universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

On the Explanatory Value of Bayesian Cognitive Modelling: A Mechanistic Perspective

verfasst von | submitted by

Matteo Mattersberger BSc BA MSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 944
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Interdisziplinäres Masterstudium Wissenschaftsphilosophie und Wissenschaftsgeschichte
Betreut von | Supervisor:	Univ.-Prof. Mag. Mag. Dr. Georg Schiemer

Abstract

Der Bayesianismus hat sich in den letzten Jahren als eigenständiges, fruchtbares Paradigma in den Kognitionswissenschaften etabliert. Basierend auf der mathematischen Grundlage der Wahrscheinlichkeitstheorie ist dieses Paradigma in der Lage, zahlreiche Facetten von Verhalten und Kognition mathematisch präzise zu modellieren und vorherzusagen. Doch worin liegt der Erklärungswert dieser Modelle? Können bayesianische Modelle die von ihnen modellierten kognitiven Prozesse auch erklären? Die vorliegende Arbeit dient dem Zweck, diese Frage aus einer philosophischen Perspektive zu erörtern. Hierzu wird zunächst versucht, einen philosophisch adäquaten Begriff der Erklärung im Kontext der Kognitionswissenschaften zu etablieren, um diesen dann als prüfenden Maßstab an das Feld der Bayesianischen Kognitionswissenschaft anzulegen. Das zentrale Argument hierbei ist, dass Erklärung in den Kognitionswissenschaften im Grunde mechanistisch ist, und Bayesianische Modelle nur Erklärungswert besitzen, insofern sie in der Lage sind, mechanistische Grundlagen der Kognition zu beleuchten.

Introduction

Despite significant progress in cognitive science, shedding light on some of the inner workings of the mind, there is still no scientific consensus on the fundamental principles underlying cognition. However, proponents of a newly emerging trend in cognitive science suggest that such an underlying principle could be provided by the mathematical framework of probability theory, more specifically, Bayesian inference. They claim that the mind functions as a probabilistic inference machine, making inferences about the distal causes of its sensory input, and that this seemingly simple principle can account for large parts, if not the entirety, of our mental lives. The aim of this thesis is not to evaluate the content of this claim. Instead, it seeks to assess the epistemic potential of its structure, or more specifically, the structure of the models it motivates. In other words, the present thesis aims to investigate to what extent Bayesian models in cognitive science are genuinely explanatory in nature (assuming they prove empirically adequate on a large scale), as opposed to being merely descriptive or predictive.

Before delving into developing an answer to this question, let us very briefly elucidate its components. First, Bayesianism is a mathematical framework for making inferences under uncertainty, which is rooted in probability theory. It operates by incorporating new empirical evidence into a prior model of the domain under consideration. The second major component of our guiding question, explanation in science (understood in a pre-theoretic sense not yet committed to a particular philosophical viewpoint beyond an everyday understanding of the concept), is the collective term for proper and adequate answers to question of the type “Why is X the case?” or “How does Y work?”. This provisional characterization is pre-theoretical because a more detailed definition of scientific explanation can hardly be given without already taking a stance on the contested question what exactly constitutes scientific explanation. An investigation of this question will follow in due course.

For the time being, this very cursory introduction to the notion of scientific explanation and the core idea of Bayesianism shall suffice, as first, the core question being asked should be sufficiently graspable even without a lengthy introduction to its core elements, and second, both notions will be duly introduced and discussed at length in the respective sections devoted to them.

What is well worth delving into in some more depth at this point, however, and what therefore shall constitute the bulk of the present introductory remarks, is the notion of a unified scientific discipline termed "cognitive science". The field of cognitive science emerged in the second half of the previous century out of concurrent developments in psychology, linguistics, computer

science (especially AI research), and philosophy (Bermudez, 2014). The unifying tenet binding these heterogeneous approaches together was the belief in the power of cognitive explanations, i.e., the notion that the best way to understand a cognitive system is by explicating the information processing operations it performs (Bermudez, 2014). In its early days, this conviction equated to what is now called classical cognitivism or representationalism, i.e., the idea that cognition is the syntactic manipulation of meaningful symbols, much like the operating principles of a digital computer (Bermudez, 2014). However, besides this shared origin story and a general dedication to studying cognition (with significant variation among proposals about what this term actually denotes), there is little holding present-day cognitive science together as a unitary field. The field of cognitive science at the beginning of the 21st century is fragmented in at least three senses of the term.

First, cognitive science is a multiparadigmatic discipline. Following Kuhn's (1996 [1962]) classical usage of the term, a "paradigm" denotes a (at least almost) universally recognized set of model solutions to a given scientific problem within a community of researchers. This model specifies the type of question that can sensibly be asked in the field, as well as the type of answer that is deemed acceptable. In a sense, a paradigm defines the ruleset of a field of inquiry. The field of cognitive science, however, lacks such a commonly shared and universally accepted ruleset.

Other than a vague agreement on a commitment to cognitive explanation, few common assumptions are shared by all its proponents (the following overview follows Bermudez, 2014 in its content and selection of major paradigms). The historically oldest among the competing frameworks within the field is the previously mentioned paradigm of "classical cognitivism", a view which understands cognition and the action wherein it results as the result of the syntactic manipulation of mental symbols. Inspired by Rosenblatt's (1958) work on the multilayer perceptron, a development that was famously initially shunned by proponents of classical cognitivism in computer science (Minsky & Papert, 1969), connectionism emerged as a second tradition in the field in the 1980s, trying to explain and model cognