optimal') as the simplest optimal regulator(s) but which are unnecessarily complex. In this regard, the theorem can be interpreted as saying that although not all optimal regulators are models of their regulands, the ones which are not are all unnecessarily complex. Conant and Ashby, 1970

What I think we want to actually say is that unnecessarily complex regulators are *bad* models, but they are still models. But what exactly is *unnecessary* complexity? Even though two models may have the same shattering capacity in principle, having the same VC dimension, the theorem attempts to state the intuition that still one model may be better than the other. The original formulation of the theorem doesn't seem to address the case that a regulator (ANN) can be a model but still be unnecessarily complex—for example it's performance is unimproved when more regulatory capacity (parameters) is added. Also, what the authors intend to be optimized is unclear when we try to relate the theorem to ANNs. To be fair, it was originally about abstract regulators, but I think a restatement of the theorem and a corollary may be warranted.

Definition 4. *For a set of regulators R_d with VC dimension d, we can impose an order $\leq$ on them according to their ability to reduce some relevant costs, increase some performance measures, and the number of trainable parameters they contain.*

With such an ordered set $\langle R_d, \leq \rangle$ we can clarify the model-mapping between the regulator's actions and events in the environment. It is taken as given that even if a regulator is not optimal we can always construct a model-mapping, however convoluted it may be. We just want a way to rank the models. Additionally, we need a measure of complexity to rank the complexity of a given representation of a regulator. We want the simplest representation of the regulator from the class of representations C_R available. Measuring the representation of the model by Kolmogorov complexity allows us to define what I think is a sufficiently updated GRT:

Definition 5. *The simplest optimal regulator R_O is both (i) the upper bound in the partially ordered set $\langle R_d, \leq \rangle$, and (ii) represented by $c_o \in C_R$ such that $K(c_o) \leq K(c_i)$ for all other $c_i \in C_R$.*

The simplest optimal regulator will be an optimal regulator with the lowest Kolmogorov complexity. When doing a machine learning task, we might scoff at the fact that there are millions of trainable parameters. This may give us the impression of a black box, populated by unnecessary amounts of parameters, inefficient in their VC capacity. However, for ANNS, these parameters are not hidden, and epistemic clarity about ANNs can still be enhanced further. Some

examples of more specific kinds of ANNs not only reduces the epistemic opacity worry, but it also trains our intuitions about the good regulator theorem—and what kinds of complexity is unnecessary.

There are many different kinds of ANNs. Data scientists use them for different purposes, in part informed by the motivations that lead to their construction. Convolutional neural networks (CNNs), for example, learn filters (kernels) to pass over an image, detecting *features* such as edges. The learned structure of these networks fulfill to some relevant extent what we perceive to be the goals of receptive fields in human vision. They have a lot of parameters, of course, but we can understand why they are good at classifying images. Unnecessary complexity in this context means something like extra convolutional layers or large kernel sizes which do not increase performance. Or, information which is compressed out by an effective compression of the network. CNNs do however fall prey to classifying images with the correct features, even though the features are not in the correct arrangement. So there is some structure or complexity which is absent, if we intend to truly regulate the superclass of image data with CNN models. Capsule networks improve CNNs for some tasks by preserving spatial relationships between features in images (learning a pose matrix). See e.g. Sabour, Frosst, and Hinton (2017). There is a great computational complexity cost for capsule networks to overcome such a drawback, yet the argument can be made that the extra complexity in the capsule network is necessary. It is improving some relevant aspect of the model with respect to the target problem, not just adding parameters willy nilly.

These examples continue to deflate the epistemic opacity worries of ANNs, providing an insight into the meaningful structure behind a sea of regulatory parameters. We could say that an ANN which did not have these additional structures, yet had the same VC dimension, would need to be unnecessarily complex. Or, at the very least, the extra complexity would be valued poorly by an expert user compared to the structured complexity found in advanced networks like a capsule network. It would lack the model-specific organization we think is relevant for regulating the typical (i.e. training) distribution of environmental disturbances. Or, worse, it would just memorize its task, a classic worry of ANNs. In any case, I am supporting the intuition that *if* we could measure the Kolmogorov complexity of an unnecessarily complex model it would be greater than that of the model with the additional structure *even if* it has more parameters. In the case of a CNN or capsule network, our descriptions even use shortcuts: feature detectors, Gabor filters, pose matrices, etc. These shortcuts represent concepts which have already reduced the opacity of the ANN, because they are specific mechanisms.

So, why do we use certain ANN architectures for certain problems? Such a question will for many have an obvious answer: the particular mechanisms involved in the different architectures are better suited for handling different

classes of data. In the language of cybernetics, they are better regulators for certain kinds of inputs (e.g. images) because their internal structure more closely resembles a (good) model relevant for processing the specialized input distribution. Temporal patterns might be better recognized or classified by a recurrent neural network. Image classification utilizes convolutional or capsule neural networks. The fact that we can distinguish between these different kinds of ANNs should tell us immediately that their construction is not arbitrary, and that at least somebody knows more or less the mechanisms we will find in each of them. Their complexity is composed—at least to some degree—of labeled *parts*, not just a random sea of parameters. The layman should be aware, even absent an understanding of the terms used to label these mechanisms, that the experts do use these shortcuts to refer to the organization and behavior of parts in the ANN.

Effective ANNs provide an enormous variety of trainable parameters, organized in distinguishable mechanisms, enabling them to fulfil Ashby's Law of Requisite Variety. Furthermore, their structure is not unnecessarily complex, additionally qualifying them as good regulators under my reformulation (even if they aren't the simplest optimal regulators!). The parameters in these regulators are transparent, and thus any epistemic opacity in our understanding of them cannot be like the opacity in black boxes which are just defined by inputs and outputs. The kind of opacity we are faced with for ANNs is typical for the study of complex systems, such as the brain.

This deflates claims of epistemic opacity in machine learning with ANNs, and highlights the comparable epistemic situation we have when studying complex computational systems such as the brain. Any opacity we feel in our understanding will be primarily due to our practical inability to follow and assign meaning to the high number of transformations in parameters, and not due to an inherent limitation of understanding any mechanisms involved. Epistemic opacity of ANNs is thus better conceptualized along the lines of Rube Goldberg machines, and not Black Boxes.

4.5 Regularization

The foundations of cybernetics provides an accessible picture of how systems *like* ANNs achieve what they can, and *why* they can. We can now cash out the above argument, and see how the picture provided helps us grasp some techniques which are common in data science when utilizing ANNs. Training an ANN can be thought of in terms of *constraint*. It aims to reduce the variety of outputs flowing from variety in input signals (disturbances, data examples). We feed in, for example, thousands of distinct images of cats and dogs—but to be an effective regulator it needs to squeeze the output into a much smaller set of desired outcomes (two classes). Furthermore, even though the variety in

the inputs might be high under one measurement or perspective, under another there is significant constraint in the organization and features of the images. Otherwise, we wouldn't expect the learning regulator to ever learn how to make effective predictions.

Ashby goes so far as to say that if "something is predictable implies that there exists a constraint." He continues, drawing a parallel to associative learning. He states that "learning is possible only to the extent that the [set of training examples] shows constraint" and "learning is worth while only when the environment shows constraint." (Ashby, 1958, p. 132-134) If the reguland or environment has no constraints (such as random input) we would not expect the regulator to be able to learn how to regulate properly. However, we should not underestimate the ability of a regulator to successfully adapt. A sufficiently robust ANN can even learn to classify images whose labels are *random*, as long as there are some constraints on the distribution of images. That is, even though the back-propagation of error is uninformative, high accuracy (at least on the training set) can still be achieved.

Perhaps more interesting, such networks can even achieve high accuracy on labelled images of random noise! See for example the overview in Zhang et al., 2016. This is a dramatic example of a problem known as *overfitting*. In fact, ANNs are *so good* at regulatory tasks that there are many methods employed by data scientists and machine learning researchers to reduce performance (e.g. accuracy). To understand why, first consider that a set of data examples are typically split into three groups: training, validation, and test data. ANNs learn first on just the training examples, cross checking the updated model agains validation data. If the model is too adaptive, it may learn to regulate *specifically* the training distribution, getting worse on the validation set. Likewise, the model will perform worse when making predictions on the more general distribution which includes the test examples.

One method utilized to increase the robustness and generality of ANNs is called *dropout*. Dropout consists of skipping over nodes (or weights) in a network with some probability during a training epoch. Training occurs on a sub-network. These nodes are then returned for the next training epoch with their values (and dropout may be applied again). The dropout technique helps avoid overfitting when training a neural network model. Intuitively, we might think of this as leading to less overfitting since the network will always need to compensate for a fraction of lost connections, and the connections which survive will be less specialized. Thus, there may be some value in a regulator *forgetting*, and remaining more adaptively flexible.

I have characterized neural networks as cybernetic regulators. What does it mean for a cybernetic regulator to *over*fit? This is a bit of a misnomer under the reconstruction I am offering. Cybernetic regulators simply *fit* their training distribution. From the control theory perspective, when the error is

reduced to zero, by definition the regulator has done what it was supposed to do. (Leigh, 2012, p. 26) An error controlled regulator will—if properly damped and allowed a sufficient strategy profile (measured by the VC dimension)—tend to match the distribution of disturbances with a distribution of actions which *model* the disturbances. This results in a reduction of information (precisely measured Shannon information) 'flowing' through the system, squeezing the state of the reguland into a desired state. If the regulatory model is to be used for a task outside of the training distribution effectively, then they need to be equipped with a method to forget, change structure, or make errors. Dropout is obviously such a method, which is why it is effective at reducing overfitting. If they regulate too well the originally intended regulands, they will poorly adapt to novel regulands.

Another way we can understand why a method like dropout works is by talking again in terms of constraint. There are many ways in which the various parameters in a regulator could be adjusted. There are, however, *fewer* ways they can be adjusted which results in effective regulation of the regulands. Thus, the actions of the regulator are highly constrained if it is to function properly as intended. On the other hand, by *reducing* constraint our resulting regulator will be more flexible. This may not be desired for a training distribution, but if we intend to use the device outside of this distribution it may be better. For an ANN, then, we see that dropout *reduces* the constraint in actions taken to regulate (fit) the regulands. That is, there is more variety in its action than optimal. It would fit better if we didn't use dropout, but we actually don't want the best fitting regulator. More precisely, we don't want the best fitting regulator *for a limited distribution of inputs*. This indeed lines up with many analyses of dropout as a regularization procedure which averages among many sub-networks. At the end of several epochs where dropout occurred, we use the entire network on test data. The entire network can now be considered an ensemble of sub-networks added together.

Another way to mitigate over-fitting (perfect regulation) is by early stopping. During the training period, or the period in which negative feedback informs the regulator, we decide at a certain level of performance to *stop* training even though we know it can still be improved. In convolutional networks training on images, we can also augment the image set by doing transformations like cropping on the images to achieve more robust representations. We could do better, but we don't want to do better, because it would come at a cost of doing worse on the test data.

These are all known as 'regularization' procedures, commonly understood as critical to help prevent overfitting and increase generalization of trained ANN models for test data. However, there is arguably still an explanatory gap since relatively simple networks with more parameters than the training data can overfit even on random labels of data. (Zhang et al., 2016) Explaining

what and how ANNs do what they do is of course one of the primary reasons why people are concerned with epistemic opacity as highlighted by Humphreys. (Humphreys, 2009) By looking to Ashby's cybernetic regulators to understand the kinds of objects ANNs are, we might see his LRV and GRT as the fundamental concepts to understanding the power and limitations of ANNs. They encode the intuition that a high number of parameters may be able to cope with complex tasks, but what is more important is the nature of the model.

4.6 Retaining and Managing Learned Structures

This final section provides a short transition to the topic of the next chapter on transfer learning. I mentioned that Ashby built a mechanical system to illustrate how a cybernetic regulator can be implemented in practice, adjusting its internal structure in response to disturbing stimuli in order to remain stable. Behaviorally, we can understand the system as 'learning' or adapting. This was called a homeostat.

As a follow-up to Ashby's homeostat, his Dispersive and Multi-Stable System (DAMS) intended to scale up the table top experiment with the ability to freeze successful internal adaptations in order that they may not be lost by subsequent changes of the system.

> The idea behind giving DAMS a more complex nature relied on the possibility of letting the machine isolate some of its parts when they would have reached adaptation, and letting the rest of the machine keep hunting for other "friendly" variables. What was to be gained from this is that it would no longer be necessary to re-wire the whole machine to prepare it for the next task. [...]
>
> Truer learning could allegedly be achieved when already reached adaptations were kept and other portions of the organism continued coping. Malapi-Nelson, 2017, p. 165

Such a notion is, I argue, a pre-cursor to the modern practice of transfer learning in ANNs. It also alludes to the potential of a more general regulatory capacity when the machine is equipped with the ability to apply past adaptations. This will be discussed in depth in the next section. Overall, ANNs display very similar behavior to the ideal concepts present in cybernetic regulators outlined by Ashby. However, a second-order framework is required for *managing* what is learned. That is, trained ANN models are each only first-order cybernetic regulators. If they are to be more general, and justify claims of artificial general intelligence, there needs to be regulation among learned models determining which models to apply and when. Something like what scientists

do, when they utilize analogies and transfer models to new domains, is arguably such a second-order regulatory method.

Finally, I just want to briefly suggest an overview of what possible situations there are in terms of knowledge transfer for an effective regulator. This is an area I think for future research into developing the picture of transfer learning I offer in the next chapter. Consider a regulator R, its action on an input in X which results in some output in Y. R takes a $x_i \in X$ to some $y_j \in Y$ (these could be vectors in vector spaces, for example). For another space of inputs and outputs $x'_k \in X'$ and $y'_l \in Y'$, what is the desired action of R, or some other R' constructed by modifying R? Overall, there are eight potential cases we could consider (assuming the format of inputs are compatible):

$$\begin{array}{llll} 1. & R(x_i) = y_j & 5. & R'(x_i) = y_j \\ 2. & R(x'_k) = y_j & 6. & R'(x'_k) = y_j \\ 3. & R(x_i) = y'_l & 7. & R'(x_i) = y'_l \\ 4. & R(x'_k) = y'_l & 8. & R'(x'_k) = y'_l \end{array} \tag{4.3}$$

The first case is what the regulator R is intended to control. That is, R is trained on data X to squeeze into desired states in Y. It is adapted to X and Y. New data from X' are either associated to R, or to another related model R' which is not trained like R is, but constructed from it. In the special case $R' = R$. The questions are then what relationship there is between the data (distributions) X and X', and whether R' squeezes into Y like R, or into some Y'. For present purposes it is more interesting to focus on the cases involving x'_k, since such inputs are novel with respect to the original regulator R. Transfer learning seems to me to be most like cases 6 and 8, where R' is a slightly modified version of R. For example, most weights in hidden layers of R are frozen, but 'adapters' are added to R' on the top and bottom of the network to handle different sized data or a different number of classes. A full discussion of the rest of these cases is an interesting task for future research.

Chapter 5

Transfer Learning and Artificial General Intelligence

> The "designer" of a machine is simply the origin of certain parameter values. Ashby, 2008, p. 4448

This article is an attempt at sketching an approach to artificial general intelligence (AGI) based on transfer learning. The automation of current transfer learning methods have the potential to make significant progress towards robust AGI. I am specifically concerned with the automation of transfer learning methods using artificial neural network (ANN) models, and I will build on the basics of ANNs presented in the previous chapter. The sketch provided here can be taken to provide a rough vision for the kind of engineering tasks involved in creating a meaningful AGI with these methods.

As I take AGI to be largely distinct from any considerations of strong or sentient AI, this present work will be largely devoid of any speculations concerning 'the singularity' or other popular discussions of conscious robots. I will only say that it seems to me that robust AGI (however it is achieved) is certainly necessary, but perhaps not sufficient, for sentient AI. Debates around sentient AI are largely about whether it is in principle possible to achieve, whereas my discussion of robust AGI via automated transfer learning takes it to be definitely graspable as an engineering problem which is possible to be solved. Because of this, there is arguably an imperative for philosophical analysis of the issue. Such analysis is squarely within the domain of a cybernetic systems framework, and I think it is instructive—as previous examples have been—to why a *structural* systems theory approach makes particular sense.

Before getting to the specific methods relevant for AGI, I draw attention to another popular method which is already automated, hyperparameter optimization (HPO). Keeping this in mind should help to contextualize the discussion in subsequent sections. Instead of a data scientist or machine learning researcher manually trying different combinations of learning and dropout rates, for example, automated HPO can simply search through a 'grid' of combinations of a range of hyperparameter values to find (more) optimal values for training a model. We want optimal hyperparameters to improve accuracy per unit of time

(batch or epoch). That is, we use HPO to find parameters to train a model more efficiently (thereby hoping we achieve a *better* model *faster*).

Grid search is just one popular HPO method offered by industrial automation services. Other methods are also used, including forms of Bayesian optimization and genetic algorithms. Examples of automation services are found in Amazon's AWS Sagemaker, `auto-sklearn` (Feurer et al., 2015), IBM's AutoAI, Cray's `crayai`, and other competitors. These services aid in the training and deployment of production level machine learning models. They provide a variety of methods which can be used to aid in a development cycle. Although very practical, HPO doesn't seem like a significant move towards what we might expect from AGI. It just improves the performance of a specific model, one which may not be generalizable in the way we might expect from more robust AGI. Since AGI is the present focus of this work, I want to consider the automation of a method which at its core is about generalizing models, and managing the use of previously trained models.

General intelligence (GI) is surely related in some way to the adaptive capability of a (cognitive) system to successfully respond to inputs which, at least from a practical point of view, are reasonably considered to be novel compared to typical past inputs. For *artificial* GI, in the context of ANNs, this means the inputs are reasonably considered to be outside the training domain. That is, the novel data comes from a different distribution compared to the distribution the ANN model was trained to handle. The system utilizes previous experience (however unrelated) to successfully handle the novel input. To this end, I now introduce more concretely what is called *transfer learning* in ANNs. In contemporary machine learning, ANNs are motivated by connectionist cognitive science and neurobiology. Connectionist motivation for studying transfer learning is outlined by the Sharkeys:

> (i), [...] it makes little sense from a cognitive science perspective to have neural nets that are trained from scratch on every new task, and that are unable to draw on any prior knowledge;
>
> (ii), so far as neural nets are used for psychological modelling, training nets that are in effect *tabula rasa* flies in the face of the mass of evidence of the important role of innate structure in the brain.;
>
> (iii), it is simply not practical to have to create and train an entirely new net as each fresh problem is approached. (Sharkey and Sharkey, 1993, p. 314)

Transfer learning is typically considered to be the use of pre-trained ANNs for novel tasks considered to be outside of the original training domain. Specifically, this usually means re-using a set of weights (and associated ANN architecture) which have been trained on one data distribution D_1 to make (or learn

to make) predictions on data from another distribution D_2.[1] The method has become common in data science, mainly due to the impracticality of always training a new network. Crucially, it is the human data scientist who decides the appropriateness of transferring the model, or of the similarity between domains D_1 and D_2. The majority of parameters (weights) in an appropriate pre-trained network can also be 'frozen', greatly increasing the efficiency of learning the new task. The reader may wonder how this relates to the familiar paradigm of supervised learning.

Transfer learning is both technically and practically distinct from supervised learning, although supervised learning could be viewed as a special case of transfer learning. Consider a mechanism M_1 which generates data X_1 according to a distribution D_1. Supervised learning models trained on X_1 arguably expect new data (the data for which the model was trained to deal with) to be interpretable within the same distribution D_1. That is, we consider the training data X_1 to be a representative subset of the data generated by M_1, exhibiting learnable properties of the distribution D_1. We don't typically assume the model will successfully apply (generalize) to some data X_2 generated by some other mechanism M_2, which may exhibit properties of a distribution D_2 which diverges from D_1. By definition, transfer learning applies a model trained on X_1 (generated by M_1) to some novel data which comes from a data set X_2 generated by some other mechanism M_2. Here we can see that if $X_1 \cup X_2$ are consistent with D_1 and can be interpreted as generated just by M_1, we are assuming a special case. Transfer learning is a technique which requires making explicit our assumption about the relationship of a model to novel data, the distribution it comes from, and the mechanism which produces it.[2]

While one could quibble in principle about these definitions and the boundaries of certain cases, in the end transfer learning is *practically* distinct from conventional modeling because it is used under the assumption that certain capacities of the model are being transferred to a new context. If it isn't always assumed outright that the data comes from a different distribution, it is arguably so for the interesting cases. These are the cases I am most concerned with, and are clearly distinct from the special case of supervised learning. It seems extremely unreasonable to suppose, for example, that transfer learning on an image problem with a model trained on ImageNet is effectively applying a model trained to deal with the ideal distribution of images *in general*—i.e. a distribution of all images.[3] Sure, some layers in a convolutional network (e.g. edge detecting filters) may apply generally to all images, however they may be

[1] Transfer can also be achieved in other ways which do not necessarily involve re-using the exact network architecture or weights. See for example Pratt, Mostow, and Kamm (1991).

[2] Yet another reason why explicit representations in the scientific methodology are necessary.

[3] Suggesting that this is what an ImageNet model is learning is ambitious, and would seem a little like changing the goal posts if it was used to define away these kinds of transfer

images at completely different scales, generated by different mechanisms, with completely different color or feature distributions.

If successful on a new task, the ability to use a pre-trained network is more efficient than re-initializing and learning the task from scratch. It also means that a model, or meaningful parts of it, may be able to apply to increasingly diverse domains. While we will never have just one model for every use case, being able to adaptively apply models to new use cases is seen as a feature of general intelligence. Transfer learning techniques may be expected to support artificial intelligence systems for increasingly more general tasks in the near future if automated properly. That is, the artificial system itself must make judgements about when to apply pre-trained models like we do, and be able to compare problem domains. The artificial system needs an inter-domain reasoning engine, a way to manage robust transfer learning.

The purpose of this article is to philosophically situate these claims, and evaluate them in an associationist framework for cognition. I begin in section 5.1 by briefly sketching an associationist skeleton, which is fleshed out by literature on analogical reasoning in section 5.2. In section 5.3 I outline the basics of an interesting recent example of transfer learning, and then discuss whether claims of AGI resting on the promise of near-future transfer capabilities are justified. I argue that, from an associationist point of view, robust transfer learning capabilities would indeed provide the kind of information-processing mechanism necessary for what we expect from AGI. As alluded to, this would require the automation of several management tasks which are currently done manually by data scientists and machine learning researchers. In the end, I suggest that we are justified to have the expectation of increasingly widespread AGI in the near future.

Flipping the script, in section 5.4 I discuss how machine implementations of transfer learning might inform an account of inter-domain reasoning. The traditionally nebulous notion of analogical reasoning, for example, might benefit from a concrete machine view based on transfer learning. A useful sketch of how to distinguish good from bad analogies is achieved by utilizing the notion of *positive* and *negative* transfer. Overall, there is certainly interest in ways of facilitating *robust* knowledge transfer in both scientific reasoning and artificial intelligence. In section 5.5 I suggest 'general systems theory' (GST) as an already existing framework that helps provide a methodology for robust transfer.

As a methodology for transferring knowledge to new domains, GST can be viewed as a 'logic' of discovery—a conceptual (and formal) account of inductive adaptation. I wrap up in section 5.6 by briefly summarizing how I view the relationship between the philosophy of discovery in science (for example in

learning cases. A 'superdistribution' could still arguably be considered distinct from the training distribution.

Simon (1973)), computational cognitive science, and connectionist artificial intelligence. I appeal to a previously discussed relationship outlined by Thagard (1982), and conclude that a philosophical framework which incorporates and facilitates knowledge transfer (such as GST) may be of use for the development of AGI capabilities built on transfer learning, completing a theoretical triangle for engineering AGI systems.

5.1 Computational Associationism

As a framework for understanding concepts and human cognition, associationism has been around in various forms from Aristotle to Locke.[4] Associationism, for our purposes, is the assumption that a cognitive system has the capacity to 'associate', or group together, different experiences relevant for cognitive tasks. It is the capacity to draw connections or comparisons between different perceptions, situations, or problems. Most importantly, we are here concerned with the association of new perceptions, or new tasks, with old experiences. It does not need to be perfect, or an in principle mechanism that draws *the* correct connection. A pragmatic sense of association is just that it is an effective mechanism. The language of 'experiences', 'perceptions', 'concepts', and so forth, should be filled out according to whatever framework of cognition is being used. What is important is that we suppose some computational framework, consisting of inputs, functions, and outputs. Additionally, some form of memory and a retrieval mechanism must be assumed—if perception and the immediate processing of perceptual inputs does not directly perform associations.

For example, we first begin with some initialized state S of the receptive perceptual system of an organism, and an input to the system I. The output, the response of the system to stimulus, is related according to a function $O = f(S, I)$. A further assumption for the associationist, however, is that an adaptive intelligent system survives in part due to the not wholly incorrect mapping between perceptual structures and the environmental input. That is, it is a reasonable assumption that the *ways* in which a successful system perceives does not contradict the processing of perceptual information relevant to survival. For example, the system would be compromised by introducing a random gate between the perception of some data and the processing of that data with respect to some goal. Conversely, we would expect (as noted by Bertalanffy (1969, p. 239-242), among others) that systems which successfully reproduce and are naturally selected process information according to a model that has

[4] I am not concerned with an in depth treatment of the idea here, but mention what I take to be the core of associationism so as to situate the potential importance of the machine intelligence concepts to follow. For a more comprehensive overview of associationist literature, see Mandelbaum (2017). Also Cordeschi (2002, §6.4) provides more context on the history of connectionism and low level associationist ideas.

relevant similarities with the structure and distribution of the environmental inputs.

Computational associationism supposes that a computational system which is intelligent effectively performs some comparative task on inputs. That is to say, the system is capable of some associative function as the output $O_A = f(S, I, I')$, taking as arguments not only the current state and perceptual input, but some previously learned (or experienced) information I'. Notably, I and I' are related through some relation of relevance $R(I, I')$. Such a relation might range from some measure of similarity on features, a structural relationship shared (or mapped) between I and I', or a statement regarding a class or type which I and I' both belong.

Associationism embodies the capacity for the comparative perception of novel objects (and problems or situations) with what has been previously encountered. That is, the capacity to *associate*. One might think of a sort of cognitive 'bootstrapping' which is enabled by the ability to transfer knowledge or actions from previous impressions. We may formulate these claims in computational level terms, according to the levels outlined in Marr (1982) and Marr and Poggio (1976). This level of analysis is concerned with what an associative system should compute and why. An assumption about associationist problem solving is that the system expects to have a lower cost using a solution that was effective on a previous problem, which is *associated* in some relevant respects with the problem at hand. It is assumed that the system solves the problem more efficiently by effectively associating it with some previous experience (or with some problem-solving structure).

On the other hand, if perception is filtered through adapted networks in the first place then this procedure isn't so much bootstrapping as it is a necessary consequence of the efficient re-use of biased information processing systems.[5] In other words, at the implementation level the association can already be encoded into the mechanisms which process the input information. A state S might encode information about I', so call it S', then we can say it is some function of I', $S' = g(I')$, and so the association is formalized as some $O_A = f(S', I)$.

The system should associate to some relevant structure, according to this view, because non-relevant solutions will be more costly. These costs may be incurred due to decreased efficiency, for example, or failure to solve the problem (when it could have been solved by the successful implementation of an associated solution). It is presumed, therefore, that non-association is strictly worse on the whole. This very rough and brief outline of what I take to be an associationist picture will suffice for what follows. When using analogies to cash out the association, previous successful solutions are analogous or similarly structured.

[5]This picture should also not be taken to be limited to a perceptual system, but for present purposes thinking of perception along these lines is sufficient.

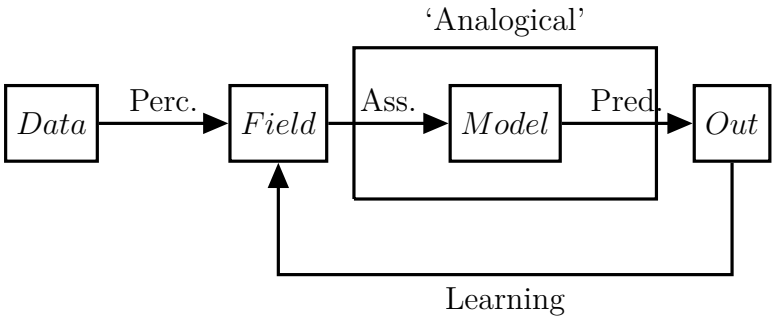


FIGURE 5.1: What has been referred to as analogical cognition may variously include parts of an associative mechanism, a cognitive system's internal model, and the application of that model for predictions and higher-order reasoning.

5.2 Analogy in Cognition

Analogical relations are perhaps the most notorious associative relations. Analogy in the context of reasoning encompasses a wide range of claims. For example, claims concerning perceptions and judgments of similarity, about cognitive and computational tasks, and also claims regarding the ability to transfer situational and methodological solutions to new (i.e. novel and previously unencountered) problems. Analogical reasoning, from an associationist perspective, can be seen as fundamental to human cognition since it is one way to cash out the associative function. That is, an analogical relation is one candidate for the relation $R(I, I')$ mentioned above.

In this present section I focus on some particular claims concerning the role of analogy in human cognition, in order to contextualize the arguments regarding transfer learning in following sections. These cannot be representative of all claims regarding analogy, but are a selection relevant for illustrating how analogy is one way to cash out associationism. Holyoak and Thagard (1997, p. 35) state that "The analogical mind is simply the mind of a normal human being." More elaborately, Hofstadter (2001, p. 499) waxes poetically about the centrality of analogical reasoning in the "broad blue sky" of cognition. Analogy is "the very blue that fills the whole sky of cognition".

He continues, and we can see the close relationship that analogical reasoning has to perception in an associationist framework for cognitive science:

> The triggering of prior mental categories by some kind of input—whether sensory or more abstract—is, I insist, an act of analogy-making. Why is this? Because whenever a set of incoming stimuli activates one or more mental categories, some amount of slippage must occur (no instance of a category ever being precisely identical to a prior instance). Categories are quintessentially fluid entities;

> they adapt to a set of incoming stimuli and try to align themselves with it. The process of inexact matching between prior categories and new things being perceived (whether those "things" are physical objects or bite-size events or grand sagas) is analogy-making par-excellence. How could anyone deny this? After all, it is the mental mapping onto each other of two entities—one old and sound asleep in the recesses of long-term memory, the other new and gaily dancing on the mind's center stage—that in fact differ from each other in a myriad of ways. Hofstadter, 2001, p. 503-504

These statements express a view that analogy, as some kind associative mechanism, is fundamental to human cognition. Analogy making is thought of not as a high-level cognitive task different in kind to perception, but rather analogy-making and perception are fundamentally intertwined. This serves to show why we are interested in machine protocols that can implement something along these lines, 'learning' in some important sense from previous experience. Carbonell (1983, p. 143) begins to provide an account of problem solving and learning that relies on previous experience. A solution method for a previously solved problem is chosen based on the similarity to the current problem.[6]

It is worth drawing attention also to the structure-mapping account of analogical reasoning from Gentner (1983). For Gentner, analogies are structural mappings between two domains (sets of features or properties and relations between them), and are found on a spectrum between mere similarity mappings and abstract generalities. (Gentner, 1983, p. 161) Whereas similarity mappings are at the object level, analogies map the *relational* structure between the objects. For a simple example, let domain X consist of a set of features $F_X = \{a, b, c\}$ and relations $R_X = \{R_{ab}, R_{ac}\}$ where the relations are spelled out like $R_{ab} = R(a, b)$. Domain Y consists of features $F_Y = \{d, e, f, g\}$ and relations $R_Y = \{R_{def}, R_{dg}, R_{fgd}, R_{fe}\}$. A similarity mapping will be a map $F_X \to F_Y$, whereas an analogy mapping will be from $R_X \to R_Y$. This mapping can then be used as a source of information, such as to assert a credence in a theory for the target system or to infer a similar property or relation also exists or holds in the target domain since it exists in the similar source domain.

For our computational level framework, analogy as explicated above must reduce some relevant cost for the computational system. Cost reduction is a weaker goal than cost minimization, but is appropriate for types of model-based reasoning which might occasionally result in poor transfer to a new problem. The effective reduction of some costs, such as reducing time for a computation, might result in less accurate computations but payoff in the long run. A

[6] A transformation in solution space from the familiar solution to a novel solution for the new problem is taken to provide an account of analogical reasoning in problem solving. The initial state in the solution space is determined by a reminding protocol, which is essentially a search for the most similar problem through a database of previously solved problems.

comparable reduction means that the class of actions prescribed by analogical cognition is supposedly preferable (i.e. expects a lower cost) to actions or solutions not rooted in an analogy (or similar associationist) protocol.

While the structure-mapping account of analogical cognition may be understood in symbolic terms, I take structure mapping to be consistent also with a connectionist picture. The structure of an artificial neural network, for example its trained weights, can also be usefully mapped to new problems in the same way that analogical mapping occurs. These weights encode relational structure.

5.3 Transfer Learning

While transfer learning may be a general concept in cognition, I am mainly concerned in this paper with the implementation of the concept in machine learning—particularly with artificial neural networks. Transfer learning protocols utilize a learned neural network as a model for application to another learning problem. This has also been called *adaptive generalization* by Sharkey and Sharkey (1993, p. 314), where the main question is "understanding *when* knowledge can be transferred between nets: identifying the circumstances under which pre-training on one task will assist (or interfere) with the performance of a subsequent task.".

In some of the first publications concerning transfer learning, transfer in neural networks is conceptualized as using the learned weights of a network at the outset—either directly or by using them as modulators to create other weights. See for example Pratt, Mostow, and Kamm (1991) and Pratt (1993). Weights from a hidden layer to the output layer have also been shown to transfer classification capacity by Sharkey and Sharkey (1993, p. 322). Modern deep ANNs have abundant hidden-to-hidden layer weights as well, and these are generally what are pre-trained in transfer learning techniques. The input and output layers can be adjusted, but the largest number of transferred parameters are in the hidden layers.

Currently, an entry level data scientist does not need to train a new ANN for every task she encounters. With `keras`, a deep learning library for the Python programming language created by Chollet (2015), there are pre-trained deep networks (*models*) available to be downloaded and used. Similarly with `pytorch`. These models are essentially sets of weights of certain dimensions, which have been trained on paradigmatic sets of data. As an example, for image recognition, there are a number of models to pick off the shelf which have been trained on the ImageNet dataset from Deng et al. (2009). This is a large dataset of images, and the models already have 'experience' classifying the images in this set. These pre-trained models are useful for experimenting on new image data sets, and for when the number of data samples available on the new classification task are small. Transfer learning is also making progress in other areas, such as

natural language processing (NLP) with LSTMs. Notably Howard and Ruder (2018) recently introduced a method called Universal Language Model Fine-tuning which "enables robust inductive transfer learning for any NLP task, akin to fine-tuning ImageNet models". The field is evolving rapidly, and perhaps future methods will involve some automated code writing from more generally capable models as well. For example, Microsoft is using exclusive access to the GPT-3 model developed by OpenAI (Brown et al., 2020) to translate natural language into code.

It takes a lot of time and energy to train ANNs, and by using pre-trained models the potential efficiency gain for solving new machine learning tasks is large. This is crucial, as it begins to close the gap with the ability of humans to adaptively generalize from very small sample sets. Transfer learning techniques are becoming more robust, as data scientists learn to apply pre-trained models in more diverse ways to more and more cases. So, the appeal of robust transfer learning includes not just to be able to transfer from a previous context to a new context, but in certain cases to be able to learn *faster* and *more efficiently* than by training a new model from scratch. If the pre-trained weights are frozen, this efficiency is largely due to a significant reduction in parameters which are 'hot' and need to be trained.

A measure of transfer can be positive or negative. Positive transfer occurs when the protocol is more efficient or accurate than an alternative protocol. The biased structure of the network encodes a model, and we expect positive transfer because this model is relevant to the structure of the input information. Negative transfer occurs when the transfer protocol makes learning less efficient, less accurate, or take longer to learn a new task. To be clear, I take negative transfer to be actively worse than useless transfer or indifference on a new task. It should be thought of as really anti-useful or impeding performance compared to a baseline (i.e. random initiation or training a separate network from scratch).[7] It is desirable to find ways to enable positive transfer learning. I will focus for the remainder on attempts which utilize pre-trained (deep) neural networks. That is, the weights or connections have been previously learned or pruned in some procedure resulting in non-random weights or non-uniform connections. Positive transfer via pre-trained neural network models can be considered as *successful* adaptive generalization. Otherwise, the network may result in negative transfer as it is biased for certain tasks and against others. (Sharkey and Sharkey, 1993, p. 326) How has connectionist transfer fared in recent work?

One interesting example, PathNet from Fernando et al. (2017), learns

[7] If our performance metrics include time, then a 'useless' transfer would most likely imply actual negative costs, since we wasted time training on a task. Unless of course we needed to train the model anyways, or were efficiently trying to reuse an old pre-trained model (perhaps we have a bag of old trained models at the ready).

weights in a deep ANN playing the Atari game *Pong*, and then 'freezes' important parts of the network for the new task of playing the Atari game *Alien*.

> The parameters contained in the optimal path evolved on the first task are fixed, and all other parameters are reset to their random initial values. Fernando et al., 2017

Keeping the learned structure from the previous task, the researchers achieved *positive* transfer learning. This has been done with deep convolutional neural networks which learn image filters (convolutions), and take images (frames of the Atari game display) as data types. In this case, the transfer was measured against the same network's ability to learn *Alien* from scratch.[8] While learning both games (*Pong* then *Alien*) was longer than just learning *Alien*, comparing the time it took to learn *Alien* after learning *Pong* shows a positive transfer effect. In other words, this case illustrates a *positive* transfer effect. It may also be argued to exemplify the efficient re-use of an ANN.

It is important to emphasize that the domain comparison is being done by the data scientist from the outside. That is, we could say there is a tutor or teacher telling the network what its pre-trained structure might positively transfer to. Even though the authors have automated one procedure for setting the structure of the network for the next task, the hard part is to figure out how to automate (with reasonable success) which problems to apply the network to. That is, we don't just want the *Pong*-network to be used for *Alien*, and other handpicked Atari games, but also other similar problems that we do not anticipate (or do not want to supervise).

So, the hard work has already been done. The domain comparison is put in by hand, we are externally plugging in a problem that *we* view as similar. The scientist is feeding to the network a problem which stands in an analogical or similarity relation to the previous problem. Also, our matrix and vector dimensions are compatible, and the data types (such as pixel colors) are commensurable. Both data sets are of the Atari type. Such compatible data formats are a very special case. Future transfer learning, if it is to fulfill the expectations of artificial general intelligence, will need to robustly automate domain comparisons and data formatting from various domains. To some extent, our evaluation of transfer learning capacities (and, I argue, of AGI capacities) will depend on our own perception of the 'distance' between the source and target domains. It will be debatable whether this case is really inter-domain transfer, or just *intra*-domain. This is likely to be a chronic issue of interpretation, while the steady march of incremental progress continues.

Furthermore, we will not have a *single* network be capable of meaningfully successful results on a wide variety of inter-domain tasks. In other words, there

[8] Or more precisely, the same configuration of a network.

is no free lunch: performance on one class of problems will necessarily come at a cost of performance on some other class of problems. This intuitive result is a gross oversimplification of the widely-cited formal no free lunch theorems by Wolpert and Macready (1997). Nonetheless, this oversimplification is sufficient for our discussion. A more likely route would be to have a variety of pre-learned networks and a way to manage which one to apply in a given circumstance. This problem I will call the hard problem of AGI with transfer learning, and it may also be similar to what Korb (2004, p. 435) calls the "meta-learning problem". Tackling this management problem requires a robust 'structure mapping engine', which has arguably been the focus of computational approaches to analogy. See for example Falkenhainer, Forbus, and Gentner (1989) for an influential early work in the field, computationally implementing the structural mapping ideas of Gentner. The work of Pickett and Aha (2013a) and Pickett and Aha (2013b), for example, is a more recent attempt in the field (although not for managing pre-trained ANN models). The spontaneous discovery of structural relations seems to me to be on the right track towards a transfer management engine.

The recent work of Lu, Wu, and Holyoak (2019) may be viewed as indicative of a trend to bridge the gap between the old computational view of analogies and a more modern view based on deep learning methods. In some ways we might see this as part of the larger trend away from symbolic AI towards sub-symbolic or connectionist AI. However, for my present purposes, works like these don't acknowledge the inherent structure mapping which is going on at a 'low level' in ANN models. With this in mind, it becomes clear that automated transfer learning will be a powerful structure mapping engine. An AGI based on modern ANN models must learn and be trained at the management level, that is, it must learn to effectively transfer models and apply them successfully. Currently, human data scientists and researchers are the managers of various models, and when to apply them.

Significant problems, such as the incommensurability of representations (or differences in feature dimensions), must be adjusted for and made similar—and so we are looking for a framework which is built to handle these kinds of problems. Making representations structurally similar, or abstracting away from irrelevant features, will need to be outlined and automated. I will consider in the final sections a theoretical framework for reasoning in science called *general system theory* (GST), which might just fit the bill. I argue that the methodology of GST is at its heart assuming a structure mapping engine, and provides us at least an intuition of what such an engine in a modern machine learning context must satisfy for AGI.

As a final comment, we would like to understand not only the practical limitations of when a transfer can be applied, but in general about whether a particular transfer *should* be allowed to occur. This is a rather involved

practical problem to solve, with some progress already being made, but it does lead to some ethical questions about AGI. Is the machine intelligence going to make a bad classification or decision based on inappropriate transfer? Will it inadvertently damage the environment or harm humans? These worries are not due to the intelligence of the machine, but to the *ignorance* of the machine concerning appropriate transfer. I will argue later that GST also helps provide some guidance on this issue.

5.3.1 Artificial General Intelligence

We are not yet at the stage of having robust artificial general intelligence (AGI), but data scientists increasingly generalize results and models in machine learning. The current state of the art still involves the interactions between a data scientist, her perceptions of a machine learning problem, and automated machine learning techniques. There are still many aspects which need to be manually done by the data scientist. For example, in a transfer learning problem, this involves the perception of similarity between the domain a pre-learned network was intended to perform on (i.e. the distribution of data that the network was trained on), and the domain for a problem. AGI, on the other hand, is the *automation* of those aspects which are currently manually done. For an AGI to be based on transfer learning techniques, perceptions of similarity (as well as other manual tasks, like formatting the data or adjusting network dimensions) need to be automatically done by the artificial system.

So, we hope to automate the role data scientists are currently playing. As our techniques for transfer learning become more robust and sophisticated, we are clarifying what exactly needs to be automated for an AGI system. Keeping in mind, of course, that there will never be a *single* network useful for all problems (there is no free lunch), the automated and robust use of a bag of networks seems like a plausible expectation for the near-future capabilities of AGI systems.

Can we achieve AGI without transfer learning? It might be possible, but I will argue that robust automation of transfer learning is sufficient to achieve reasonably impressive AGI. If we grant some form of associationism in human cognition, as I have outlined earlier, we would have justification for expecting that a machine learning environment which can achieve some means of association and transfer would be more powerful than a machine learning environment without such capabilities. In this case, more robust AGI is just what we would expect from automating robust transfer learning.

As a note, a general machine intelligence needs the ability to generalize from previous problem encounters and solutions to new ones, but it does not need to generalize to *any* potential future problem. Only the next one, or the next one in a domain of perceived similarity. Future work might find additional progress

in furthering this research program by utilizing *transductive* machine learning techniques, as mentioned in Harman and Kulkarni (2007).

This outlines what I take to be a minimum account of what a machine protocol must satisfy to be considered a candidate for AGI. What should we expect from AGI in the near future? Are claims about AGI stemming from transfer learning research justified? To be more precise: should we expect that robust AGI can be achieved by focusing on developing more and more robust transfer learning techniques? I answer this question in the affirmative, for the reason that transfer learning is a way to cash out what it *means* for a machine intelligence to be *general*. However, what is interesting for philosophers, data scientists, and AI theorists, is the explanation for *why* it makes sense that more robust transfer learning capabilities will lead to AGI. The reason associationists can offer is that transfer learning is what human cognitive systems do at some level. Additionally, there are lessons I think we can take away for a philosophical account of analogical reasoning.

5.4 Machine Analogies

Transfer learning via pre-trained artificial neural networks might be a way for philosophers to get a handle on forms of inter-domain reasoning like analogy. From the literature discussed earlier, I sketched a picture of analogies which is closely linked with the notion of structural transfer. It is perhaps the most intuitive way to visualize what transfer learning can do. Analogies are, in a sense, the prototypical way we transfer knowledge from one domain to another. Turning this on its head, we can take the story of transfer learning in ANNs to inform a philosophical account of analogy.

I argue that machine protocols doing something arguably *like* analogical reasoning, namely inter-domain transfer learning in ANNs, can inform the expectations of traditional normative and explanatory goals in philosophical analyses of analogical reasoning. There have historically been several attempts at implementing programs that are inspired by the cognitive theory of analogical reasoning. That is, they aim to implement structure mapping (*ala* Gentner) or analogical discovery. For a very useful overview, see Bartha (2013, §3.4-3.5), and also the recent work of Pickett and Aha (2013a) and Pickett and Aha (2013b). We can turn these models from being computational studies aimed at informing us about analogical cognition, and as models of interest for artificial intelligence, to models that can inform a philosophical account of analogical reasoning. Similarly, transfer learning in ANNs can do the same (they are arguably not even wholly distinct from these previous works). Paul Bartha recognizes that a schism is already tangible between computational and classical philosophical analyses:

> One concern is that there appears to be no way to extract any specific norms of analogical reasoning from the algorithms and coding conventions. [...] But computational models are meant to challenge traditional epistemological objectives. Efforts to develop a quasi-logical theory of analogical reasoning, it might be argued, have failed. In place of faulty inference schemes such as those described earlier, computational models substitute procedures that should be judged on their performance rather than on traditional philosophical standards. Bartha, 2013

He seems optimistic that some philosophical theory of analogical reasoning might be born out in a computational approach. I maintain the view he alludes to, that a machine account can justifiably forego providing (or assuming) some normative account of good and bad analogies. Rather, we can provide a means of distinguishing good analogies from bad ones by measuring the degree of transfer that the inter-domain relation enables. In other words, measuring analogies by their performance in some formal sense. Of course, this is easier said than done. It is unlikely that such a measure would be formally feasible for the high-level reasoning Bartha is concerned with, yet we can still tell a story (and construct models) populated with characters from the picture outlined here. Maybe good (high-level) analogies, such as in scientific reasoning, are ones which reduced some relevant costs in solving a new problem. Perhaps they enabled science to proceed efficiently, without frustration, or the alternative was not to 'solve' the problem at hand.

This is arguably a natural impulse if one is thinking of analogy as shorthand for an associative cognitive mechanism akin to perception. Such a contention is explicitly appealed to in a more recent computational approach to modeling analogical cognition in Pickett and Aha (2013a) and Pickett and Aha (2013b), with the goal of spontaneously recognizing structural isomorphisms (defining an analogy) in a data set in a manner more efficient than a search over the concept space—because of an assumption that the brain associates analogical structures efficiently. The authors achieve this by "representing relational structures as feature bags, [...reducing] the problems of analogy to problems of surface similarity."

The discussion of spontaneous analogy should be kept distinct from the case of a *granted* analogy. That is, the problem of finding a structure to map is different from the procedure of using an established mapping for some purposes (like inferring a new structural property holds in the target domain). The first sense is something like a discovery, while the second is more like an induction where both source and target domains are already known to some degree (and presumed to be related structurally). Both of these aspects play a role in computational models of analogy, but as Pickett and Aha (2013a) note the "more difficult problem is finding the analogs to begin with."

These computational approaches are different from the pursuit of an analogical classification scheme *a priori*. Perhaps at a later point there will be a rejoinder between the *classical* philosophical view and the *machine* view, however this does not concern me here—as Bartha says there is "room for both computational and traditional philosophical models of analogical reasoning".[9] At least for now. I will return to this discussion in the next section. In an important sense, the discussion of norms we threw out earlier for a philosophical account of analogy comes into play when we are faced with the task of regulating transfer. It is still not needed to define or make sense of transfer from the beginning.

For better or worse, humans make 'bad' analogies all the time. Thus, a philosopher might not be satisfied with computational models of analogy inspired by 'imperfect' human cognition, and want to categorize proper from improper analogies. Nonetheless, 'good' and 'bad' might instead be learned through interaction, and through negative responses from the environment. That is, analogies might be characterized by how they model the environment—and the relations between perceptual systems and environmental input. We might then judge them by how they impact performance on classes of transfer learning tasks. For example, a bad analogy might enable a bad prediction—resulting in a large distance between what is expected and what is eventually observed. Alternatively, a good analogy might reduce such a measure.

Thus, what we care about is the *effectiveness* of the analogical reasoning procedure. 'Good' analogies are ones that result in *positive* transfer learning, and 'bad' analogies result in *negative* transfer learning. Positive and negative will, of course, vary with the ways we measure costs, the computational model, and with the choice of a target problem. Traditional philosophical measures may classify certain effective (and positive) cases of transfer learning as bad analogies even though in the situation, the agent might have effectively reduced some relevant cost such as time. Another cost reduction we see in examples of transfer learning in ANNs is in the number of samples needed to train (learn) a new problem.

Consider the following statement in a recent machine learning textbook, noting the ability of humans to quickly generalize and transfer with small training sets:

> Humans do not require tens of thousands of images of a truck, to learn that it is a truck. [...] This suggests that humans have much better ability to generalize to new settings as compared to artificial neural networks. [...] In other words, humans are masters of transfer learning both within and across generations. Aggarwal, 2018, p. 453-454

[9] For further details on perhaps the most up to date philosophical work on analogy, the interested reader may wish to take a look at Bartha (2010) as well.

The author supports this contention with an evolutionary argument about the evolved (*pre-trained*) structure of the neural networks which we suppose are present in the brain. Thus, in a certain sense, AGI via transfer learning extends a long-standing philosophical claim about cognition and human intelligence into the frontier of machine learning. However, as I have noted, robust transfer learning includes much more than the transfer itself. It also includes the ability to curate, format, and interpret a wide range of data—and the selection of an appropriate structure to map onto the problem.

This is all not to say that there can be no normative component in an account of analogy. For example, I think we can agree with Carbonell (1983, p. 148) on a basic norm that *any* machine account of analogy should satisfy. He talks about an experimenter placing bananas out of reach, while the subject (a monkey) is allowed to watch the bananas being placed. The experimenter places the bananas out of reach with the help of a stool, and then removes the stool. A general norm in this case is something like "...a 'smart monkey' ought to learn from his observations ..." and follow the observed manner in which the bananas were originally placed as a solution to retrieve them. In Carbonell's example, if the monkey does not learn from its observations of an experimenter, we can make a strong argument that it will be less likely to solve the problem of obtaining bananas which the experimenter has placed out of reach. The relevant version of such a norm will be, for machine accounts of analogy, something along the associationist lines I have outlined.

To reiterate, this doesn't necessarily tell us anything about good or bad analogies. We would have to formulate the statement with regard to a cost in order to say something about bad analogies. It is simply a normative condition concerning information processing in the most general sense.

5.5 A Methodology for Robust Transfer

Above, I have alluded to some of the extra bits and pieces that are required for *robust* transfer learning to be automated. Since it is achieving the robustness that will provide the largest hurdle for technical implementation and software development, some further discussion is warranted. Additionally, we may want to guard against bad transfer, and recover some of the normative intuitions that philosophers have.

Is there a rigorous enough way to rule bad transfer out such that a machine protocol could also achieve normative classifications of domain comparisons and transfer? Does there already exist a methodology which might provide a framework for solving these problems for robust transfer learning? I suggest that general systems theory (GST), as outlined and advocated for by Bertalanffy (1950) and Bertalanffy (1969), provides us with a framework useful for

discussing robust transfer learning. GST is motivated in part by a call for scientific generalists after noting the inadequacy of reductionism in biological and social sciences, and the failures of specialists as noted by Bode et al. (1949). GST aims to facilitate general modeling practices in science by focusing on relevant similarities between different systems in different fields. The methodology is intended to guard against improper transfer between domains:

> [G]eneral system theory can serve as a regulatory device to distinguish analogies and homologies, meaningless similarities and meaningful transfer of models.
>
> The homology of system characteristics does not imply reduction of one realm to another and lower one. But neither is it mere metaphor or analogy; rather, it is a formal correspondence founded in reality inasmuch as it can be considered as constituted of "systems" of whatever kind. Bertalanffy, 1969, p. 85

His distinction between analogy and homology stems from biology, where an analogous organ in an organism has a similar function to another organ (in another organism) but is unrelated in a structural or causal way. In this case, a shared structure could be a shared evolutionary path. A homologous organ has more structural similarities to another organism's organ due to shared evolutionary ancestors, but may differ in function. The main point for our purposes is that this distinction focuses on managing transfer of relational structures: being conscious of whether or not a transfer is justified.

GST may help in regulating analogy-use in science. Perhaps more interesting, can it shed light on the proper role of transfer in artificial intelligence systems? Since GST is a framework that emphasizes structural knowledge transfer between domains, it must guard against improper transfer or it is at risk of becoming useless. It also must work on ways of abstracting, idealizing, and translating terms and concepts in one domain in order to be comparable with terms and concepts in other domains of science. These are remarkably similar to what we would want for robust transfer learning and AGI. That is, we would ideally want non-trivial comparisons between problem domains, as well as the ability to make non-uniform data types compatible, and models which are successful across tasks.

While "meaningless similarities" may not be an issue for an account of analogy based on performance (we might just imagine a poor measure of transfer, or no increase in problem-solving ability), they do seem to present an issue for implementing AGI by transfer learning techniques. Thus, the norms for my present account become significant *a posteriori* when considering the practical implications of a concrete methodology. The reason we look for the norms are

for harm reduction. They are, however, unnecessary for the theory of (analogical) structure mapping—only to deal with the consequences of an AGI system that is based on transfer learning.

GST provides a philosophy of science framework explicitly suited to analyze knowledge transfer between systems and, therefore, between modeling domains. There is a focus on expressing structural properties common between systems. As such it allows inter-theoretical transfer of knowledge for a scientist, modeler, or philosopher. A practicing systems theorist may take the knowledge gained from one system (i.e. a method of problem solving, or an idealized mechanism) and apply the method or mechanism to another (typically less well-understood) system. The GST framework thus has baked into it a robust method of inter-theoretic (or inter-domain) knowledge transfer.

As a final comment, it should be noted that there are two typical means for applying the methodology of GST. The first is a top-down approach, starting with a very general formalism which captures a particular class of systems. Alternatively, when the class of systems is not well known, one can begin with a case-by-case method and build up a system theory which generalizes as the cases are addressed. As a more recent example, I take the approach advocated for in Goldstone and Wilensky (2008) to illustrate the benefits of a case-by-case method of general systems theory promoting knowledge transfer in the learning process (particularly utilizing agent based modeling).[10]

> We have proposed transferring complex systems knowledge by having students rig up their perceptual systems to perceive situations in a manner informed by a provided or constructed rule set and then simply "leave this rigging in place" when presented with new situations. Goldstone and Wilensky, 2008, p. 505

I am thus an advocate for a case-by-case approach to general systems science rather than a top-down universal systems approach for two reasons. First, the approach is arguably more relevant for unsolved problems and anomalies in science as well as descriptive of certain important cases in modern systems science. Second, the knowledge transfer between restricted classes of systems provide more meaningful explanations of target system behavior. Future work on the differences in implementation of these approaches in artificial systems would be interesting, perhaps providing justification in turn for refining GST as a philosophy of science framework (specifically focused on structures).

[10] "The principles of complex systems can be expressed generically, but we do not advocate this as a stand-alone pedagogical procedure. The principles are typically very difficult to understand when presented only in a generic form but highly intuitive when instantiated in a case study." Goldstone and Wilensky, 2008, p. 474

5.6 A New Logic of Discovery?

To wrap up, I would like to connect the present discussion with previous work on the logic of discovery. Analogical reasoning, and by extension transfer reasoning, is many times associated with *discovery*. That is, an agent learns something new through some (not always straightforward) procedure. In the philosophical discussion, the debate is whether we as philosophers can characterize a *logic* of discovery. Is there a way to formalize the process of discovery—and how does this relate to creativity and revolutionary thought especially in scientific reasoning? Of particular interest is whether such a procedure is *deductive*.

One might think that, by definition, certain revolutionary discoveries are not deductive. We romanticize *Eureka!* moments as spontaneous events of genius insight, and so it seems intuitively unlikely that we will be able to reconstruct a step-by-step procedure that leads to such an event. On the other hand, as discussed by Simon (1973), for other cases of discovery we might not have the same intuition. It might be possible to reconstruct a logical procedure for "everyday" discovery, for example the discoveries of normal science or in human cognition. Whereas for a logic of revolutionary discovery we would seem to need to solve a problem similar to that of induction, a logic of normal discovery would not. If we characterize normal science as systematic model-building and testing, for example, we can imagine logical steps to follow that might lead to discovery.

For the present purposes, discovery is most interesting when it enables learning about a new domain. Thus, a suitable formal reconstruction of transfer learning of some sort is a reasonable candidate for a logic of discovery. While one might find traditional formalizations of analogy as unsatisfactory, transfer learning in neural networks for machine learning provides a state of the art example that is arguably more interesting for a number of reasons. We can tell a plausible story about associationist cognition, as well as the usual story about neural networks being biologically inspired and similar in some relevant respects to human cognition.

The GST story of a logic of discovery is about being ready to interpret new problems as having structural similarities to old problems. GST is in the modeling business of making representations more similar, and in the machine learning context it can be seen as a framework advocating to not only re-format past experience but also bias future perception in ways that will promote positive transfer.[11] This is of course not perfect, and will result in bad transfer more or less depending on the implementations, but the basic idea is that some transfer will in general be better than no transfer.

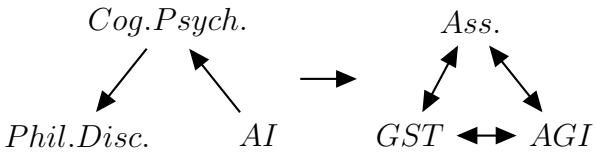
I will end with a simple overview of the kind of reasoning utilized for my arguments here. In Thagard, 1982, we are provided with a description of the

[11] See the discussion in Goldstone and Wilensky, 2008, p. 492-495 for more on the relationship between perception and similarity of representations.

relationship between AI, philosophy of discovery, and cognitive science:

> The reason that AI provides more than interesting analogies [between AI and philosophy of discovery] is that along with cognitive psychology it is concerned with many of the same kinds of problems that arise in philosophy of science, so that solution of those problems can only come through progress in all fields. Thagard, 1982, p. 167

Thagard summarizes his view of the relationship between these fields by a graph (left), which I modify (right) to summarize the direction advocated for in this article:



For my present argument, I am retrofitting the triangle for the purposes of AGI by robust transfer learning. As a sketch of the methodology here, I am plugging associationist cognitive science and general systems theory into the graph. I think it is safe to extend Thagard's graph to one where there is symmetric influence between all three fields. Then, we plug associationism into the cognitive science node, general systems theory (preferably structural systems theory) into the philosophy of discovery node, and AGI (with robust implementations of transfer learning) into the AI node.

Under this theoretical triangle, I believe that indeed some claims of near-future AGI based on robust transfer learning are justified. Furthermore, philosophical accounts of model-based reasoning, including inter-domain analogical reasoning, might benefit from this kind of inter-dependent theoretical framework. At the very least, I hope to have illustrated why, as Aggarwal, 2018, p. 454 notes, "Developing generalized forms of transfer learning [...] is a key area of future research" in data science, and why philosophers should be involved with the discussion as well.

Chapter 6

Bayesian Confirmation from Analog Models[1]

> When an experimenter comes to a difficult patch in the particular system he is investigating he may, if an isomorphic form exists, find that the corresponding patch in the other form is much easier to understand or control or investigate. And experience has shown that the ability to change to an isomorphic form, though it does not give absolutely trustworthy evidence (for an isomorphism may hold only over a certain range), is nevertheless a most useful and practical help to the experimenter. Ashby, 1958, p. 97

The word analogy comes from the greek ἀναλογία, meaning proportion, and analogical reasoning has been discussed since at least the time of Aristotle. Following the modern treatment found in Hesse, 1966, the form of an analogical argument is similar to solving proportion problems in mathematics. We are given the relation between two terms, and one term from a second relation. From these three terms, and the relation between the first pair, we can infer a fourth term completing the second pair. Using fractional notation such problems are of the form

$$\frac{A}{B} :: \frac{x}{D} \tag{6.1}$$

For simple mathematical ratios, we could have for example

$$\frac{2}{3} :: \frac{x}{9} \tag{6.2}$$

where we read out "two is to three, as x is to nine". This expression represents a transformation A such that $A(2,3) = A(x,9)$. In this case, if A is a division operator, x is determined by preserving the quotient on both sides of the equality. We look for an x that, under A as the first argument and 9 as the second argument, provides the same number. Solving for $x = 6$ involves interpreting the :: as an equality, and we simply multiply the diagonals yielding

[1] This chapter is the result of joint work with Roland Poellinger.

$3x = 18$. However, without some further clarification three possible interpretations might seem equally likely, that $x = 4, 6$ or 8. The latter option interprets the relationship between 3 and 9 as the addition of 6, and applies the same to the numerator. The first option interprets the relationship as squaring 3, and thus squaring 2 is 4. By convention, and use in analogical reasoning, we say that in this case $x = 6$ since it is the *proportion* that must remain the same—that is, it is not the relationship between 3 and 9 but the relationship between 2 and 3 that should be preserved on the right hand side for an analogical relation. This is a structural relation, since we could replace the 2 and 3 with any two numbers which are invariant in the quotient.

We can think of this example as a special case for this mathematical proportion problem, to be relaxed for analogies. We just require instead two transformations, A, A', where A' is relevantly similar to A. This is plausible since the arguments (or objects) being related may differ in important respects, or the relations themselves are not the same. In the mathematical example, the division operator is presumed to be implemented by the same algorithm for both pairs, which are composed of the same kinds of objects: integers. This need not be the case, and indeed analogies generally will not be precise enough to fulfill such object-level and relational level identities. However, under reasonable levels of abstraction we can understand analogical reasoning as a mapping of relevant transformational structure. That is, there is some kind of mapping or morphism $T : A \mapsto A'$.

Relational structure mapping is the key to analog relations according to Gentner, 1983. When we use the form of (6.1) to talk about scientific reasoning from an analog model, we might read out "A is related to B, which is similar to the way C is related to D". That is, there is some relation $Rel_1(A, B)$ that is similar to a relation $Rel_2(C, D)$. This relational structure is what is mapped between representations in an analogy, on Gentner's account, and is denoted here by the double colon notation '::'. The fraction bars are indicating some general relation (as opposed to actually representing division of integers, as in the mathematical example). As discussed in chapter 3, we can say that the kind of mapping relation between an analog model and a target system is just a special case of a model-target relation.

6.1 Analogy and confirmation

Probabilities are subjective degrees of belief. Theory confirmation in science is represented in the Bayesian framework by confirmatory probabilities—that is, a theory obtains higher probability given confirming evidence compared to the theory without this evidence. Hypotheses and Evidence are considered to be random variables, visualized as nodes in a network. These variables and their epistemic relations are represented in DAGs—Directed Acyclic Graphs.

For nodes H and E, we say that E confirms H when $P(H \mid E) > P(H)$. If H is a hypothesis and E some evidence for the hypothesis, we would expect our subjective degree of belief in H to be positively influenced by an observation of relevant evidence. In a causal Bayesian network representation of the situation, we intuitively expect E to be a descendant of H—that is, there is some causal or other directed connection between H and E. This aspect of a Bayesian network will be slightly revised later in order to accommodate our view of analog confirmation, since *analog* evidence E' is not, under minimal assumptions, a descendant of H nor is it causally connected. It is also not a direct *prediction* by the hypothesis.

In his book, Paul Bartha comments on analogical arguments and Bayesian epistemology, noting the related problem of old evidence:

> For Bayesians, it may seem quite clear that an analogical argument cannot provide confirmation. In the first place, it is not obvious how to represent an analogical argument as an evidential proposition E. Second, even if we can find a proposition E that expresses the information about source and target domains used in the argument, that information is not new. It is "old evidence," and therefore part of the background K. This implies that $E \wedge K$ is equivalent to K, and hence that
>
> $$Pr(H \mid E \wedge K) = Pr(H \mid K) \tag{6.3}$$
>
> According to the definition, we don't have confirmation. Instead, we have an instance of the familiar "problem of old evidence" (Glymour 1980). Third, and perhaps most important, analogical arguments are often applied to novel hypotheses H for which the "prior" probability $Pr(H \mid K)$ is not even defined. Again, the definition is inapplicable. (Bartha, 2010, p. 31)

Bartha goes on to suggest that the role of analogy in Bayesian epistemology is to raise prior probabilities of a considered hypothesis. While we agree that analogical considerations may impact prior probabilities, we think that Bartha has only discussed one way in which analogical reasoning may be used by scientists—and thus only one way in which analogies may shape Bayesian models. There is an important difference between a situation in which a specific analogical (or modeling) relation is *granted* by a scientist, and a situation in which the discovery or establishment of an analogy is taken as evidence. The latter is arguably what Bartha is referring to, whereas we will mainly be concerned with the former. These may just be two phases of one reasoning process, however we think they are conceptually and formally distinct.

We argue that Bartha's idea is captured in an epistemic network in which the analogy is modeled as a node—a random variable representing the possible existence or non-existence of an analogical relation. Confirmation from an analog model, on the other hand, *grants* the existence of an analogical relation. We think this is better modeled as an informational link (a *contour*) between theoretical domains. We will see how analogy *does* play a role in evidential statements which do not succumb to the old evidence problem. Therefore, traditional Bayesian confirmation is also accounted for in our approach, showing that analogical considerations can also impact posterior probabilities by taking into account evidence predicted by an analog model.

6.2 Analog simulation

First, we provide some background on the kinds of cases motivating our analysis. In a recent paper on analogical inference in physics, Dardashti, Thébault, and Winsberg, 2017 discuss an analogy between an accessible table top fluid system (the model or source) and a less accessible target system we want to gain insights about, namely black holes. The source system is prepared, manipulated, and observed. It is built not just to verify and exhibit control, but to make predictions and justify inferences about the target system.[2] The authors introduce the formal concept of *analog simulation*, explicating the intuition that the model can make predictions relevant for confirming the target theory. We will look at another example, also from the philosophy of physics, but first let us introduce a formal reconstruction of analog simulation as we see it, omitting the specific details to the case:

1. The target system T is represented as $\mathcal{M}_T$ in a suitably chosen modeling framework $\mathcal{L}_T$;

2. $\mathcal{M}_T$ is constrained by certain background assumptions $\mathcal{A}_T$, summarizing theoretical and empirical knowledge as well as the domain of conditions where the model is intended to apply;

3. $\mathcal{M}_T$ can be used to predict phenomena E_T, and will in turn be confirmed by evidence in accordance with E_T;

4. The accessible source system S is represented as $\mathcal{M}_S$ in a suitably chosen modeling framework $\mathcal{L}_S$;

[2] It seems to be a matter of opinion whether constructing the model results in any surprising new knowledge. As an anecdote, the present authors remember a physicist being unconvinced in spending any time on analog gravity models since *of course* they will exhibit the behaviors of interest because there are the obvious isomorphic equations describing the model system. This seems to be similar to the problem of old evidence Bartha is alluding to. We will, like (Dardashti, Thébault, and Winsberg, 2017), push forth under the assumption that model-building is informative for a significant group of interested parties.

5. $\mathcal{M}_S$ is constrained by background assumptions $\mathcal{A}_S$, and the domain of conditions which the model is intended to apply;

6. Just as on the T side, $\mathcal{M}_S$ can be used to predict phenomena E_S and will in turn be confirmed by evidence in accordance with E_S.

The source system S will now allow *analog simulation* of target T's behavior if (i) there exist exploitable structural similarities between $\mathcal{M}_S$ and $\mathcal{M}_T$ sufficient to define a robust syntactic isomorphism, and if (ii) this isomorphism is prompted by and based on a set of *model-external empirically grounded arguments*, abbreviated as MEEGA. The first condition is concerned with establishing the capacity to *control* and make relevant *predictions*. The scientist who builds the model system is doing so to test knowledge, and knowledge is demonstrated or gained through the ability to control the system and make accurate predictions. The second condition concerns a means to justifiably relate two apparently separate theoretical domains for the purposes of scientific reasoning. A scientist may feel justified for any number of reasons, from bare perceived similarities to expert knowledge of the general behavior of large classes of similar systems. Our reconstruction will focus on the first (and weaker) of the two, where there is some stake in building a model system for investigation and validation of the target theory.

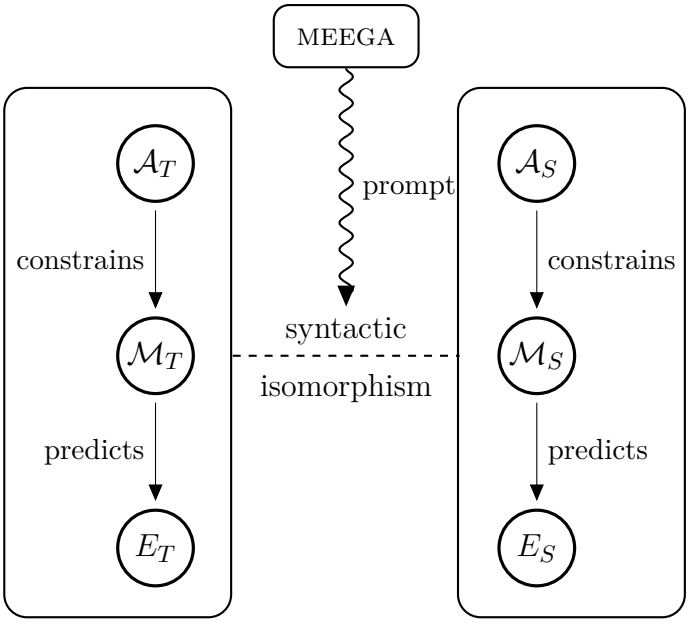


FIGURE 6.1: The analog simulation scheme: Framework $\mathcal{L}_T$ (left box) is used to model target system T in model $\mathcal{M}_T$; source system S is treated accordingly in framework $\mathcal{L}_S$ (right box).

Figure 6.1 relates the above sketch in a conceptual graph. MEEGA prompt the establishment of a bridge between theoretical networks in the form of a

syntactic isomorphism as translation between the systems' components. For Dardashti, Thébault, and Winsberg, 2017, observations of phenomena E_S in table-top fluid systems boost confidence in theoretical assumptions $\mathcal{A}_T$ about gravitational phenomena described in framework $\mathcal{L}_T$. The syntactic isomorphism (motivated by extensive expert knowledge about the underlying physics and mathematics of both frames) allows for the transfer of knowledge about acoustic Hawking radiation in the fluid system to Hawking radation in black holes. The model demonstrates knowledge by demonstrating control, and this is why it seems plausible there is a confirmatory boost to the target theory by observing evidence predicted in the model.

But suppose we are non-experts, unsure about the high-level physics and mathematics justifying the analogy. We want to be able to model confirmation from analog models even when they are not very pretty, and even when there might be obvious dissimilarities between the systems. It is worthwhile to find a way that Bayesian reconstructions in these cases can inform a general Bayesian account for model-based knowledge transfer and analogical reasoning in general. If our following case study (also an analog model from physics) does nothing else, we hope it helps triangulate an approach for modeling knowledge transfer in cognition and artificial systems.

6.3 Water wave analog of the Casimir effect

The example of an analog model we will consider here is the table-top fluid model of the Casimir Effect investigated by Denardo, Puda, and Larraza, 2009. The model is a physical or material analogy. The quantum Casimir effect is produced between two very small (and very thin) uncharged parallel metal plates that are placed close together in a vacuum. At certain distances d the plates are pushed together, at others they are pushed apart. How do we explain this? In quantum theory there is a non-zero energy associated with the ground state of each mode in the quantum vacuum $hf/2$ (where f is the frequency of the harmonic oscillator associated with the mode and h is Planck's constant). We can account for Casimir effect behavior in quantum electrodynamics by calculating the relative difference in pressure between the force of electromagnetic radiation outside the plates and inside the plates, since the closeness of the plates excludes certain wavelengths in the background spectrum. The spectrum of zero point frequencies is infinite both on the outside and inside of the plates, but after renormalization Denardo, Puda, and Larraza (2009, p. 1095) note that the result of the calculation gives a force of $\pi^2 \hbar c/240d^4$ per unit area.

The water wave analog model that the authors construct consists of a table-top bath which is vertically vibrated according to a range of frequencies (10-20 Hz) which excites surface waves. Two acrylic or PVC plates are hung in parallel above the bath and dipped into the vibrating fluid bath (VFB). The

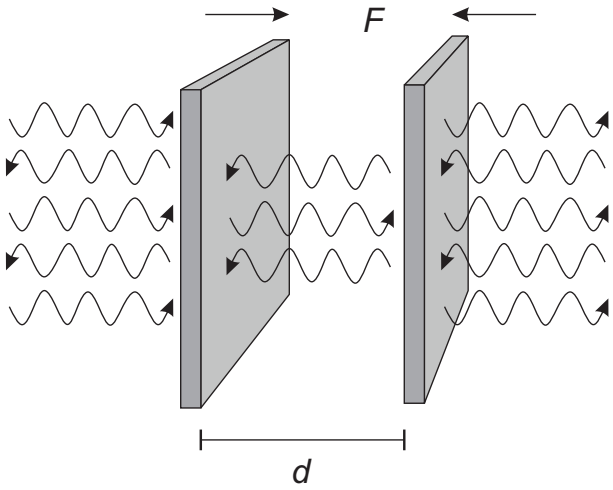


FIGURE 6.2: The abstract scheme underlying both the quantum Casimir and the fluid Casimir effect; F indicates the attractive force, d the distance between the two plates.

surface waves and VFB are analogous to the zero point fluctuations in the quantum "vacuum". The PVC plates experience an analog Casimir force, which we explain in terms of the difference in pressure due to the exclusion of certain waves between the plates.[3]

We can model the relationship between these two systems in the following way according to the similarity relation discussed above. Say the phenomena of two parallel plates being pushed together when dipped in the VFB is E_F. To differentiate, let us say that E_Q is the *similar* effect on Casimir plates. The VFB for the analog model is B_F, and it has a causal relationship to E_F in that it causes the relative pressure due to wave motion to be greater on the surface area outside the plates than on the surface of the interior. Thus far we can say—in the form of a fractional notation:

$$\frac{B_F}{E_F} :: \frac{X_Q}{E_Q}. \tag{6.4}$$

Our choice for X_Q, if we had no other knowledge of the target system, would arguably be an X_Q that fulfills similar conditions as B_F (that is, X_Q should be some causal explanation like B_F, i.e., $X_Q = {B_F}'$). It seems that our expectation for an $X_Q = {B_F}'$ is *greater* than other options—given that we have established some justification for applying the analog model in the first place (i.e., we have at the very least $E_F :: E_Q$), as well as subjective perceptual and theoretical justification for building the model in the first place.

[3]Also mentioned in both Denardo, Puda, and Larraza, 2009 and Barrow (2002, §7) is a larger instance of macroscopic 'Casimir' effects where parallel ships rolling on a swell will result in the destructive interference of waves between the ships, thus allowing the relative pressure difference from the waves on the outer surfaces to push the ships closer together.

Importantly, this set up is different from that considered in Dardashti, Thébault, and Winsberg, 2017 since it is not the existence of a particular effect that is the unknown variable in the analogical argument. In their discussion, the effect of Hawking radiation is what is inaccessible—black holes are presumed to exist. Here, what is inaccessible is not the effect (we already have observed Casimir plates coming closer together). Rather, what is inaccessible is the background medium or field which is supposed to be part of the causal mechanism of how the observed phenomenon is produced—*but which cannot be directly observed.*[4] What is confirmed in their example is a theory of black holes and Hawking radiation. What is potentially confirmed here by the argument $X_Q = B_F{}'$ is any quantum theory of the vacuum which gives an ontology of non-zero energy density using a mechanism or term that is similar to the vibrating fluid bath (i.e., a $B_F{}'$). We will now refer to the quantum mechanism as B_Q.

In our example, we know that the analogical relationship at least breaks down with respect to the *dimensions* of the analog model and the target system. The authors consciously note other deficiencies in the analogy Denardo, Puda, and Larraza, 2009, p. 1095:

> The analogy of our water wave system to the Casimir effect is not exact. Because the water waves are driven, the energy density of the spectrum is not infinite, so a regularization procedure is not needed. Furthermore, we are primarily concerned with the case of closely spaced plates, which yields a force that is independent of the separation distance d. This behavior is in contrast to the Casimir force, which has a $1/d^4$ dependence due to the divergence of the ω^3 spectrum at high frequencies.

Furthermore, there are terms in our formal representation of the fluid such as viscosity and surface tension which have unclear analogical relationships with the quantum world. However, in our view these are not particularly troublesome. The analogy concerns the relative difference in pressure between the exterior and interior of two parallel surfaces in an oscillating medium composed of a range of frequencies. We are now taking this analogical relationship as *granted*. We are no longer in a discovery phase, but applying the "discovered" model to see how we learn, how our beliefs are affected.

Before introducing a Bayesian network in the next section which can appropriately model confirmation from B_F (the analog model) we should first like to know the kinds of probabilities that should hold in such a network in order to

[4] "Although [zero point energy] cannot be directly observed, the presence of the plates discretizes the spectrum between and transverse to the plates, which causes the imbalance of the radiation force."Denardo, Puda, and Larraza, 2009, p. 1095

ensure confirmation. As mentioned earlier, we can consider quantum electrodynamics T_Q to be the theory of electromagnetic radiation in a quantum vacuum—the relevant theory for explaining the Casimir effect. Unfortunately, measuring the zero point fluctuations directly is not possible, and thus—obviously—the existence of such an ontology is independent from the theory. In other words, since we have not observed B_Q, then $P(T_Q \mid B_Q) = P(T_Q)$. This is a problem since B_Q is supposed to give us the causal explanation of E_Q, and surely we should have that $P(T_Q \mid E_Q) > P(T_Q)$—i.e., that observing the Casimir effect confirms a quantum theory of the vacuum.

Considering the analog model, however, it seems plausible that as a Bayesian agent the mechanism offered by the fluid system confirms (increases the probability of) an analogous mechanism in the target system. It should thus be that $P(B_Q \mid B_F) > P(B_Q)$, as well as $P(B_F \mid E_F) > P(B_F)$—i.e., observing evidence predicted by a model of the bath system should confirm the model. In the end we will want that $P(T_Q \mid E_F) > P(T_Q)$, that the analog phenomenon confirms the target theory. This stems from the final positively correlated module $P(T_Q \mid B_Q) > P(T_Q)$.

So the Bayesian network should allow the following assumptions to hold:

$$P(T_Q \mid B_Q) > P(T_Q) \tag{6.5}$$

$$P(B_Q \mid B_F) > P(B_Q) \tag{6.6}$$

$$P(B_F \mid E_F) > P(B_F) \tag{6.7}$$

In the example of the fluid analog model of the Casimir Effect, the theories of fluid mechanics T_F and quantum electrodynamics T_Q have evidence domains concerning macroscopic fluids E_F and electromagnetic radiation in a vacuum E_Q respectively. The difference in orders of magnitude justifies by itself the assumption that the overlap between evidence of these theories is, under normal conditions, nonexistent. Our situation is notably different from the overlapping evidence in the networks discussed in Dizadji-Bahmani, Frigg, and Hartmann, 2011. However, it is seen that *after* the analogical argument is admitted into the network, both B_F and E_F are common descendants of T_F and T_Q. This is the first preliminary look at one Bayesian model which affords the kind of confirmatory relationship we are looking for.

An analogy is made with the relative pressures of waves on parallel partitions in the (practically) random 'bath'—either surface waves in a macroscopic fluid or electromagnetic radiation in a quantum vacuum. These waves are assumed to be linear, since non-linear mechanics begin to appear upon sufficiently high amplitude vertical oscillations of the fluid bath when droplets are ejected. See Denardo, Puda, and Larraza, 2009 and Terwagne and Bush, 2011, for example.

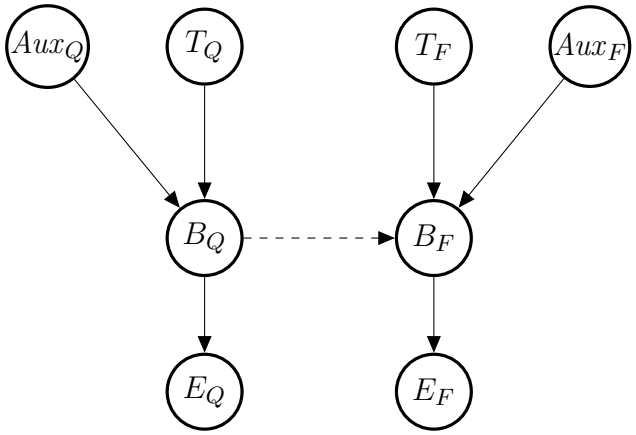


FIGURE 6.3: A Bayesian network representing the relationship between the analog Casimir effect E_F and the Casimir effect E_Q. The dashed edge marks analogy as asymmetric link in order to capture the subjective sense in which $P(T_Q \mid E_F) > P(T_Q)$, by making E_F and B_F descendants. Also $P(T_Q \mid B_F) > P(T_Q)$.

Thus, the limitation of the system to approximately sinusoidal waves is a relevant aspect of the analogy that we can call an *auxiliary* condition. These and other auxiliary conditions from the experimental set up that produces both the Casimir effect and the analog effect, plus the respective relevant theories, gives us descriptions of the fluid and quantum systems B_F and B_Q respectively, which are related through the analogical argument given earlier. However, it still isn't exactly straightforward how to represent the analogical relationship between B_F and B_Q in the network. A possible choice for directing an edge between these nodes in the Bayesian network would be from B_Q to B_F: In this way, B_F and E_F are descendants of T_Q, but the arrow also introduces asymmetry in the graph.

6.4 Bridging models

The above network fails to capture the epistemic information required of an analog relation based on similarity between target and model systems. It represents, if you will, only *half* of a superposition of *two* Bayesian networks, each yielding answers to queries for different inferential directions—if B_F and B_Q behave analogously, evidence for either side should inform the other side in a symmetric way. The purpose of the present work is to show that such a case should be admitted as normal into the Bayesian formal framework. The structural similarities between these domains is a matter of course and not an exception to the probabilistic reasoning Bayesian epistemology aims to reconstruct. In this section we provide an alternative account of Bayesian confirmation, utilizing undirected edges to bridge theoretical frameworks and preserve the symmetric capacity of analog relations. The undirected edges are a minimal extension to

normal Bayesian networks, convenient shorthand for preserving the symmetrical information.

6.4.1 Directed and undirected relations

Our choice to direct the edge from B_Q to B_F in Fig. 6.3 is rooted in the epistemic status and goals of a scientist (or philosopher) in a particular situation. Simply put, we want to confirm T_Q, not T_F. If we wanted to confirm T_F we would switch the arrow the other way around.[5] So, E_F and B_F must be descendants. If the arrow pointed in the other direction and they weren't descendants, the collider structure at B_Q would d-separate T_Q from B_F and E_F, giving us $T_Q \perp\!\!\!\perp B_F, E_F$. In other words, if the collider were present then $P(T_Q \mid B_F) = P(T_Q)$, where instead we want that $P(T_Q \mid B_F) > P(T_Q)$.[6]

For a specific inquiry, confirmation from an analog model can be represented in a DAG. Yet, we may wish to have confirmation flow the other way (e.g., to T_F) and our formal system should be prepared with the extra information on hand to do this. If we were to use *only* the above DAG in our representation of the problem, then we would lose the information relevant to the symmetries of an analogy and the potential to confirm in the opposite direction (i.e., $P(T_F \mid E_Q) > P(T_F)$). Since we are not looking for a case-by-case account of analog confirmation, we want to preserve this information in a more general framework that ties in with Bayesian confirmation theory. As shown, a standard Bayesian network is insufficient to adequately handle representing confirmation from analog relations as we have construed them. The question remains: How can intertheoretical, symmetric relations be integrated in a formal model from which genuine (Bayesian) confirmation claims can be derived?

We suggest a new type of edge—a non-directed, non-causal, informational link. Furthermore, it is not to be deactivated by (causal) interventions in the model. It should work like synonyms, mathematical inter-definitions, or logical relations (which certainly all belong to the pool of knowledge we use for decision making). A deterministic function, for example, results in the special case of a maximal mutual information between the variables. Information gained about one results in the same amount of information about the other. There is, so

[5] However, the unobservability of the quantum "bath", B_Q, may present some issues. For example, it seems like it should be the case that $P(T_F \mid B_Q) = P(T_F)$. If we can't observe B_Q directly (this is the reason the analogical argument was made in the first place) then we can't condition on it like it is observed evidence.

[6] This argument similarly goes for an edge between E_Q and E_F, were we to choose to express the analogical relationship between the two frames at the evidential level. Also, our previous discussion used analogical reasoning to map the structure between B_F and E_F to suggest B_Q. If the fluid bath was not granted as analogical, a mere similarity of evidence would arguably not justify the strong intuition that we might want to confirm T_Q. One might be getting similar evidence from *dissimilar* systems—e.g., mere numbers from point measurements, similar but unstructured. What is important is precisely the structural mapping between the descriptions of systems.

to say, a communication channel between the representations of systems. In standard statistical modeling, extensionally equivalent variables would certainly be collapsed into one single variable (node, respectively). For our purposes, though, we would like to disambiguate in the model the intensional distinction between connected variables. Consequently, the final model ought to contain two separate nodes and mark these nodes as tightly, functionally dependent.

In briefly discussing the possibility of embedding such non-causal links into causal Bayes net structures, Verma and Pearl acknowledge the usefulness of such hybrid models:

> The ability to represent functional dependencies would be a powerful extension from the point of view of the designer. These dependencies may easily be represented by the introduction of deterministic nodes which would correspond to the deterministic variables. Graphs which contain deterministic nodes represent more information than d-separation is able to extract; but a simple extension of d-separation, called D-separation, is both sound and complete with respect to the input list under both probabilistic inference and graphoid inference. Verma and Pearl, 1988, p. 75

Symmetrical inter-theoretical relations like analogies are in our view a paradigm case study to implement such an extension. We thus propose to model analogy as a relation between strictly correlated variables. It is a non-causal and non-directional relation constructed on top of a syntactic isomorphism (formalized as a 1-1 function) in an extensions of a Bayes net causal model. Such hybrid structures have been discussed in philosophy (Poellinger, 2012), as well as statistics (e.g., as chain graphs in Lauritzen and Richardson, 2001). We can extend a standard causal model triple $M = \langle U, V, F\rangle$ to a quadruple $\mathcal{K} = \langle U, V, F, C\rangle$, where U is a set of exogenous variables, V a set of endogenous variables, F a set of functional causal mechanisms (cf. Pearl, 2000, def. 7.1.1, p. 203). The extension, C, is a set of *epistemic contours*: a set of 1-1 functions $i_{j,k}$ that take the value of some variable V_j and assign the value $i_{j,k}(V_j)$ to some other variable V_k in the pattern. Importantly, intervening on one of these entangled variables will not break the contour.

Contours possess the properties we want for our analog relations. Yet, embedding entangled variables of this kind in Bayesian networks precisely renders them non-Markovian.[7] In the general case, the inferential framework must be

[7] When Pearl claims that "[t]he Markovian assumption [...] is a matter of convention, to distinguish complete from incomplete models"(Cf. (Pearl, 2000, p. 61) he naturally has Bayes net causal models (with distinct variables) in mind, which we just dismissed.

tweaked to retain soundness,[8] but in our special case with a single intertheoretical bridge, we only require the idea of utilizing an undirected functional link to join two probabilistic chains (i.e., our two frames). So, how can we spell out analogical inference across this newly introduced bridge?

6.4.2 Analogical inference across symmetric links

In our proposal, the model-external postulate (or assumption, or also perception) "B_Q is similar to B_F (in certain known respects)" prompts the inclusion of a translation relation rather than the insertion of a new node. Analogical reasoning begins with a domain comparison which we characterize as the insertion of an inter-theoretical bridge.[9] Figure 6.4 is a rendition of the Casimir effect example discussed above with the contour i marking the analogical relationship between the frames at the level of systems B_Q and B_F.

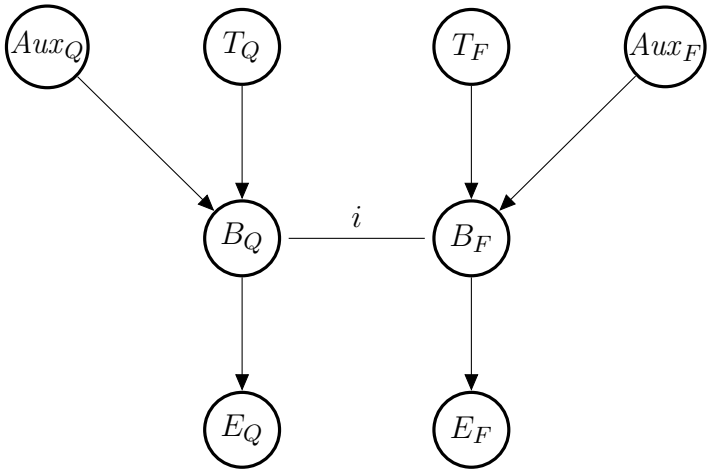


FIGURE 6.4: System-level analogy as epistemic contour i with the intertheoretical bridge i between B_Q and B_F.

In this graph, the undirected edge between B_Q and B_F, along with the formal explication we have introduced, provides a means for implementing analog confirmation as we have construed it. A scientist or an artificial system can obtain $P(T_Q \mid E_F) > P(T_Q)$ while retaining the information represented by the undirected edge, and the ability at some later time to provide confirmation for T_F. While in this case we might not need or want to confirm T_F, a general account *should* provide for this.

[8] For consistent reasoning and efficient computation of causal knowledge patterns to remain possible at all, acyclicity, independence (as expressed in the graphical d-separation criterion), and the *identifiability of causal effects* receive new explications. Poellinger, 2012 introduces a further graphical criterion, the *principle of explanatory dominance*, to define under which conditions the Markov requirement can be reclaimed and extended Bayes nets can be utilized for causal inference.

[9] This insertion can formally be understood as a structural mapping by which two frames are related.

One might wish to instead represent the analogical contour between T_Q and T_F (Figure 6.5). However, this would be a much stronger claim since B_Q and B_F are determined to an additional extent by auxiliaries. After all, confirmatory boosts in probability may be divided between the theory and auxiliary assumptions as per the Duhem-Quine problem. An analogy at the theory level is, in some sense, an analogy that could be a step further towards unification than one at the system level. We will return to briefly discuss (pre-)unification in Sec. 6.5.

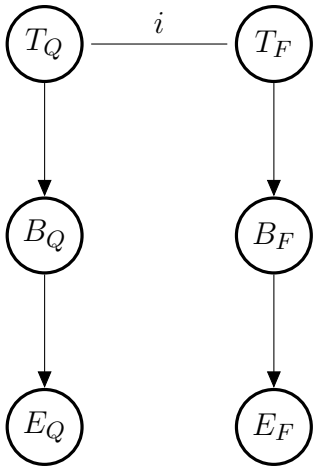


FIGURE 6.5: Theory-level analogy as epistemic contour i with the intertheoretical bridge i between T_Q and T_F.

A potential option would also be to insert a collider structure between the model (system) levels representing an analogy. However, as we have argued, granted analogies should not be represented as a node. It is unclear what the content of the node would be, and the values it could take would arguably depend upon a meta-level analysis of the network (i.e., it would be a self-referential node). We think our approach of modeling the analogy as a functional relation is more consistent with the case study, as well as mathematically useful for future applications of the method. Furthermore, there is support from the cognitive science literature treating analogy as an associative perceptual mechanism (i.e. structure mapping). In this way, the transfer of knowledge from one domain to the other is *direct*, not necessarily the result of a top-down generalization.

6.4.3 Translation via relevant sub-isomorphisms

We take analog contours to be an expression of a modeling relationship between frameworks. It can be thought of formally as a translation relation based on a *relevant sub-isomorphism*, which has been anticipated in the literature on models and representations in science, cf. Frigg and Hartmann, 2012:

> One version of the semantic view, one that builds on a mathematical notion of models, posits that a model and its target have to be isomorphic (van Fraassen, 1980; Suppes, 2002) or partially isomorphic (Da Costa and French, 2003) to each other. Formal requirements weaker than these have been discussed by Mundy, 1986 and Swoyer, 1991. Another version of the semantic view drops formal requirements in favor of similarity (Giere, 1988; Giere, 2004; Teller, 2001). This approach enjoys the advantage over the isomorphism view that it is less restrictive and also can account for cases of inexact and simplifying models. However, as Giere points out, this account remains empty as long as no relevant respects and degrees of similarity are specified. The specification of such respects and degrees depends on the problem at hand and the larger scientific context and cannot be made on the basis of purely philosophical considerations (Teller, 2001).

We follow this line of reasoning and formulate a relevance filter in order to capture the purpose-driven selection of theoretical entities to be translated. Of course, basing analogy on a purpose-driven relevance concept makes the concept of analog models context-specific. The important point is that there is *some* structural relationship which is mapped between the model and target systems. We can embrace this and call B_F an analog model of B_Q *relative to*

1. a relevance filter Rlv;
2. a bijection between the relevant properties of B_Q and B_F (an isomorphism between sub-structures of B_Q and B_F).

The filter function Rlv, an indicator function over the descriptive elements of both frameworks, selects for each semantic category (for individual objects and each set of n-ary relations between such objects) subsets of equal magnitude; i.e., for each category:

$$|| \, Rlv(B_Q) \, || = || \, Rlv(B_F) \, || \, . \tag{6.8}$$

If B_Q and B_F behave alike with respect to relevant parts (i.e., parts selected by Rlv) that are described by $P_Q(x)_Q$ and $P_F(x)_F$ (with properties P of objects x in the respective models), then the following formula explicates the analog relation between frameworks via translation i:

$$\forall P_F, \mathbf{x}_F (P_F(\mathbf{x}_F) \leftrightarrow P_Q(\mathbf{x}_Q)) \, . \tag{6.9}$$

Note that this isomorphism might be the result of iteratively fine-tuning non-bijective translations between the frameworks.[10]

[10] We are thankful to Mark Colyvan for valuable discussions about the nature of this morphism.

Having defined the propagation of information across the epistemic contour in this way, tracing confirmatory support in Fig. 6.4 yields the following:

$$P(B_F \mid E_F) > P(B_F) \tag{6.10}$$
$$P(T_Q \mid B_Q) > P(T_Q) \tag{6.11}$$
$$P(i(B_F) \mid B_F) > P(i(B_F)) \tag{6.12}$$

where $i(B_F)$ represents specific information about the properties of B_Q relevant for the analogical inference (i.e., as chosen by the filter function).[11] Eq. 6.12 exploits the characterization of contour i as 1-1 function: Learning B_F tells us more about the *Rlv*-selected properties and objects at the core of B_Q, thereby raising our degree of belief in those B_Q that are compatible with $i(B_F)$. Now, by transitivity, 6.10, 6.11, and 6.12 together entail

$$P(T_Q \mid E_F) > P(T_Q), \tag{6.13}$$

which was implied by our list of desiderata above—Eq. 6.5, Eq. 6.6, Eq. 6.7, chained together. Formula 6.13 is an instance of Bayesian confirmation—this time across theoretical frameworks, though, and it encodes what we set out to achieve: Bayesian confirmation from an analog model.

6.5 Analogy and (pre-)unification

In her structure-mapping account of analogical reasoning, Dedre Gentner makes an important distinction between mere similarity, analogy, and abstract generalities or law-like statements. These represent different stages of learning about the relationship between two domains, moving from early similarity comparisons, to analogies, to generalizations:

> This sequence can be understood in terms of the kinds of differences in predicate overlap discussed in this paper. In the structure-mapping framework, we can suggest reasons that the accessibility and the explanatory usefulness of a match may be negatively related. Literal similarity matches are highly accessible, since they can be indexed by object descriptions, by relational structures, or by both. But they are not very useful in deriving causal principles precisely because there is too much overlap to know what is crucial. Potential analogies are less likely to be noticed, since they require accessing the data base via relational matches; object matches are of no use. However, once found, an analogy should be more useful in

[11] As soon as one learns of a specific instantiation of $i(B_F)$, i.e., the relevant core of B_Q, those theoretical entities not in the *Rlv* mapping must be updated in line with consistency requirements.

> deriving the key principles, since the shared data structure is sparse enough to permit analysis. [...] To state a general law requires another step beyond creating a temporary correspondence between unlike domains: The person must create a new relational structure whose objects are so lacking in specific attributes that the structure can be applied across widely different domains. (Gentner, 1983, p. 167-168)

This contextualizes our approach and the way in which analogy can be seen as the pre-unification of two theoretical frames. The way we have modeled the link between these frames may be understood by Gentner as a "temporary correspondence between two unlike domains". The Casimir example discussed is arguably in the middle of Gentner's spectrum between bare similarity and abstract generality. There are literal similarity matches—but some features in representations of the respective domains must also be thrown out as not similar. There is relational structure being mapped—but the objects still have enough specific attributes that it may be unwarranted at this stage to make any conclusive generalization about an entire class of systems. Although, in this case, it may be appropriate. Sliding the undirected contour up or down in the extended Bayesian network seems to plausibly correspond to the levels of analysis Gentner has outlined.

That said, we can imagine that such a class could be built up in a case-by-case manner, and eventually justify a unified claim regarding the structure of all domains in the class. Indeed, there are cases in science where strong or systematic analogies can be thought of as almost unificatory (see Bartha, 2013). We think there is strong motivation for interpreting contours at the level we have utilized them (i.e., between model systems) as pre-unificatory analogies. This might be contrasted with, for example, an approach that models the inter-theoretical relationship as a sort of common cause (i.e., a parent node of both structures at the uppermost theory level). We think that these two approaches can coexist, representing different stages of epistemic modeling. An *analog knowledge pattern* can precisely represent a scientist's nuanced view of an inter-theoretic relation *before* she might wish to consider that the theories under question should somehow be unified into one theory.

As a final note for additional follow-up work by fellow philosophers of physics, we might consider confirmation of the general class of quantum theories such that with regard to the quantum vacuum they contain a B_Q—which is analogous to the vibrating bath in the analog model. In this sense, a particular interpretation of the Casimir effect seems to be implied by the conceptual aspect of the explanation: the argument $X_Q = B_F{}'$ supports a field-theoretic explanation in terms of a spectrum of modes rather than that the effect is due to Van der Waals force. This is perhaps a problem for explaining the Casimir

effect in terms of the Van der Waals force.[12] Furthermore, there are many other systems which exhibit similar phenomena (Denardo, Puda, and Larraza, 2009, p. 1100):

> Casimir-type effects occur, in general, for two bodies in a homogeneous and isotropic spectrum of any kind of random waves that carry momentum. A net attractive force occurs between two parallel plates in the typical case where the radiation force is reduced between them.

The various analog models of Casimir-type effects seem to provide some sort of unifying explanation, whereas an alternative explanation of the Casimir effect in terms of the van der Waals force is in contrast to such models.[13] However, the systems are not all described by a single unifying theory—and thus the weaker analogies between models might be taken to supply a form of *pre*-unificatory explanations in the cases where well-defined structural similarities might eventually lead to theoretical unification. Grounded in an analytic approach towards the concept of analogy, our constructive proposal provides a novel template for making such epistemological advances explicit.

6.6 Conclusion

The extended subjective Bayesian network presented here is able to account for confirmation from analog models and analog simulation. Thus, slightly modified, Bayesian confirmation theory is able to meet the challenge offered in Dardashti, Thébault, and Winsberg, 2017. Importantly, our account preserves the informational symmetries involved in analogical reasoning, as demonstrated in an application to a case study from philosophy of physics, the Casimir effect. It should be reemphasized that, in this case, what is inaccessible about the target system is not the phenomenon—both the target and analog systems have shown the plates moving closer together—but rather the theoretical and ontological explanation of why the target system produces the phenomenon. Thus, we are confirming a theory of a phenomenon by offering an explanation of an analogous phenomenon.

[12] Also, an explanation in terms of 'virtual particles' flitting in and out of existence seems inferior, given these results, to an explanation in terms of crests and troughs in a fluctuating medium.

[13] Unless van der Waals and Casimir forces were shown to be equivalent, but our understanding is that they are distinct. We remain open to further discussion on this point.

Chapter 7

Decoherence and Survival

> The existence of laws of similar structure in different fields makes possible the use of models which are simpler or better known, for more complicated and less manageable phenomena. Therefore general system theory should be, methodologically, an important means of controlling and instigating the transfer of principles from one field to another, and it will no longer be necessary to duplicate or triplicate the discovery of the same principles in different fields isolated from each other. At the same time, by formulating exact criteria, general system theory will guard against superficial analogies which are useless in science and harmful in their practical consequences. Bertalanffy, 1969, p. 81

The physical world is ultimately quantum, and our interactions with it are governed by quantum measurements, quantum uncertainty, and quantum logic. What does this mean for human cognition and decision-making? One might suggest first looking at the agents whose survival depends most on correctly discriminating states in the quantum world: physicists. We find out rather quickly that physicists do not have to use quantum mechanics to solve *all* problems. They do not use it to make sense of the everyday world they live in. Quantum decoherence is a feature of quantum dynamics that many physicists appeal to in order to supply a story or explanation why the world humans live in is best characterized *classically*. Classical physics is the physics of large objects at low-speeds—the world of billiard balls and bears. The most efficient way to solve problems in this regime is to utilize classical mechanics.

The objects humans interact with, internally model, and have beliefs about are *typically* classical. That is, they are subject to classical measurement, classical definiteness, and classical logical relationships. There are exceptions. Some macroscopic systems are characterized quantum mechanically. This is, however, *atypical*. These are highly technical situations that involve, for example, quantum fluids or superconducting systems. There is little to no justification to think that human agents have ever had to make decisions over these kinds of systems until the last hundred years or so.

Arguably, it is by far a rare exception that some human belief state might be appropriately modeled as 'quantum'. That is, unless one is talking about a special class of phenomena or problems where a quantum belief state appropriately mirrors the distribution of these inputs. That is to say, a *quantum physicist* should have quantum belief states when she is trying to distinguish quantum states and quantum phenomena in the quantum world. If one wants to be as certain as possible about what a quantum state is (or will be), a modeler will use quantum representations, quantum measurements, and quantum logic.

A quantum state can exist in a so-called 'superposition' of states, and this is not merely an uncertainty between classical states. That isn't to say that macroscopic states of objects like water are atypical because of the superposition property which holds between multiple waves on a fluid surface. Most objects, most tasks, most situations encountered were surely better modeled classically. A superposition or wavelike model would impede decision-making in the same way that a physicist would be impeded trying to distinguish some macroscopic state by parsing a high-dimensional wave function.

What, then, should one make of recent quantum-like models (QLMs) of cognition? Are we wrong about what is a typical task for human agents? Does the success of these models suggest that quantum-like representations are better than classical alternatives? I argue that they do not do better, and that the intuitions about classical representations being optimal for the classical world are correct. Even granting that maybe there is some use for these models in a way which does not contradict 'classical' cognition, I argue that decoherence should still occur among realistic sets of un-isolated belief states.

There are already well known decoherence objections to actual quantum mechanisms in the brain. Tegmark, 2000 Proponents of QLMs want to remain agnostic about what carries their quantum-like mechanisms of cognition, and just use the quantum formalism:

> We remain agnostic with respect to this question, and instead focus more on the application of the mathematics, of the formal core of quantum theory, to the behavioral results obtained from cognition and decision experiments. Busemeyer and Bruza, 2012, p. 25

Unfortunately, it is clear that they also appeal to certain concepts that are not easily divorceable from physical interpretations in quantum theory. They are also not divorceable from the *formal* mechanisms which are sufficient for a quantum-like decoherence argument. The debate over decoherence in quantum mechanics represents precisely the kind of interpretational problem which plagues the quantum formalism. One does not need *actual* quantum physical systems to interfere with the environment (such as other quantum states) to make a decoherence argument against coherent quantum models of cognition. Even 'agnostic' formal models will need to justify one set of formal procedures

which act on sets of (quantum) belief states over another. If I assert a no-collapse interpretation of quantum-like belief dynamics, who can stop me?

Nevertheless, I argue that the procedures these models rely on are not as agnostic as they seem. Indeed, I think these models lie squarely in the algorithmic level of Marr's hierarchy (see e.g. Marr, 1982; Marr and Poggio, 1976). There are a set of assumptions, and a sequence of operations with these assumptions, which result in the ability to model (or produce) formal results aligning with behavioral data. In particular, behavioral data which shows widespread mistakes like the conjunction fallacy. That being said, I will grant for the most part that these models may be construed as potential computational-level alternatives to Bayesian models.

7.1 Classical Bayesianism

Bayesianism provides a strong story for what a human agent should compute, and why. It is, essentially, supposing that cognition is classical in its logical structure to reflect the classical structure of typical stimuli in the environment. If it weren't, the mismatch with reality would result in certain losses which could be avoided by modeling the world appropriately (classically).

In a Bayesian decision theory, one interprets probabilities as subjective degrees of belief, and update in accordance with Bayesian updating. Central to this update procedure is Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7.1}$$

This theorem assumes $P(B) > 0$, and is proven by the commutativity of set intersection, where abstractly defined events A, B are elements or subsets of the entire space of events Ω. One might wonder, in the context of *subjective* decision theory, whether this commutativity is actually justified. No one seems to question the commutativity of abstract objects in set theory, and neither do I, but when A and B are events in a subjective belief space this might be questioned. In other words, there is non-trivial philosophical work to do to support the claim that subjective degrees of belief obey commutativity as well.

Proponents of quantum like decision modeling emphasize the non-commutativity of projection operators corresponding to incompatible events in an agent's cognitive process, it is important to see that it is crucial in the proof of Bayes theorem.

Proof of Bayes thm:

Definition Conditional Probability $P(A|B) = \frac{P(A \cap B)}{P(B)}$

1. $P(A|B)P(B) = P(A \cap B)$

2. $P(B|A)P(A) = P(B \cap A)$
3. By commutativity of the set intersection, we can substitute so that $P(A|B)P(B) = P(B|A)P(A)$
4. $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Again, in terms of static or logical set-theoretic properties, it does not seem justified to question commutativity. These events are by definition compatible, and they will commute. This brings up the discussion in Birkhoff and Neumann, 1936, where the event structures of classical vs. quantum experimental propositions were being considered. The structure for experimental propositions in quantum mechanics is different from the corresponding structure in classical physics. The way an agent, like a physicist, interacts with the physical world obeys a slightly different set of logical relations for quantum events compared to macroscopic or classical events. Without going into detail here, one could characterize the difference as *boolean* and *non*-boolean. If the agent is to be successful at *doing* physics, then these logical relations should be adhered to in the respective domains. Decoherence is an explanation for why classical descriptions are appropriate for macroscopic physics despite the fact that the world is actually quantum underneath.

Are the kinds of problems that are typically encountered by a (typical) agent more like those found in quantum or classical mechanics? Most importantly is the issue of whether an agent *should* somehow act upon representations which commute, or whether they just descriptively do not always do so. I will return to this discussion in later sections when discussing the success of a physicist as an agent who "regulates" physical phenomena, and how decoherence provides the key to understanding why this physicist will be the best regulator by respecting quantum decision principles *only* for quantum problems.

Proponents of QLMs want to revise arguably the best computational level accounts, those which provide some computational story of minimizing expected error/loss and why the system *should* do so, without providing any alternative norms. Take, for example, the following quotes:

> There is another line of research in which quantum physical models of the brain are used to understand consciousness (Hammeroff, 1998) and human memory (Pribram, 1993). We are not following this line; instead, we are using quantum models at a more abstract level, analogous to Bayesian models of cognition. Busemeyer et al., 2011, p. 193

Elsewhere, proponents acknowledge the computational foundations (i.e. a "top down" approach in Marr's hierarchy, eliciting the *hows* and *whys* of cognition) of the Bayesian paradigm.[1] The authors clearly contextualize their approach as an *alternative* to the Bayesian paradigm. Pothos and Busemeyer, 2013, p. 257

> We seek to infer the computational (and possibly process) principles of cognitive processing. Pothos and Busemeyer, 2013, p. 319

I do not think I am out of line in interpreting the QLM program as attempting to replace the (Bayesian) classical normative picture at the computational level. However, the normative force of classical accounts goes much deeper into Marr's hierarchy than QLM proponents are ready to admit. Consider again the following statements I discussed in chapter 3 from von Neumann concerning the capacity of physical components in a computational device to be interpreted in multiple ways:

> "The electromechanical relay, or the vacuum tube, when properly used, are undoubtedly all-or-none organs. Indeed, they are the prototypes of such organs. Yet both of them are in reality complicated analogy mechanisms, which upon appropriately adjusted stimulation respond continuously, linearly or non-linearly, and exhibit the phenomena of "breakdown" or "all-or-none" response only under very particular conditions of operation." Neumann, 1963, p. 297-298

For present purposes, most of the quote appears to relate to Marr's algorithmic and hardware levels. However, notice that a physical system doesn't just 'compute'. A *purpose* is implicitly assumed. For, without a purpose of the computational system, how could one talk of a *proper use*? A device is properly used when it is applied to a problem which is from a typical distribution of problems.

For a device to compute—and for someone to *talk meaningfully* about something computing—I interpret von Neumann as supporting the claim that any given computational system must be interpreted with a "proper use" in mind. He continues, providing some conditions that are relevant for understanding this proper use:

> "By an all-or-none organ we should rather mean one which fulfills the following two conditions. First, it functions in the all-or-none manner under certain suitable operating conditions. Second, these

[1] I agree with many of the sentiments in the critical articles in response so far. A number of good points are made in the responses of Pothos and Busemeyer, 2013.

> operating conditions are the ones under which it is normally used; they represent the functionally normal state of affairs within the large organism, of which it forms a part. Thus the important fact is not whether an organ has necessarily and under all conditions the all-or-none character—this is probably never the case—but rather whether in its proper context it functions primarily, and appears to be intended to function primarily, as an all-or-none organ. ”
> Neumann, 1963, p. 298

These two conditions can be applicable to more than just “all-or-none” organs, but to how one constructs the notion of normativity and typicality in computation, (and regulation in a cybernetic system generally). His first condition, that of suitable operating conditions, allows one to define sufficient initial conditions, i.e. a ‘set up’, such that it is possible to follow the meaningful steps (an algorithm) that proceed afterwards. It also allows one to specify why the system *should* function according to the (potentially user-defined) use. This constitutes the *why* of a computational analysis. The second of von Neumann’s condition concerns what it means for a system to have a ‘proper’ use. If we don’t know the typical conditions in which a system should function, we will likely have no clue as to what its proper function would be, i.e. what is being computed. In practice, one may just have to *assume* typicality in order to reverse-engineer the system under study.[2]

For a given device, in order to determine when it is failing to compute, one assumes typicality. In the conjunction fallacy example, discussed shortly, the Bayesian paradigm can regard the problem as atypical, and the failure as irrational. Without the normative framework determining what is typical, one lacks the ability to categorize mistakes. This is precisely the problem that QLMs run into for the conjunction effect, where it is unclear whether they really render it as irrational.

Quantum-like models, on the other hand, would allow for incompatible representations which can result in cognition that is fundamentally context sensitive when it shouldn’t be (by classical lights). There is, as of yet, no story of why human cognition *should* be so. However, I will entertain the possibility that it is possible. In this case, proponents may be justified in their research program. It is clear that the mathematical machinery is at least more general than classical probabilities.

[2] “[. . .] the nervous system is not accompanied by a manual explaining the principles of operation. The blueprints and the early prototypes were thrown away a long time ago. Now we are stuck with an artifact, so we must try to reverse engineer it.”Mead, 1990, p. 1630

7.2 Quantum Basics

The quantum formalism begins with the concept of a vector space. Consider a plane as a two dimensional vector space, V. Let V be a set such that vectors $\mathbf{v}, \mathbf{w} \in V$, represented visually by arrows, can be identified with a pair of coordinates $[x, y]^T$ from an origin (where T represents the transpose operation, making the in-line row into a column vector). We can concatenate elements of V to form a new element of V. That is, $\mathbf{v} + \mathbf{w} = \mathbf{u} \in V$. The '+' symbol could be thought of as addition, but it is more precise to just think of it as a linear combination of the two vectors.

The dimensions of the space are determined by the number of basis vectors required to span the space. A spanning basis is a set of linearly independent vectors which, when stretched (multiplied by a scalar $a, b \in \mathbb{R}$) or combined any number of ways can reach any point in the space. A set of vectors are linearly dependent when one of the vectors can be written as a linear combination of other vectors. Two vectors are sufficient to achieve a basis for a plane. Take the so-called computational basis vectors $\mathbf{v_1} = [1, 0]^T$ and $\mathbf{v_2} = [0, 1]^T$ (i.e. the unit arrows in the x and y directions, at 90° to one another). Convince yourself that every point in the x-y plane can be reached by combinations of scalar multiples of these basis vectors. That is, any vector (or point) $\mathbf{w}$ can be written in the form

$$\mathbf{w} = a\mathbf{v_1} + b\mathbf{v_2} \tag{7.2}$$

Now I move on to the vector spaces used in the Hilbert space formulation of quantum mechanics. This space has an inner product defined in addition to using more general complex numbers $\alpha, \beta \in \mathbb{C}$ for scalar multiplications.[3] Here I follow the simple conventions in the field of quantum computation, by representing a quantum wave function in Dirac notation with column vector basis states defined according to observed classical bits $|0\rangle$ and $|1\rangle$.

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad\qquad |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{7.3}$$

A qubit will be represented by a linear combination ('superposition') of the form

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \tag{7.4}$$

where the complex-valued coefficients α, β ('amplitudes') are supposed to evolve coherently in time evolution of the state $|\psi\rangle$, and determine observed probabilities in quantum mechanical experiments according to the Born Rule

[3] A complex number, in Cartesian coordinates, is a number c of the form $x + iy$, where $i = \sqrt{-1}$. These numbers make up the complex plane, and vectors are defined similar to those above.

calculation $|\alpha|^2 + |\beta|^2 = 1$ (where $|\cdot|$ denotes the modulus of the complex number). Superpositions (linear combinations) and amplitudes are general features of wave mechanics—such as surface waves on water. When two waves overlap, we represent them mathematically as a linear combination. They are 'superposed', and result in constructive and destructive *interference*. The amplitudes in a quantum state, which determine observed probabilities, can also interfere with one another. When using the formalism of quantum mechanics, we will have coherent superpositions of n qubits according to the general form

$$|\psi\rangle_n = \sum_{x=00\ldots0}^{11\ldots1} \alpha_x |x\rangle \tag{7.5}$$

where x is an integer in binary representation. The coherent dynamics of this superposition during time evolution will mathematically appear as a function on each and every one of the 2^n amplitudes. It is a philosophical question what exactly the entities are which the amplitudes α_x and states $|x\rangle$ represent in the quantum world. These amplitudes will interfere, and because of their role in determining probabilities they are sometimes referred to as *probability amplitudes*. This term is at best a convenient reminder, but at worst it is a misnomer. Derived probabilities reflect results in measurements or observations and not properties of coherent dynamics of a state (unless carefully qualified as *counterfactual* properties). One may find the term "quasi-probability" more helpful to avoid the conflation between probabilities and amplitudes.

The qubit is a state represented by an $n = 2$ dimensional Hilbert space $\mathcal{H}$. Measurements (which will determine answers to questions in the cognition and decision contexts) are for present purposes represented as projections $\mathbf{P_i}$ onto subspaces $\mathcal{H}_i \subset \mathcal{H}$, with probability $p(i) = \langle\psi| \mathbf{P_i} |\psi\rangle$.[4] In general, projectors will not commute, i.e. for all $\mathbf{P_i}$,$\mathbf{P_j}$ it will not be the case that $\mathbf{P_i P_j} = \mathbf{P_j P_i}$. Also, $\sum_i \mathbf{P_i} = \mathbf{I}$, the identity projection of the space.[5] We can characterize a measurement device D as another two dimensional quantum system with eigenstates $|\uparrow\rangle$ and $|\downarrow\rangle$, where the interaction between the device and the quantum state $|\psi\rangle$ obey the following transitions after time evolution:

$$|0\rangle |D\rangle \rightarrow |0\rangle |\uparrow\rangle \tag{7.6}$$

$$|1\rangle |D\rangle \rightarrow |1\rangle |\downarrow\rangle \tag{7.7}$$

That is, the device will show 'up' if the state was $|0\rangle$ and 'down' if the state was $|1\rangle$, whatever the state of the detector was beforehand. Thus the

[4] $\langle\psi| = |\psi\rangle^{*T}$, where $*$ denotes complex conjugation.

[5] Projectors are not even the most general form that a measurement can take in quantum information theory. Rather, the most general form is a positive operator valued measurement (POVM). However, the proponents' main claims utilize projective measurements as far as I am aware.

superposed state $|\psi\rangle$ will transition like so (again omitting amplitudes):

$$(|0\rangle + |1\rangle)\,|D\rangle \rightarrow |0\rangle\,|\uparrow\rangle + |1\rangle\,|\downarrow\rangle \tag{7.8}$$

Uncertain beliefs regarding sets of quantum belief states in the quantum modeling framework are called *ensembles* consisting of the set and associated (classical) probabilities $\{p_k, |\psi_k\rangle\}$. Many potential ensembles of quantum belief states correspond to and are represented by density matrices. Pure states are special cases in the density matrix formulation. Coherent beliefs are represented as superpositions $|\phi\rangle = |\psi_1\rangle + |\psi_2\rangle$ in some basis, and these are pure states. The choice of basis is critical in assessing the model. Note that one is not uncertain about whether the belief is *actually* $|\psi_1\rangle$ or $|\psi_2\rangle$ in a superposition, but the belief state *is* the superposition. In other words, it would be misleading to think of measurement on a superposition as reducing uncertainty about two potential states (one could measure in a basis where the superposition is an eigenstate). Rather, this is the interpretation for an ensemble. If $|\psi_1\rangle$ and $|\psi_2\rangle$ are potential states in an ensemble, then we can interpret the probabilities for measurement as uncertainty of which one has been 'prepared'.

A multi-step transformation on a coherent quantum state will not necessarily benefit from a Born Rule calculation (that formal procedure which obtains values interpreted as probabilities from the amplitudes) until at the end when one is ready to make a measurement upon the state in an experiment, since the amplitudes will interfere in the meantime. Such a calculation indicates probabilities for what one *would* have measured at *that time*. However, the term probability amplitude does help capture the difference between how we come up with quantum probabilities compared to classical probabilities.

7.3 Quantum Conjunction Effect

Quantum like models of cognition consist of a 'belief' state-vector $|\psi\rangle$, which can be written as a linear combination in an infinite number of orthogonal (spanning) bases. A pure state will be a point (in this two-dimensional example) on the circumference of the unit circle. Proponents use this to analyze results in psychology such as the famous Linda Problem, where agents seem to process information *irrationally* from the perspective of Bayesian norms of rationality. In these cases of a "conjunction fallacy", respondents in an experiment will judge conjunctions as more likely than one of the conjuncts, which violates Bayesian norms. Respondents are told a story about a woman named Linda or a man named Bill, and asked to rank the probability of certain statements about the character. The prompt for Bill is the following:

> Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics, but

> weak in social studies and humanities. Tversky and Kahneman, 1981

Respondents in experiments of these type widely commit a conjunction fallacy, for various stories and for more or less explicit probabilities. In the Bill example, the probability rankings for three statements are sufficient to show the conjunction effect. These propositions are A): Bill is an accountant; J): Bill plays jazz for a hobby; $A \wedge J$): Bill is an accountant who plays jazz for a hobby. The experiment was constructed so that the story was more representative[6] of A than J, yet the probability ranking judgments of a large majority of respondents (even by students educated in advanced statistics courses) were $A > (A \wedge J) > J$. This violates classical probability norms, since categorically the probability measure of a conjunction of events is strictly less than or equal to the measure of a conjunct.

The same is true for the Linda story, where conjunctive judgments of bank teller and feminist are ranked as $F > (F \wedge T) > T$. In QLMs, there are two bases we are concerned with, where $\{F, \neg F\}$ and $\{T, \neg T\}$ represent in the Linda problem the beliefs that Linda is a feminist or a bank teller. Or, rather, we can express the belief state vector $|\psi\rangle$ either as an uncertainty regarding an outcome of measuring in the $\{F, \neg F\}$ basis, or as uncertainty regarding an outcome of measuring in the $\{T, \neg T\}$ basis. These bases are at an angle Θ with respect to each other. They both span the space.

With this, QLM proponents suggest rendering conjunctions in quantum-like cognition as a procedure of subsequent projections onto basis states. While it is ambiguous which basis to project onto first, they suggest to first project onto the axis that the belief state vector is closest to. One of F or T is 'measured' or cognitively processed first based on the nearest distance from the cognitive state $|\psi\rangle$ (a unit vector) in the space to a basis state. That is, the basis state whose projector results in the highest probability will be evaluated first in the conjunction. This is an assumption which allows the model to reproduce the desired results. In the Linda example, we are to assume that the subject "projects" attention onto F first, being more likely to land on $|F\rangle$. Then, from $|F\rangle$ we project onto $|T\rangle$. The probabilities calculated for this transition can differ from the probability of projecting in the other order, given the angle Θ between the basis rays in the space.

The angle between the T and F bases is also a parameter relevant to constructing a quantum like model, in this case there is a presumption that a $|\neg T\rangle$ state is closer to $|F\rangle$ than it is to $|\neg F\rangle$. This is a way to 'bake in' correlations presumed by the modeler. This is akin to assuming representativeness, as in Tversky and Kahneman (1981). One way to look at quantum like models in

[6] That is, so that it seems like A is representative of the distribution alluded to in the story, and J is not.

this case is to say that they are modeling the way that an individual might give the 'wrong' answer in a single multiple choice question. It is unclear whether one could use the framework to illustrate the *rate* at which the agent chooses one answer over another. Proponents appear to assume that this is sensible. It seems like the correlation angle between the bases should update between trials, and multiple trials must be presumed if one were to speak of the probability of an individual choosing an answer, unless discussing an *actual* quantum-like mechanism which behaves like the quantum world and is liable to be interpreted with a Born-like Rule.

However, the method for this is unclear. If there is no such method, then it is unlikely that the probabilities can be interpreted as a frequency. If this is the case, the only likely alternative is that the numbers are degrees of belief. Then, the two flavors of Bayesianism, where degrees of belief are assumed in the framework, are subjective and objective. If the authors think there are rational constraints telling us how the agent should orient the angle of correlation between the bases, they are objective. If any angle is possible, they are subjective. If the latter is the case, an account of the statistics of a population of test subjects committing the conjunction fallacy is problematic, since one expects a roughly even distribution without some bias in orientation.

Long story short, QLMs are assuming a structure to the problem: the positive correlation between F and T. They are also assuming that a population of test subjects themselves assume this structure for computing the answer to the Linda question.

In the formalism of quantum mechanics, calculating the projection tells us the probability (through Born's rule) that our measurement will result in observing the state represented by the basis projected onto. In a given outcome where one has observed T for example, one then assigns the state of the system to be $|T\rangle$. In quantum mechanics, this formal account of measurement is at best an involved epistemological discussion about how one represents knowledge of the quantum world, and at worst a tangled mess of epistemology and ontology about how, whether, and to what extent the representations reflect entities in the quantum world.

Unfortunately, there is no additional story for why, in a QLM, one should initially represent the belief state vector in the way that we have above. It seems that the reason for representing the state this way is such that the union of a set of its projectors are identity and have associated probabilities sum to one. The reason to do this is arguably so that we remain classically coherent for the measurement characterized by these projectors. This is just what Bayesian normativity tells us, only the way the coherent probabilities are being computed has changed. If we had a story why to represent the feminism and bank teller bases with incompatible observables, one would have a story which compels loss under certain betting circumstances. In other words, it would be a norm which

would be ineffective and irrational for classical tasks.

Furthermore, there is no clarification of why the method of choosing a sequence of projections is not arbitrary. Is there a consequence we want to avoid by first projecting onto the nearest basis? Do we also consider that this projection is due to an attention measurement that was in that basis (i.e. the agent asks himself "Is $|\psi\rangle = \{F, \neg F\}$?")? Why, for example, doesn't one instead first assume that the agent asks herself such a question, and project onto that basis regardless of whether the state has a more likely projection in another basis? Why *should* one model the agent as doing so, and why *should* the system do so?

The importance of providing this story should not be underestimated. QLMs are giving us a *procedure* to recover certain data. They don't tell us why such data should or shouldn't be the case (i.e. what is right or wrong, typical or atypical). To do this, proponents would need a normative component I think they are lacking.

In the case that QLMs are just pure math, there is no reason to associate any quantum concepts—only the math is relevant. Unfortunately, quantum-like modeling proponents do not remain agnostic with respect to quantum concepts—invoking such concepts like measurement, wave-particle duality, and collapse of the wave function. Are all of these concepts justified in an agnostic use of mathematics? I don't think so, as they are heavily dependent on foundational debates over the correct interpretation of quantum mechanics. Even so, it is not clear that total agnosticism about the nature of the quantum belief states precludes certain mathematical operations from happening. For example, if QBSVs are defined, and projective measurements are defined (whatever measurement means in an agnostic context), it seems likely that a decoherence-like procedure is also defined.

7.4 Quantum-like Decoherence

Quantum-like modeling proponents seem to take their models to possibly be a general computational level analysis of cognition, akin to Bayesian cognitive science. They acknowledge that in order to avoid decoherence objections which have been leveled against hardware level theories, they must be at such an abstract level. Otherwise, if the belief states were to exist somehow as physical states, the question of decoherence would rear its ugly head.

The reason decoherence is so devastating for quantum theories at the hardware level, is that the brain is noisy. Any coherent quantum state would effectively be measured by the noise (thermal or otherwise) and subsequently decohere. What would it decohere into? The observables would decohere into states which appear classical.

Given a quantum measurement device, there will only be certain eigenstates that are observed. The interaction between the device and the quantum state being measured is the subject of the measurement problem. What happens to the entity in the world when we update our epistemic state to reflect the state observed in the measurement? Where does the extra information go? Is there a dynamical process which occurs? What happens to the *coherent* wave function?

There are collapse theories, no-collapse theories, many-worlds accounts, many minds accounts, Bohmian accounts, and many variations in-between attempting to reconcile the fact that the accurate description of the quantum state after interaction with a measurement device is an eigenstate of the device. However, what many of these accounts can at least agree upon, is that there is some important difference between the appearance of the classical world (and classical descriptions of classical problems) and the quantum world (and associated quantum descriptions).

Quantum states interacting with other high-dimensional quantum systems can be considered to be in some sense *measured* by these systems. Noise from the environment dampens the coherent phases. Thus, a coherent quantum state can be modeled as receiving so-called *phase kicks*, eventually resulting in a state described by a density matrix with off-diagonal terms equal to zero. (Nielsen and Chuang, 2010, §8.3.6) These off-diagonal terms are the coherence terms, corresponding to varying amplitudes (and thus varying measurement probabilities at different times). In other words, the interaction of the open quantum system will find a basis in which the density matrix will be diagonal, and the probabilities will be constant in time. This is called *einselection*, environmentally induced super-selection (Zurek, 2007). This means that our description is now *mixed*, with the terms interpretable as classical probabilities. Decoherence as a mechanism thus explains why it is appropriate to describe macroscopic systems classically, even if ultimately there is still a coherent universal quantum state. One can thus consider decoherence, at least for agnostic formal models, as a formal mechanism not dependent on corresponding physical mechanisms, since there is only an *apparent* collapse of the wave function.

Thus, if one considers quantum *belief* state vectors as QLM proponents do, one has two options. One could invoke some idealization in which these purely formal objects do not interfere with each other and are isolated. This seems highly implausible, if one transfers the domain knowledge of how a physicist makes sense of the apparent classical world. Alternatively, one could make the models more realistic with QBSVs in higher dimensional belief spaces. This means in turn that either we are dealing with global (or universal) QBSVs—or relatively isolated low-dimensional QBSVs interfere with the rest of an agent's beliefs. In this case, it seems that these QBSVs will decohere. Such an issue has also recently been noted in a review of QLMs, but the importance of the agnostic status of quantum belief states is overlooked:

> Quantum effects are generally not robust when states are allowed to interact with an environment, or when the evolution is otherwise not well controlled. This has two implications; first, it suggests that care might be needed to ensure cognitive states remain quantum during experimental manipulations. Failure to do so could mean no quantum effects are visible. Second, it suggests an explanation for why some cognitive variables do not show quantum effects, perhaps certain preferences/beliefs are just to hard to isolate, and the inevitable interaction between them and other thoughts quickly kills off any quantum behavior before it can be observed. This is a worthy subject for future research.
>
> More generally, this shows how in some sense CP-maps allow one to move smoothly between pure quantum and pure classical Markov models. Evolution under open system dynamics may kill off quantum properties of a system, and the long time dynamics may be essentially classical. Yearsley, 2017, p. 36-37

This is not so much an interesting area for future research, as it is an achilles heel for quantum-like models in cognition. It is not possible to have truly idealized platonic quantum belief states which one is agnostic about, while simultaneously invoking a quantum-like measurement procedure (attention measurements in judgement). They are incompatible since the mechanism of measurement *is* the mechanism of decoherence, and I can argue for a no-collapse version of the dynamics of quantum belief states. Decoherence is a transition of the same form as measurement. For a two-dimensional environment E, we can characterize its behavior just like a measurement device. Decoherence can just be thought of as measurement by the (also quantum) environment in a higher dimensional system (see again Zurek, 2007 for more details):

$$|\psi\rangle\,|E\rangle = (\alpha\,|0\rangle + \beta\,|1\rangle)\,|E\rangle \rightarrow \alpha\,|0\rangle\,|E_{\uparrow}\rangle + \beta\,|1\rangle\,|E_{\downarrow}\rangle \tag{7.9}$$

A similar transition can be written also in the case that $|\psi\rangle$ is a combined state plus detector as in (7.8). All three quantum systems (environment, detector, and the original state) are then irreversibly correlated with one another. The result when the environment dominates the interaction (i.e. it is represented on a higher-dimensional Hilbert space) is a state classically describable by a reduced density matrix.[7] The interaction between the systems has reduced off-diagonal coherent quantum terms, and we can consistently talk only using classical probabilities of a decohered state.

[7] "[...] irreversibility could also arise from more familiar, statistical causes: Environments are notorious for having large numbers of interacting degrees of freedom, making extraction of lost information as difficult as reversing trajectories in the Boltzmann gas." Zurek, 2007

The formally agnostic 'dynamics' of quantum models simply contain the mechanism for decoherence from the get go. We would thus expect that non-isolated belief states will almost *always* decohere under realistic scenarios. The result is that the state can be described by a reduced density matrix, where diagonal terms are classical probabilities. Predictions and judgments of the quantum-like cognition system not only *should* obey classical principles, but if correct in the above argument, even the agnostic formal models just *will* result in classical cognition. That is, unless something goes wrong and the incompatible representations are allowed to persist coherently. Agnosticism about the ontological or implementational status of quantum belief states seems to smuggle in the coherent magic of quantum mechanics without paying the toll of decoherence. Unfortunately, if you want to bring *measurement* into your modeling framework, then decoherence is your baggage.

That said, belief state vectors which have not decohered may be one way to understand failures of cognition. Mistakes are made because the state has not 'resolved' into a classical state, and thus for the purposes of a classical task (which are most typical!) the cognitive system is 'uncertain' when it shouldn't be. This is the best outcome for QLMs of cognition: it is a quantum-like model of *failure* to meet classical norms. Why should quantum models be taken as preferable over other alternative modeling frameworks which can also recover classicality *and* illustrate failures? Quantum models are hardly more parsimonious.

In the cognition and decision context, an agent makes a judgment on a set of (perhaps uncertain) beliefs. Realistically, the set of beliefs is much larger than beliefs about specific propositions in the agent's immediate sphere of attention.[8] That is, lets say that for a task an agent relies explicitly on k beliefs. Plugging this into the quantum modeling framework, the set of online beliefs is represented by a quantum belief state vector $|\psi\rangle \in \mathscr{H}_k$, where $\mathscr{H}_k$ is a Hilbert space of dimension k. For *all* beliefs, we would have a universal belief state $|\Psi\rangle \in \mathscr{H}_N$ where the dimension of the space is larger than the beliefs which are online $N > k$. One can imagine that, even if judgments are not made on $\mathscr{H}_N$, they in general may be made on $\mathscr{H}_m$ such that $k < m < N$.

A superposition of n qubits (two-state quantum systems) will have 2^n amplitudes. More quantum states in superposition increases exponentially the complexity of coherence. The risk of decoherence, and of noise affecting the coherent system, becomes exponentially larger as the size of the system increases. The same is true of high-dimensional QBSV for quantum-like models. Only idealized and isolated quantum belief state vectors will not decohere under realistic assumptions of the complexity of online beliefs. Imagine a platonic idealized fluid bath where waves never dissipate.

[8] Attention is anyways considered by QLM proponents to be like projective quantum measurements.

This presents a problem for a model of the world, or representations of beliefs about the world, based on quantum representations. These representations will become unnecessarily cumbersome as the number of simultaneously held beliefs or features in a model increases. While again this may not be an issue for quantum belief states in an unrealistically ideal realm, it is an issue for any algorithmic or hardware level account. Just grant for the sake of argument that it is plausible to have quantum-like mechanisms in an ideal realm, and ignore for the moment implementation levels. I emphasize again the point being made here, in case it is not clear: ideal states interfere with each other as mathematical objects, and this is sufficient to show that a realistically high-dimensional belief set *will* decohere. It also *should* decohere.

7.5 Classical Norms for the Classical World

Classical (Bayesian) norms stay intact even in quantum like models, since one still *ought* to represent events compatibly. Decoherence still *should* occur. The conjunction fallacy is still an *error*. There are issues with the way proponents anticipate this exact issue:

> "Bayesian models of cognition are claimed to be founded on a rational basis (Oaksford & Chater, 2009). In fact, Bayes' rule is derived from Kolmogorov's axioms for classic probability theory. Quantum models of cognition are based on von Neumann's axioms, which reduce to classic theory when all the variables are assumed to be compatible. So why do we need to use incompatible events, and is this not irrational? In fact, the physical world obeys quantum principles and incompatible events are an essential part of nature. Nevertheless, there are clear circumstances where everyone agrees that the events should be treated classically (such as randomly sampling balls from urns). However it is harder to argue what is rational for cases like the Linda story, because one cannot refer to any empirical relative frequencies for a singular or unique event. Furthermore, it remains an empirical question whether quantum or Bayesian methods are more useful for modelling probabilities of very complex sequences when the joint probabilities are largely unknown." Busemeyer and Bruza, 2012, p. 141-142

The statement above is typical of the mistakes made concerning quantum like models and their relationship to normativity and the computational level. First, yes—Bayesian models are founded theoretically on considerations of rationality. They are founded on *normative* considerations of rationality. It isn't just that Bayesianism states some nice relationships that degrees of belief might have—but that an agent (living in a classical world) *ought* to display and be

consistent with those relationships *or else*. The idea is not just that someone else thinks you ought to comply with these rationality constraints, but that if you don't one can demonstrably show that you yourself will suffer a loss you could have otherwise avoided.[9] The authors seem to be aware of this, yet do not offer any consolation for those who might concur with the sentiment that to represent non-quantum events incompatibly seems irrational.

A second issue with the above statement is that without some clarification, it is more or less irrelevant that "the physical world obeys quantum principles and incompatible events are an essential part of nature." One might interpret this as implicitly arguing for physical reductionism, in the same way that one might for an argument from physics for revising logic (discussed previously). However, according to the proponents, one is supposed to only be considering the formalism of quantum mechanics independent of any physical mechanisms which may or may not be present in the brain—which demolishes the soundness of such a revisionary argument.

If the physics of the brain doesn't matter, why should it matter in general what the physical world is ultimately like at the most fundamental level? It seems more relevant to talk of the physics of the brain than it does to comment about the physical world generally—yet proponents are avoiding the former while seemingly using the latter to support their project. What is relevant, however, is that the world which makes up the typical types of problems that a cognitive agent encounters is *classical*, not quantum. It appears decohered.

Third, just because it might be difficult to come up with Bayesian models for cases like the conjunction fallacy doesn't mean we need a new computational level framework—even if QLMs provided one. We could rather characterize these cases as *non*-typical applications of the computational level framework. QLMs may, however, give an algorithmic level story.

7.5.1 Algorithmic Comparison

The focus on sequences of steps by the authors indicates strongly that the proper level for QLMs is *algorithmic*:

> "A compatible representation requires us to assume that it is meaningful to assign probabilities to conjunctions of events, and an incompatible representation assumes that this cannot be done, and instead we need to assign probabilities to ordered sequences of events."Busemeyer and Bruza, 2012, p. 37

In this section I compare the kinds of steps found in QLMs as well as algorithmic level accounts implementing BDT. One see that the kinds of steps

[9] Again, the first idea is external while the second is internal to the foundations of Bayesianism.

involved in QLMs are directly analogous to those more familiar steps already found in algorithmic implementations associated with Bayesian cognitive science. Thus, I will conclude that QLMs are an algorithmic level framework.

Part of the problem contextualizing alternative approaches to cognitive modeling might be that aspects of Bayesianism happen to span both the computational level (with its normative story) and the algorithmic level with updating dynamics. The updating dynamics are *also* algorithmic—as well as any number of alternative methods for providing the correct relationship between probabilities in a distribution (i.e., fulfilling the normative constraints). Thus, it might seem that if quantum models can also do some similar things with updating, they might also be at the computational level. In a certain sense, they are, but *only* because classical probabilities and compatible events are special cases of the Hilbert space formalism—and there is a normative story compatible with this *sub*-formalism.

Quantum models piggy-back, if you will, onto the normative constraints already present in Bayesianism—while simultaneously arguing to compete with and, ultimately, replace it. This would only pull the rug out from under their project, rendering them without even the special-case computational-level support that classical norms supply to the formalism. The computational level is not about sequences of calculations, it is about what *should* be calculated by *any* algorithm, and *why*.

To broadly recall the steps in a QLM of the conjunction fallacy: First, one chooses two bases representing (incompatible) observables (answers to questions). Then, one assigns a belief state vector (a superposition or linear combination of a basis). Assume that this vector is not equidistant (uniformly superposed) from all relevant events, and an answer will be given by projecting (with comparably higher probability) onto nearest axis representing one of the variables in the conjunction. Then, project onto the axis of the other variable. The probability of this projection is given by the squared modulus of the projector on the belief state vector. The entire transition probability can be greater than a single projection. This protocol is clearly algorithmic in nature. Take the following statements as evidence:

> "There are several ways to justify the assumption that the more likely event is processed first. One is that the more likely event matches the story better and so these features are more quickly retrieved and available for consideration. A second reason is that individuals sometimes conform to a confirmation bias and seek questions that are likely to be confirmed first."Busemeyer and Bruza, 2012, p. 124

These statements belong to the algorithmic level, in particular the level of algorithm specificiation, while the projections used to carry out these intuitions

are at the second algorithmic level in Marr's hierarchy. They are particular *mechanisms* that carry out the specification of the algorithm. The following table summarizes the similarities (denoted '::') between algorithmic aspects of these protocols and the kinds of steps in algorithmic BDT. The aspects range over both algorithmic levels.

Determine Event Space	::	Fix Hilbert Space Dim.
Choose Priors	::	Assign State Vector
Check Coherence	?	Represent Events Compatibly
Update on Information	::	Subseq. Proj. onto Subspaces

A given algorithm implementing BDT may, for example, choose a uniform prior distribution. It is a good example of a particular specification of how to choose priors.[10] Similarly, there may be multiple ways to update coherently. BDT does not, at the computational level, tell us how to do this—only that one should obey the probability calculus when doing so.

The thing is, Bayesians can construct an algorithmic account of the conjunction fallacy just as well as QLMs can. For example, one just picks non-coherent sets of belief. The contention between these frameworks is whether they both consider the decision behavior *irrational*. It would seem that both *do*, since it is doubtable that QLM proponents will want to say that conjunction fallacy behavior is *rational*. However, it should be mentioned that a non-coherent set of beliefs is *outside* of the prescriptions of BDT, whereas it is not outside the prescriptions of idealized QLMs. If this were a feature of a computational level theory, then it would be a *bad* computational level theory. What is more reasonable is that QLMs are primarily algorithmic level accounts.

7.5.2 A Quantum Agent Should Have Compatible Beliefs When Possible

One could say that the only difference in a normative story for QLM is that normal Bayesian normativity only holds in the special cases where events are compatible. But one could go a step further and reconstruct the Bayesian norms by requiring that an agent *should* represent events compatibly wherever possible. Their beliefs *should* decohere. In my exposure to proponents of quantum like models, I have not found sufficiently clear and emphasized statements of their view on rational normativity in their model. However, a few statements indicate that any contribution they might have is in terms of an agent's ability to represent events compatibly.

[10] Some Bayesians may not be satisfied with just the coherence requirement, and thus they may stipulate even more normative conditions on the algorithmic implementation of BDT. For example, they could say that one *should* utilize a uniform distribution of priors. Something along these lines is what is meant by the so-called *principle of indifference*. Another example of a normative condition on priors is that *if* there is an observed frequency among multiple samples of an event, one *should* have a prior probability corresponding to this frequency.

> "When all the events are compatible, quantum probabilities satisfy the same properties as classic probabilities and meet the same rational standards in this restricted case. The main question of rationality arises when incompatible events become involved. "Busemeyer and Bruza, 2012, p. 347

It might be the case that people *don't* always fulfill optimal rationality requirements (i.e. Bayesian coherence protecting against Dutch Books). Furthermore, one could grant that the QLM approach could be a fruitful way to model 'non-rational' behavior. Still, this is consistent with and does not replace the Bayesian normative story. However, there might still be something to say about normativity for quantum-like modeling in that an agent *should*, whenever possible, represent beliefs as compatible with one another. When this is fulfilled, traditional norms follow.

When beliefs are incompatible, one can always construct a Dutch Book, or show probabilistic fallacies such as the conjunction fallacy. It is also possible to construct a Hilbert space model which captures this data. This is the most plausible results concerning normativity. Simply put: agents *should*, when able, represent events compatibly. Otherwise, a violation of *classical* normativity and classical losses can occur. In other words, then a Dutch Book can be made against the agent. In those cases where the agent does not represent events compatibly, it is possible to construct a quantum model for the agent. QLMs utilize a strictly more general framework, and thus they are capable of modeling at least as much as other models realizing BDT. Still, as noted above, work has to be done on justifying the quantum stories.

This being said, however, one still needs some independent justification as to *why* these concepts are incompatible (or rather, that they should be represented by incompatible observables). It does not seem like bank teller and feminism are incompatible concepts incapable of defining together a joint event. It seems as if incompatible is being defined post hoc, for those cases where classical violations occur. If this is the case then QLMs can no longer *explain* the conjunction fallacy since incompatibility is being presumed in order to make the model fit. In other words, one cannot conclude that feminism and bank teller are incompatible if this is being assumed to begin with. However, it seems that these concepts *are* compatible in general, which is why respondents are irrational (and judge themselves as irrational) when committing the conjunction fallacy.

Furthermore, the normative aspect of BDT simply *implies* compatibility (i.e. decoherence) in the algorithmic handling of events in the first place. Since if an agent does *not* represent events compatibly in processing, sure losses may ensue.

If QLMs were at the computational level, the only way to justify their status at this level would be to reproduce the exact same normative or problem-oriented statements as BDT. There are not sufficient reasons to suppose that

QLMs can be construed as providing an alternative framework at the computational level. Not only do the quantum stories involved in QLMs not establish any further normative considerations for a computational level reading, but they are formulated almost solely at the algorithmic level. Agents committing the conjunction fallacy discussed above *should* represent events compatibly, and this corresponds with the normative story already present in Bayesian decision theory with classical probabilities. As argued, even if it is granted that QLMs are at the algorithmic level, it is still plausible that classical normativity emerges from the quantumness through decoherence. Thus, in the end it seems like quantum models—properly transferred—bolster a classical analysis at the computational level even further.

7.6 Decoherence and Survival

Proponents of QLM try to avoid the decoherence objections that plague hardware level claims of quantum processing in cognition, since they seem to see themselves at the computational level where there is no quantum mechanisms involved (and thus no interactions which could lead to decoherence). A decoherence objection is still possible for an agent which processes information in a quantum-like manner at the algorithmic level. Decoherence illustrates the improper (or incomplete) mapping of quantum theory involved in the modeling procedure by QLM proponents, and where traditional assumptions about typicality in cognitive tasks break down. Considering a cognitive agent, one can see how decoherence is actually related to survival in a crucial way.

Take, again, a physicist as an example agent. This agent has to make decisions and answer questions as reliably as possible given a variety of *typical* physical propositions as inputs. These inputs are all in the realm of classical physics. They are macroscopic situations and problems, the optimal solutions of which will be dictated by classical physics. For example, questions about large objects at low speeds. The success and performance of this physicist are directly correlated (lets say) with her survival. Doing poorly on certain physics problems will negatively impact her survival.

One can formulate this in terms of regulation, and think of the agent as a cybernetic regulator. See for example a useful introduction in Ashby, 1958. A cybernetic regulator is a complex system tasked with mitigating disturbances (or solving problems) such that a regulatory goal is achieved. The variance of problems encountered should be reduced by regulatory actions, and ideally result in a subset of possible outcomes that align with the regulatory goal. One says that this regulator will have a *model* of the environment, and importantly that this model will be largely determined by *typical* environmental inputs (or

problems). It will be more successful at regulating if it is optimized for the typical class of regulatory tasks or inputs.[11]

If a physicist has a regulatory model based on classical physics, he or she will do pretty well to regulate (i.e. answer correctly) classical problems. If this agent has a regulatory model based on quantum physics, he or she will do well *only* if decoherence is applied first. Imagine, for example, trying to regulate classical problems (such as a bear attack) under the assumption that each macro-state is in a coherent superposition. Is the bear dead and alive at the same time—or is the dead bear in an alternate branch of the universe? This is of course a rhetorical simplification, but the point is that interpretational debates and quantum weirdness still matters for a psychological theory. The issues arise due to nature of the concepts and formalisms in quantum theory, not the order of magnitude at which they apply. Physicists use decoherence as a story to explain why the world appears classical even though it is actually still coherently quantum.

It seems unlikely that an agent whose psychological method of representing and processing information defaults as *quantum* would survive in the world. Or, survival would be worse off in comparison to a regulatory model which reflects the classical structure of typical inputs. As noted, when this processing does *not* coincide with a classical regulatory model (for example, when violating Bayesian norms) there are real world consequences. The classical world cannot be made sense of by a quantum regulator absent some processing technique which respects classical norms.

Thus, the regulatory model—if it is to be quantum—must decohere in order for it to be effective. The quantum-like processing must yield classical results. The survival of the system would be compromised otherwise. Or, one could say, the regulatory model would be unnecessarily complex. This observation comes from Conant and Ashby, 1970 and the so-called "good regulator theorem". The good regulator theorem illustrates why a physicist should rather switch between quantum and classical regulators, instead of utilizing a quantum regulator for all tasks. Practically, quantum descriptions are cumbersome and highly complex for macroscopic systems, even when decohered. In principle, a regulator operating on these descriptions could be optimal, however it will be unnecessarily complex for classical problems.

This provides us with a good way to understand how a non-decohered (i.e. quantum) regulatory model for classical tasks would be inefficient *at best*. Classical events could be treated compatibly, agreeing with classical norms, but the regulator would still be unnecessarily complex (and therefore expending more resources). Of course, in the platonic realm of mathematical formalisms, one might not care about such resources. The other case is worse, where events

[11] I am sidelining for the present any considerations of overfitting or over-regulating and generalizable adaptability.

are represented incompatibly leading to violations of the classical norms. The conjunction fallacy example discussed above is a prime example. In survival terms, this is clearly worse because it leads to sure losses in typical classical tasks. Again, algorithmic level mechanisms may well be quantum-like, but the computational level should match the structure of the distribution of tasks in the classical world.

In the discussion of decoherence, one can see that the structure of quantum theory itself involves a story of how the classical world appears at higher orders of magnitude. Quantum theory, if it is to be a physical theory applied by a physicist, applies even to macroscopic objects—*given* a story about how to relate coherent unitary dynamics to what appears in macro-level approximations. Compatible observables will be used by the physicist to characterize the vast majority of the class of problems for large objects at low speeds. Arguably, the typical class of problems to be encountered by the (typical) cognitive agent is much closer to this class of classical physical problems than it is to the set of quantum physical problems. This is the class that the system is evolved to regulate.

Now, one may expect a response along the following lines. It could be that psychological beliefs about various statements are indeed incompatible and elicited via non-commutative projectors. The urge to think of this as saying something about the computational level of the system must be resisted. As repeatedly noted, a system whose regulatory model treats sets of beliefs in this manner will be subject to penalties from the environment. These penalties are not only avoided by a more plausible (and efficient) classical model, but such a classical model matches the structure of quantum theory as applied by a physicist concerned with responding to and regulating tasks in physical science. A regulatory model concerns what should be regulated, how the system *should* respond to a variety of inputs in order to achieve the regulatory goal—i.e., survival.

The above discussion takes significant inspiration from the methodology in philosophy of physics. It also illustrates that there are nuances and concepts that have been neglected in the transfer of the theory and formalism of quantum mechanics to the decision and judgment contexts. Thus, the flawed methodology employed by quantum modeling proponents can be understood as improper model transfer.

For example, algorithmically it doesn't allow us to define what is typical or atypical for the cognitive system to compute. This is because it doesn't allow us to categorize what is 'rational' or 'irrational' in any novel way (aside from fulfilling classical norms). It misidentifies—or, even worse, is agnostic about—the proper kind of error or uncertainty that the system should be concerned with. For example, it doesn't give us an account for *why* quantum uncertainty should be minimized as opposed to classic disutility.

Furthermore, there *are* accounts concerning the minimization of classical error. There are accounts for *why* an agent should represent events classically, and why certain classical rules should be followed. The Dutch Book account, perhaps the best known, spells this out in behavioral terms. There are other accounts, such as proper scoring rules. These are normative accounts which provide the context for what *should* be reasoned, or computed, by an agent.

It only *sounds* nice to note that quantum theory (and associated mathematics) is made to handle uncertainty, and does so 'better' than e.g. classical probability. Uncertainty in quantum theory is not the same kind of uncertainty that we are concerned with in normative aspects of computational level theories for cognitive systems.

> Our senses did not evolve for the purpose of verifying quantum mechanics. Rather, they have developed in the process in which survival of the fittest played a central role. There is no evolutionary reason for perception when nothing can be gained from prediction. And, as the predictability sieve illustrates, only quantum states that are robust in spite of decoherence, and hence, effectively classical, have predictable consequences. Indeed, classical reality can be regarded as nearly synonymous with predictability. Zurek, 2007

Uncertainty in quantum theory is about knowledge of quantum states, and knowledge of properties of quantum objects. Uncertainty in cognitive science is about knowledge and beliefs of everyday (classical) states, and the knowledge of properties of everyday objects. There is no apparent reason to suspect that cognitive systems should minimize uncertainty *about quantum states*.

Only recently have scientists even begun to understand macroscopic quantum behavior—not to mention that we are barely a century into the quantum revolution. Even though there now exist macroscopic systems set up to exhibit quantum-like properties, it hardly bears at all on the kinds of things the human system encounters regularly.

Of course it *might* be the case that the *way* in which a cognitive system computes at the algorithmic and hardware levels is minimizing uncertainty in a way that one might characterize as similar to that in quantum information theory. Certainly, though, one would not expect that the reasons *why* the system computes in this way are tied to the way in which cognition functions. The first could just be an artifact of physics, chemistry, or biology—while the second might be linked to evolution and the need for the entire system to survive (and not just the isolated computational parts, i.e. neurons).

Quantum models cannot do better for the typical class of inputs, and proponents would have to argue that what was once the typical class of inputs encountered by the regulatory system is no longer typical, or that the system is

built to handle the atypical (hence why they wish to account for classical fallacies such as the conjunction fallacy). The fact remains that the system might *not* be able to handle the atypical, because the regulatory system is adapted to handle the *typical*.

7.7 Conclusion

The structure of quantum theory implies that a decoherence interpretation (and objection) can also apply to quantum-like processing models. This can actually be construed to support the appropriateness of classical norms at the computational level. In short, it is difficult to make evolutionary sense out of the idea that our cognitive system minimizes uncertainty about quantum-like objects—the effective regulation of classical inputs would likely be compromised, due to unnecessary complexity or assured losses, and effective regulation is what it means to *survive*.

Quantum like models in cognition and decision theory are, at the very best, taking for granted the established and thoroughly discussed normative character of classical (Bayesian) cognitive theory. At worst, they claim to overwrite classical theory without recognizing that what has to be overwritten is also that which would grant them computational level status as construed here. This can be explained by the improper modeling transfer of the structure of quantum theory, and understood by examining decoherence as one crucial aspect which is not transferred from the quantum theoretical domain correctly.

Using the mathematical structure to model the agent's judgements and decisions might be appealing, but methodologically it represents a poor example of transferring knowledge from models of physical systems to models of cognitive systems. A scientist might model an agent using the mathematics of (some) of quantum mechanics, but the agent should not use quantum-like models as internal models. Stated another way, any internal model which is coherently quantum-like should decohere. If it doesn't, it is an unnecessarily complex model (or regulator) which compromises the agent's survival.

Earlier in this dissertation I laid out a pragmatic view of analogy—it is effective transfer of control via a mechanism similar to perception, applying an associated model. So perhaps quantum-like models are just pragmatic, and the formalism just transfers well because it 'works'. This is a fair point, but I argue it ignores Bertalanffy's regulative ideal against superficial analogies. I contend that, while perhaps not superficial, quantum-like models are certainly not *homologous* in their application to human cognition. This I have argued can easily be seen by considering an idealized physicist who must 'regulate' physics problems in a lab. Such an agent is, after all, the kind of agent for which the mathematics of quantum mechanics was developed—as a regulatory tool for

interacting with the quantum world, to solve quantum problems. A physicist uses classical models for classical problems, not quantum models.

Chapter 8

Towards a Structural Systems Theory

> What is "structure"? Bourbaki has undoubtedly given the answer, but what does it mean in my terms? Ashby, 2008, p. 5305

I began this dissertation by introducing cybernetics as a systems theory framework focused on models and structure and which sees knowledge transfer as the transfer of control capacity. I followed the thread of model-based transfer from computation to connectionist networks. Then I noticed how artificial general intelligence is like the automation of a general systems methodology. In the end I brought the discussion back up to the level of scientific reasoning with examples knowledge transfer (structural mapping). As I have demonstrated throughout the previous discussions, influential accounts involve an inescapable appeal to *structure*. There is clearly a structuralist attitude present in the foundations of relevant systems sciences that should be explicitly acknowledged. This is especially clear in the work of Ashby. If general systems theory was not originally primarily concerned with structures, it is at least plausible to forge a modern alternative which is. I argue, however, that that the way in which it has been explicated by its most influential proponents is at heart a structural view of systems. I now conclude by suggesting that a structural approach to analyzing knowledge transfer in cognitive systems is warranted, and much of the analysis of knowledge transfer I have presented would not be possible without diverging from the neo-mechanist and anti-representational takes to systems analysis.

First, as a recap, General System Theory (GST) as outlined by Bertalanffy (1969) is a philosophy of science framework for analyzing reasoning in scientific practices, particularly in non-fundamental sciences populated with modelers whose subject matter are complex systems. GST also aims at developing an outline of a methodology for how one should reason about classes of systems, particularly using mathematical tools which can describe many different kinds of systems. In this respect, it focuses on the structural and formal similarities between models of systems. This methodology may implicitly be found widely in practice, and GST is not wholly distinct from what are called dynamical

and complex systems theory today. Cybernetics is, likewise, relevant for the foundations and history of control theory, where feedback and error control are central. I have established that this systems approach appears to be in contrast to what Craver (2007, §4) calls the "systems tradition", which he references when building his mechanistic view. His overview of the systems tradition lacks any reference to Bertalanffy or Ashby, where there is a recognition of the importance of the kinds of representations and structures that modelers use. I think the term *structural* systems theory may be more accurate in signifying what Bertalanffy and Ashby intended.

If not for Bertalanffy, certainly I think it is quite plausible that a structural view of mechanisms was intended by Ashby. I can offer a brief glimpse of what I think Ashby had in mind for outlining a structural systems theory. For Ashby and the foundations of cybernetics, as a paradigm example of a GST, his appeals to the Bourbaki group of mathematicians make this re-branding justified. Aspects of set theory and group theory are discussed by Ashby throughout his writings, including in his books *Introduction to Cybernetics* and companion *Design for a Brain* (see for example Ashby (1960, §19)). Modern structuralism, group theory, and what is now category theory, although useful, might however be a bit too much for what Ashby thought to be the *fundamental* notion of structure for systems.

Of primary importance to Ashby's notion of structure are the intertwined concepts of transitions and constraint. As I have noted, this involves an explicit appeal to Shannon's communication theory. Constraint is the non-random patterns which appear in a message or signal, and quantified by an entropy measure. The system is fundamentally composed of state transitions, and an analysis of these transitions by appropriate manipulation will display some constraint (information). This information, when fed into a regulator (a cognitive system), can be learned or adapted to. Then there has been a transfer of constraint from the environment into the system.

> It seems that this *structure* (e.g. of linkages between departments) is essentially a constraint. What I suspect is that when [the] Environment presents, in succession, samples from a *class* of problems, so that constraint or restriction is present in the problems, the system, if it seeks optimal conditions for solving, will develop a corresponding constraint, which is the structure. [...] Ashby, 2008, p. 4514

For Ashby, structure is fundamentally *constraint* in information theoretic terms. There are parts and organizations and behaviors of a system in the world, and if upon disturbance of the system the resulting output is *constrained*, then we can speak of some structure. This is because our representation of the system—just in 'black box' terms of inputs and outputs—repeatably (or reliably) reduces information (surprise). That is, a measure like entropy

$-\sum_i p_i log(p_i)$ of probabilities p_i of transitions in the 'channel' tends towards zero. This is because there is structure *there in the world* if there is demonstrable constraint there. This also means that the system is *controllable* in principle if each input to the system can be reliably manipulated. Thus, a strictly information-theoretic analysis of inputs and outputs, and a characterization of state transitions, is fundamental to the notion of structure employed in structural systems theory. Only then are category and group theories going to enter into the picture, and whatever other mathematical structure a modeler wishes to use to *explain* the structure or information which is in the world. Given the importance of knowledge transfer in cognitive systems, I think we should be able to explain the transferability of models by whatever formalisms are utilized.

After discussing how a class of military problems were approached in World War II, Ashby says that "[...] *different classes of problems demanded different organizations.*" (Ashby, 2008, p. 4515) Ashby then brings us back to the concept of an idealized 'designer' imparting structure to the system.

> The next question is how this structure is being arrived at. Is the designer to lay it down initially? Notice that this structure implies selection (from other structures) and it involves a setting of parameters. If it goes far enough it will finish with the designer specifying *everything* in [the] 'cortex' and leaving the cortex nothing to do. In other words, the designer, if he puts in part of such structure, *is putting in part of the solution.* Ashby, 2008, p. 4515

For our purposes, the designer can be thought of as a data scientist, a scientist or modeler, or just the previous experience of a cognitive system. He continues:

> If the environment puts up a problem whose solution is part known and part unknown, a designer may, if he wishes, shorten the machine's work by building in a structure that is suitable for the known part, leaving the machine only the unknown part to work out.
>
> If the designer has done this, then to be precise we must define whether the 'designer' is or is not to be considered in the 'machine'. If 'in', the knowledge that went to the building of the permanent part must have been acquired earlier; so the *real* start of the solution (the designer's training) was much earlier than the nominal switching on of the machinery.
>
> The designer can, if he pleases, extend this 'helping' of the machine to any extent. He can, for instance, help it all the way; in which case the machine is left with nothing to do but carry out the programme. This is the case of the ordinary computing machine; it

> solves nothing; all the solving was done beforehand in the designer's brain, and constructed in as programme.
>
> The designer might, it is worth noting, stop just short of full programming. He could then build a machine capable of 'discovering' the calculus, say. He could build all the machinery so that the only thing left to be explored was whether a certain sign was positive or negative. The machine would make this trial in a microsecond and would make the decision. [...]
>
> So if we inspect a machine and find that the 'cortex' shows a strongly built structure such that the further trials are taking place within its constraint, we can deduce that *some partial knowledge has already been obtained.* [...]
>
> If the environment throws up a series of problems having common features, the results of the earlier attempts may provide partial knowledge for the later [attempts]. This partial knowledge can be used for the construction of a (semi)-permanent structure for [dealing with the problems]. Ashby, 2008, p. 4516-4518

This, I argue, is the foundation for the notion of knowledge transfer offered in this dissertation. It brings together a model-based notion of computation, the transfer of parameter values in connectionist networks, and why analogical cognition is so closely linked to discovery. The structural foundation for knowledge transfer in cognitive systems provides a coherent account across multiple levels of modeling and explanation. Ashby provides a very succinct summary: "*Built-in structure in an adapting system implies that part of the solution has already been obtained.*" (Ashby, 2008, p. 4519) Thus, knowledge transfer in cognitive systems is effective because it utilizes built-in structure. It is effective at controlling for, or adapting to, novel problems or situations which arise.

As one comment for future work, discussions by Bertalanffy (1969, §4) and Ashby (1991a) distinguish between two methodological approaches for a general systems theorist. The first is characterized as 'empirico-intuitive', where a modeler goes from one case to another, perhaps noticing similarities of a new case to some familiar model. The second approach is more 'deductive', defining some framework (or structure) which applies universally to a large class of systems. Bertalanffy characterizes Ashby (1958) as an example of the latter, but more work on these methodological and modelling schema is warranted. Perhaps in different situations different approaches should be (or are) utilized. Discovery for example may proceed more along a case by case basis.

I suspect there are many details about the structural view of systems that are still to be rediscovered, particularly in Ashby's work, and to be developed further with a modern perspective (and modern tools). I think that at the very

least, if we want to discuss knowledge transfer in cognitive systems as I have in this dissertation, such a view is worth researching and developing further.

8.1 How Explanations Enable Control

> Thus science works, in a sense, uphill; for it persistently rejects the primary, vivid, knowledge and seeks the colourless, complicated, patterned knowledge that is communicable. Ashby, 2008, p. 5282

Finally, this dissertation is about knowledge transfer, and began with a discussion of a prominent account of explanation. Perhaps the structural systems view, and the account of knowledge transfer as transfer of control, can shed some light on the debate over explanation.

Explanation in neuroscience and cognitive systems theory may involve several levels of explanation from multi-level causal mechanisms, as has been argued by Craver (2007). Furthermore, following Piccinini (2015), some of these mechanisms may specifically be computational mechanisms. As mentioned in the introduction, a major drawback I find in these positions is the ontic notion of explanation present in particular in Craver's work, and the anti-representational view on computational and cognitive systems. Like Wright (2012) and Wright (2015), I find the ontic notion of explanation is just a misconception which is not useful. An alternative account of explanation for a systems theorist is warranted, and can be found just by dropping the ontic distinction in the first place. Again, Craver already links his project to the notion of control. By dropping the ontic notion, Craver's picture then becomes more or less consistent with the systems theory tradition I have sketched.

In the introduction I discussed knowledge as the capacity to control and make predictions. Knowledge transfer is then defined as the transfer of this capacity from one domain to another. In other words, we can consider that the capacity is *communicated*. An explanation, I have argued, must also be communicable if it is to be useful. Explanations are a means of communicating knowledge (as control capacity) from person to person. The whole point of this pragmatic notion of explanation in systems theory is that, given appropriate means, an agent on the receiving end of an explanation could in principle use the information communicated. Explanations *represent* some information relevant for the regulation (control, prediction) of a system. Explanations enable control, as Craver (2007) often notes, but they do this in a way that is inconsistent with the ontic picture. In other words, ontic explanations do not enable control unless they are explanatory 'texts'.

I see the rejection of ontic explanations as just an obvious amendment to Craver's picture. Nearly everything else in his view, so far as I am aware, is totally consistent with the structural systems picture I advocate for. The

important point is that by describing multi-level causal mechanisms in an explanatory text, the capacity to control the system is made communicable. It is just incorrect in my view to conclude that explanations (in neuroscience) are not arguments, as Craver does:

> To explain a phenomenon, it is neither required, nor is it enough, to show that there is a strict law of nature (one that is universal, deterministic, and insusceptible to failure) between the variable and the *explanandum phenomenon*. Such subsumption under laws would be necessary if explanations were arguments, as defenders of the epistemic accounts hold. In contrast, I advocate an ontic view of explanation according to which one explains a phenomenon by showing how it is situated in the causal structure of the world. Craver, 2007, p. 200

What does it mean to say that explanations are (or are not) arguments? Well, I take the question to be about what explanations do and how they do it. First we have to figure out what explanations do, and then we can decide for ourselves whether it makes sense to talk of them as arguments. There are many kinds of arguments. Descriptions of causal mechanisms in the world are logical and linguistic entities, and they are used somehow by an explanation. If they aren't an argument to Craver's standards, I argue the description functions effectively as an argument in an explanatory structure. Explanations of phenomena are functions of descriptions of causal mechanisms. We input our (context-and-user-relative) description of multi-level causal mechanisms (*systems*) into our explanation function, and out pops an explanation of the phenomenon.

What are the kinds of things that we wish to explain? Lets take an event or phenomenon P and try to explain it. The way the word 'explain' is used in this sense presupposes that to explain is to engage in some activity, some protocol. How are we going to explain P? I do not think it is out of the ordinary to characterize explanations of P as functions. The purpose of an explanation is to enable an agent (or artificial protocol) to solve a classification problem.

Such a classification problem will be, for our purposes, a high-level abstract task. Given an explanation, an agent will decide if P is explained or not. Thus, we can characterize an explanation as a function of the event, call it $E(P)$. In the trivial case, however, P by itself is relatively uninformative. It makes more sense to consider an explanation as a function of the event and a model or description of the event $E(P, M)$.

For example, consider a binary explanation function $E(P, M) \in \{0, 1\}$. Good explanations are classified as 1, i.e. they explain, otherwise they obtain 0 and do not explain. We can expand M into much more fine-grained elements, including for example much of the criteria developed by Craver Craver (2007). It can include defeaters, context, other aspects of a complex system which we

invoke to make sense of P. With more fine-grained arguments to the explanation function, we might reasonably move to a more fine-grained classification. We can move to a continuous explanation scale, where different models explain to a degree (say on the unit interval).

The degree of explanation for a particular agent may also be dependent on assumed value structures, or perceptual limitations, or other model-external factors. Thus, in general E may be a function of multiple variables and structures (M is, after all, a structure). This shouldn't be surprising if we expect explanations to be high-level abstract tasks. Crucially, explanations as outlined here depends on the *representation* of the model, as well as a representation of the phenomenon which will likely not be model-independent.

The *function* or *purpose* of explanations is to explain, and in this sense explanations must be more than just ontic explanatory objects—they need to be communicable and interpretable for the purposes of a decision problem. We are interested in whether some protocol or function E explains an event P. It is uninteresting for our purposes (discussing explanation in science) to bother about ontic explanatory objects. They are by themselves benign. An explanation needs to transfer some capacity of control.

For the purpose of control, it is simply necessary that the explanation be represented in some linguistic structure. It has to contain *information* that is non-trivially communicated by the explainer, such that the explainee can use the information (at least in principle) to control or regulate the mechanism of concern. The information must be intersubjectively transferred.

8.2 Summary

To conclude, I have argued for a model-based notion of computation in cognitive systems, which bolsters a structure mapping account of knowledge transfer. While originally intended to apply to high-level analogical reasoning, this structure mapping account makes sense at lower mechanistic levels for cognitive processes. To explain (and study) such interesting cases, I argue, it is necessary to abandon the ontic sense of explanation. Furthermore, there are abundant reasons to resist the anti-representational notions of characterizing relevant systems. Rather, a systems theory concerned with knowledge transfer in cognitive systems must be grounded on a robust notion of representation and structure. By appealing to the cybernetic picture outlined in particular by W. Ross Ashby, I believe a structural systems theory is a viable and consistent alternative to the neo-mechanist picture without sacrificing any progress.

In the model-based picture I argue in favor of, the ontic notion of explanation does not appear to be tenable. Structure mapping is the most obvious way to characterize the way that predictive and control capacity is transferred. As I have argued, this seems to be inconsistent not only with the mechanistic

views of Craver (2007) and Piccinini (2015), but apparently also their views of explanation. I have throughout this dissertation grounded my alternative view in the cybernetics of Ashby. To me Ashby's work arguably provides a consistent alternative foundation for a *structural* mechanistic view. Ashby also lays out a consistent picture of explanation for mechanists who have more pragmatic leanings. Compared to the ontic notion, in a structural systems theory explanations must be *communicable*.

GST provides a philosophy of science framework explicitly suited to analyze knowledge transfer between systems and, therefore, between modeling domains. GST focuses on structural properties common between systems, allowing intertheoretical transfer of knowledge for a scientist, modeler, or philosopher. A practicing systems theorist may take the knowledge gained from one system (i.e. a method of problem solving, or an idealized mechanism) and apply the method or mechanism to another (typically less well-understood) system.

My research has centered the account of knowledge transfer around the themes of model-based structure relevant for control, the mapping of such structure, and how an overall framework for a structural systems theory might emerge. After a background in the scientific methodology of general systems theory and cybernetics (2) I worked my way up from a model-based notion of computation (3), through connectionist ANN models (4), and into the theory of artificial general intelligence as the automation of scientific modeling (5). Then I provided one example of how one might model successful analogy in scientific reasoning (6), and how structure mapping might go wrong (7). Finally, I concluded by more explicitly endorsing a structural systems theory and a notion of explanation as enabling control (8).

Bibliography

Aggarwal, C.C. (2018). *Neural Networks and Deep Learning: A Textbook*. Springer. ISBN: 9783319944630.

Asaro, Peter (2008). "From Mechanisms of Adaptation to Intelligence Amplifiers: The Philosophy of W. Ross Ashby". In: *The Mechanical Mind in History*. Ed. by Phil Husbands, Owen Holland, and Michael Wheeler. MIT Press.

— (2011). "Computers as Models of the Mind: On Simulations, Brains and the Design of Early Computers". In: *The Search for a Theory of Cognition: Early Mechanisms and New Ideas*. Ed. by Stefano Franchi and Francesco Bianchini. Rodopi. Chap. 3, pp. 89–114.

Ashby, W. Ross (1958). *An Introduction to Cybernetics*. Chapman and Hall.

— (1960). *Design for a Brain: The Origin of Adaptive Behavior*. Second Edition. John Wiley and Sons.

— (1962). "Simulation of a Brain". In: *Computer Applications in the Behavioral Sciences*. Ed. by Harold Borko. Prentice-Hall Behavioral Sciences in Business Series. Prentice-Hall. Chap. 19.

— (1991a). "General Systems Theory as a New Discipline". In: *Facets of Systems Science*. Boston, MA: Springer US, pp. 249–257. ISBN: 978-1-4899-0718-9. DOI: 10.1007/978-1-4899-0718-9_16.

— (1991b). "Requisite Variety and Its Implications for the Control of Complex Systems". In: *Facets of Systems Science*. Boston, MA: Springer US, pp. 405–417. ISBN: 978-1-4899-0718-9. DOI: 10.1007/978-1-4899-0718-9_28.

— (2008). *Journal (1928-1972)*. The W. Ross Ashby Digital Archive. URL: http://www.rossashby.info/journal.

Barrow, John D. (2002). *The Book of Nothing: Vacuums, Voids, and the Latest Ideas about the Origins of the Universe*. Vintage.

Bartha, Paul (2010). *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford University Press.

— (2013). "Analogy and Analogical Reasoning". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2013.

Basheer, I.A and M Hajmeer (2000). "Artificial neural networks: fundamentals, computing, design, and application". In: *Journal of Microbiological Methods* 43.1. Neural Computting in Micrbiology, pp. 3 –31. ISSN: 0167-7012. DOI: 10.1016/S0167-7012(00)00201-3.

Beebe, Cameron (2016). "Model-Based Computation". In: *Proceedings of Unconventional Computation and Natural Computation: 15th International Conference, UCNC 2016, Manchester, UK, July 11-15, 2016*. Ed. by Martyn Amos and Anne Condon. Vol. 9726. Lecture Notes in Computer Science. Springer International Publishing, pp. 75–86. ISBN: 978-3-319-41312-9. DOI: 10.1007/978-3-319-41312-9_7.

— (2018). "Model-based computation". In: *Natural Computing* 17.2, pp. 271–281. ISSN: 1572-9796. DOI: 10.1007/s11047-017-9643-0.

Bertalanffy, Ludwig von (1950). "An Outline of General System Theory". In: *The British Journal for the Philosophy of Science* 1.2, pp. 134–165. ISSN: 00070882, 14643537. URL: http://www.jstor.org/stable/685808.

— (1969). *General System Theory*. George Braziller.

Birkhoff, Garrett and John Von Neumann (1936). "The Logic of Quantum Mechanics". In: *Annals of Mathematics* 4, pp. 823–843. ISSN: 0003486X. DOI: 10.2307/1968621.

Bode, Hendrik et al. (1949). "The Education of a Scientific Generalist". In: *Science* 109.2840, p. 553–558.

Borko, Harold (1962). "History and Development of Computers". In: *Computer Applications in the Behavioral Sciences*. Ed. by Harold Borko. Prentice-Hall Behavioral Sciences in Business Series. Prentice-Hall. Chap. 3.

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners". In: arXiv: 2005.14165 [cs.CL].

Busemeyer, Jerome R. and Peter D. Bruza (2012). *Quantum Models of Cognition and Decision*. Cambridge University Press.

Busemeyer, Jerome R. et al. (2011). "A Quantum Theoretical Explanation for Probability Judgement Errors". In: *Psychological Review* 118.2, pp. 193–218.

Carbonell, Jaime G. (1983). "Learning by Analogy: Formulating and Generalizing Plans from Past Experience". In: *Machine Learning: An Artificial Intelligence Approach*. Ed. by Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell. Symbolic Computation. Berlin, Heidelberg: Springer, pp. 137–161. ISBN: 978-3-662-12405-5. DOI: 10.1007/978-3-662-12405-5_5.

Care, Charles (2010). *Technology for Modelling: Electrical Analogies, Engineering Practice, and the Development of Analogue Computing*. History of Computing. Springer London. ISBN: 978-1-84882-947-3. DOI: 10.1007/978-1-84882-948-0_1.

Chollet, François et al. (2015). *Keras*. URL: https://keras.io.

Clark, Andy (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science". In: *Behavioral and Brain Sciences* 36.3, pp. 181–204. DOI: 10.1017/S0140525X12000477.

Colombo, Matteo, Stephan Hartmann, and Robert van Iersel (2014). "Models, Mechanisms, and Coherence". In: *The British Journal for the Philosophy of Science* 1, pp. 181–212. ISSN: 0007-0882. DOI: 10.1093/bjps/axt043.

Conant, Roger C. and W. Ross Ashby (1970). "Every good regulator of a system must be a model of that system". In: *International Journal of Systems Science* 1.2, pp. 89–97. DOI: 10.1080/00207727008920220.

Cordeschi, Roberto (2002). *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics*. Ed. by James Fetzer. Vol. 28. Studies in Cognitive Systems. Springer Netherlands. ISBN: 978-94-015-9870-5. DOI: 10.1007/978-94-015-9870-5.

Corfield, David, Bernhard Schölkopf, and Vladimir Vapnik (2009). "Falsificationism and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis Dimensions". In: *Journal for General Philosophy of Science* 40.1, pp. 51–58. ISSN: 1572-8587. DOI: 10.1007/s10838-009-9091-3.

Costa, Newton Da and Steven French (2003). *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*. Oxford University Press.

Craik, Kenneth (1943). *The Nature of Explanation*. Cambridge University Press.

Craver, Carl F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Dardashti, Radin, Karim Thébault, and Eric Winsberg (2017). "Confirmation via Analogue Simulation: What Dumb Holes Could Tell Us about Gravity". In: *The British Journal for the Philosophy of Science* 68.1, pp. 55–89. DOI: 10.1093/bjps/axv010.

Davis, Randall and Walter Hamscher (1988). *Model-Based Reasoning: Troubleshooting*. Memorandum AI Memo 1059. Artificial Intelligence Laboratory, Advanced Research Projects Agency, Office of Naval Research.

Denardo, Bruce C., Joshua J. Puda, and Adrés Larraza (2009). "A water wave analog of the Casimir effect". In: *American Journal of Physics* Vol. 77.No. 12, p. 1095–1101.

Deng, J. et al. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*.

Dizadji-Bahmani, F., R. Frigg, and S. Hartmann (2011). "Confirmation and reduction: a Bayesian account". In: *Synthese* 179, p. 321 –338.

Esser, Steven K. et al. (2016). "Convolutional networks for fast, energy-efficient neuromorphic computing". In: *Proceedings of the National Academy of Sciences* 113.41, pp. 11441–11446. DOI: 10.1073/pnas.1604850113.

Falkenhainer, Brian, Kenneth D. Forbus, and Dedre Gentner (1989). "The structure-mapping engine: Algorithm and examples". In: *Artificial Intelligence* 41.1, pp. 1 –63. ISSN: 0004-3702. DOI: 10.1016/0004-3702(89)90077-5.

Fernando, Chrisantha et al. (2017). "PathNet: Evolution Channels Gradient Descent in Super Neural Networks". In: *CoRR*. arXiv: 1701.08734.

Feurer, Matthias et al. (2015). "Efficient and Robust Automated Machine Learning". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., pp. 2962–2970.

Fraassen, Bas van (1980). *The Scientific Image*. Oxford University Press.

Frigg, Roman and Stephan Hartmann (2012). "Models in Science". In: *Stanford Encyclopedia of Philosophy*.

Friston, Karl (2010). "The free-energy principle: a unified brain theory?" In: *Nature Reviews Neuroscience* 11.2, pp. 127–138. DOI: 10.1038/nrn2787.

Gentner, Dedre (1983). "Structure-Mapping: A Theoretical Framework for Analogy". In: *Cognitive Science* 7, pp. 155–170.

Giere, Ronald (1988). *Explaining Science: A Cognitive Approach*. University of Chicago Press.

— (2004). "How Models Are Used to Represent Reality". In: *Philosophy of Science* 71.5, pp. 742–752. ISSN: 00318248, 1539767X. DOI: 10.1086/425063.

Glymour, Clark (1980). *Theory and Evidence*. Princeton University Press.

Goldstone, Robert L. and Uri Wilensky (2008). "Promoting Transfer by Grounding Complex Systems Principles". In: *Journal of the Learning Sciences* 17.4, pp. 465–516. DOI: 10.1080/10508400802394898.

Harman, G. and S. Kulkarni (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. A Bradford book. MIT Press. ISBN: 9780262083607.

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

Hofstadter, Douglas (2001). "Epilogue: Analogy as the Core of Cognition". In: *The Analogical Mind: Perspectives from Cognitive Science*. Ed. by Dedre Gentner, Keith J. Holyoak, and Boicho N. Kokinov. MIT Press.

Holyoak, Keith and Paul Thagard (1997). "The Analogical Mind". In: *American Psychologist* 52.1, pp. 35–44.

Holzinger, Andreas (2016). "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics* 3.2, pp. 119–131. ISSN: 2198-4026. DOI: 10.1007/s40708-016-0042-6.

Howard, Jeremy and Sebastian Ruder (2018). "Fine-tuned Language Models for Text Classification". In: *CoRR*. arXiv: 1801.06146.

Humphreys, Paul (2009). "The philosophical novelty of computer simulation methods". In: *Synthese* 169.3, pp. 615–626. ISSN: 1573-0964. DOI: 10.1007/s11229-008-9435-2.

Jardine, N. (1967). "The Concept of Homology in Biology". In: *The British Journal for the Philosophy of Science* 18.2, pp. 125–139. ISSN: 00070882, 14643537.

Korb, Kevin B. (2004). "Introduction: Machine Learning as Philosophy of Science". In: *Minds and Machines* 14.4, pp. 433–440. ISSN: 1572-8641. DOI: 10.1023/B:MIND.0000045986.90956.7f.

Landauer, R. (1961). "Irreversibility and Heat Generation in the Computing Process". In: *IBM J. Res. Dev.* 5.3, pp. 183–191. ISSN: 0018-8646. DOI: 10.1147/rd.53.0183.

Lauritzen, Steffen and Thomas S. Richardson (2001). "Chain Graph Models and their Causal Interpretations". In: *B* 64, pp. 321–361.

Leigh, James Ron (2012). *Control Theory: A Guided Tour*. 3rd. IET Control Engineering Series. London, United Kingdom: The Institution of Engineering and Technology.

Lu, Hongjing, Ying Nian Wu, and Keith J. Holyoak (2019). "Emergence of analogy from relation learning". In: *Proceedings of the National Academy of Sciences* 116.10, pp. 4176–4181. ISSN: 0027-8424. DOI: 10.1073/pnas.1814779116.

Malapi-Nelson, Alcibiades (2017). *The Nature of the Machine and the Collapse of Cybernetics: A Transhumanist Lesson for Emerging Technologies*. Palgrave Studies in the Future of Humanity and its Successors. Springer International Publishing. ISBN: 978-3-319-54517-2. DOI: 10.1007/978-3-319-54517-2_3.

Mandelbaum, Eric (2017). "Associationist Theories of Thought". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2017. Metaphysics Research Lab, Stanford University.

Marr, David (1982). *Vision*. W.H. Freeman and Company.

Marr, David and Tomaso Poggio (1976). *From Understanding Computation to Understanding Neural Circuitry*. A.I. Memo 357. Massachusetts Institute of Technology.

Maxwell, James Clerk (1868). "I. On governors". In: *Proceedings of the Royal Society of London* 16, pp. 270–283. DOI: 10.1098/rspl.1867.0055.

Mead, Carver (1990). "Neuromorphic electronic systems". In: *Proceedings of the IEEE* 78.10, pp. 1629–1636. ISSN: 0018-9219. DOI: 10.1109/5.58356.

Mundy, Brent (1986). "On the General Theory of Meaningful Representation". In: *Synthese* 67.3, pp. 391–437. ISSN: 00397857, 15730964.

Neumann, John Von (1963). "The General and Logical Theory of Automata". In: *John von Neumann Collected Works*. Ed. by A. H. Taub. Vol. V: Design of Computers, Theory of Automata and Numerical Analysis. Pergamon Press, pp. 288–326.

Nielsen, Michael and Isaac Chuang (2010). *Quantum Computation and Quantum Information*. Cambridge University Press.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. ISBN: 9780521773621.

Piccinini, Gualtiero (2015). *Physical Computation: A Mechanistic Account*. Oxford University Press.

Piccinini, Gualtiero and Sonya Bahar (2013). "Neural Computation and the Computational Theory of Cognition". In: *Cognitive Science* 3, pp. 453–488. ISSN: 1551-6709. DOI: 10.1111/cogs.12012.

Pickering, Andrew (2010). *The Cybernetic Brain: Sketches of Another Future*. The University of Chicago Press.

Pickett, Marc and David W. Aha (2013a). "Spontaneous Analogy by Piggybacking on a Perceptual System". In: *Proceedings of the 35th annual conference of the cognitive science society*. arXiv: 1310.2955.

— (2013b). "Using cortically-inspired algorithms for analogical learning and reasoning". In: *Biologically Inspired Cognitive Architectures* 6. BICA 2013: Papers from the Fourth Annual Meeting of the BICA Society, pp. 76 –86. ISSN: 2212-683X. DOI: 10.1016/j.bica.2013.07.003.

Poellinger, Roland (2012). *Concrete Causation: About the Structures of Causal Knowledge*. (doctoral thesis). Ludwig-Maximilians-Universität München. URL: https://epub.ub.uni-muenchen.de/21384/.

Pothos, Emmanuel M. and Jerome R. Busemeyer (2013). "Can quantum probability provide a new direction for cognitive modeling?" In: *Behavioral and Brain Sciences* 36, pp. 255–327.

Pratt, Lorien (1993). "Discriminability-based transfer between neural networks". In: *Advances in neural information processing systems*, pp. 204–211.

Pratt, Lorien, Jack Mostow, and Candace Kamm (1991). "Direct Transfer of Learned Information Among Neural Networks". In: *AAAI-91 Proceedings*, pp. 584–589.

Reck, Erich and Georg Schiemer (2019). "Structuralism in the Philosophy of Mathematics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University.

Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow (1943). "Behavior, Purpose and Teleology". In: *Philosophy of Science* 10.1, pp. 18–24. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/184878.

Rubel, Lee A. (1985). "The Brain as an Analog Computer". In: *Journal of Theoretical Neurobiology* 4, p. 73–81.

Sabour, Sara, Nicholas Frosst, and Geoffrey E Hinton (2017). *Dynamic Routing Between Capsules*. arXiv: 1710.09829 [cs.CV].

Searle, John R. (1990). "Is the Brain a Digital Computer?" In: *Proceedings and Addresses of the American Philosophical Association* 64.3, pp. 21–37.

Seth, Anil K. (2015). "The Cybernetic Bayesian Brain". In: *Open MIND*. Ed. by Thomas K. Metzinger and Jennifer M. Windt. Frankfurt am Main: MIND Group. Chap. 35(T). ISBN: 9783958570108. DOI: 10.15502/9783958570108.

Shagrir, Oron (2010). "Brains as analog-model computers". In: *Studies in History and Philosophy of Science Part A* 41.3. Computation and cognitive science, pp. 271 –279. ISSN: 0039-3681. DOI: 10.1016/j.shpsa.2010.07.007.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Sharkey, Noel E. and Amanda J. C. Sharkey (1993). "Adaptive generalisation". In: *Artificial Intelligence Review* 5, pp. 313–328. ISSN: 1573-7462. DOI: 10.1007/BF00849058.

Simon, Herbert A. (1973). "Does Scientific Discovery Have a Logic?" In: *Philosophy of Science* 40.4, pp. 471–480. ISSN: 00318248, 1539767X.

Suppes, Patrick (2002). *Representation and Invariance of Scientific Structures*. CSLI Publications.

Swoyer, Chris (1991). "Structural representation and surrogative reasoning". In: *Synthese* 87.3, pp. 449–508. ISSN: 1573-0964. DOI: 10.1007/BF00499820.

Tegmark, Max (2000). "Importance of quantum decoherence in brain processes". In: *Physical Review E* 61.4, 4194–4206. ISSN: 1095-3787. DOI: 10.1103/physreve.61.4194.

Teller, Paul (2001). "Twilight of the Perfect Model Model". In: *Erkenntnis* 55.3, pp. 393–415. ISSN: 01650106, 15728420.

Terwagne, Denis and John W. M. Bush (2011). "Tibetan singing bowls". In: *Nonlinearity* Vol. 24, p. R51–R66.

Thagard, Paul (1982). "Artificial Intelligence, Psychology, and the Philosophy of Discovery". In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1982, pp. 166–175. ISSN: 02708647.

— (2010). "How Brains Make Mental Models". In: *Model-Based Reasoning in Science and Technology*. Ed. by Lorenzo Magnani, Walter Carnielli, and Claudio Pizzi. Vol. 314. Studies in Computational Intelligence. Springer, pp. 447–461.

Trenholme, Russell (1994). "Analog Simulation". In: *Philosophy of Science* 61.1, pp. 115–131. ISSN: 00318248, 1539767X. URL: http://www.jstor.org/stable/188292.

Turing, Alan M. (1950). "Computing Machinery and Intelligence". In: *Mind* LIX.236, pp. 433–460. DOI: 10.1093/mind/LIX.236.433.

Tversky, Amos and Daniel Kahneman (1981). *Judgements Of and By Representativeness*. Tech. rep. Stanford University Department of Psychology.

Ulmann, Bernd (2013). *Analog Computing*. de Gruyter.

Unruh, W. G. (2008). "Dumb holes: analogues for black holes". In: *Philosophical Transactions of The Royal Society A* 366, p. 2905–2913.

Vapnik, Vladimir (2000). *The Nature of Statistical Learning Theory*. Springer.

Verma, Thomas and Judea Pearl (1988). "Causal Networks: Semantics and Expressiveness". In: *Proceedings of the 4th Annual Conference on Uncertainty in Artificial Intelligence (UAI-88)*. New York: Elsevier Science.

Wheeler, Gregory (2016). "Machine Epistemology and Big Data". In: *The Routledge Companion to Philosophy of Social Science*. Ed. by Lee McIntyre and Alex Rosenberg. Routledge.

Wiener, Norbert (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

Wolpert, D. H. and W. G. Macready (1997). "No free lunch theorems for optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1, pp. 67–82. ISSN: 1089-778X. DOI: 10.1109/4235.585893.

Woodward, James (2016). "Causation and Manipulability". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.

Wright, Cory (2012). "Mechanistic explanation without the ontic conception". In: *European Journal for Philosophy of Science* 2.3, pp. 375–394. ISSN: 1879-4920. DOI: 10.1007/s13194-012-0048-8.

— (2015). "The ontic conception of scientific explanation". In: *Studies in History and Philosophy of Science Part A* 54.Supplement C, pp. 20 –30. ISSN: 0039-3681. DOI: 10.1016/j.shpsa.2015.06.001.

Yearsley, James M. (2017). "Advanced tools and concepts for quantum cognition: A tutorial". In: *Journal of Mathematical Psychology* 78.Supplement C. Quantum Probability and Contextuality in Psychology and Economics, pp. 24 –39. ISSN: 0022-2496. DOI: 10.1016/j.jmp.2016.07.005.

Zhang, Chiyuan et al. (2016). "Understanding deep learning requires rethinking generalization". In: *CoRR*. arXiv: 1611.03530.

Zurek, Wojciech Hubert (2007). "Decoherence and the Transition from Quantum to Classical — Revisited". In: *Quantum Decoherence: Poincaré Seminar 2005*. Ed. by Bertrand Duplantier, Jean-Michel Raimond, and Vincent Rivasseau. Birkhäuser Basel, pp. 1–31. ISBN: 978-3-7643-7808-0. DOI: 10.1007/978-3-7643-7808-0_1.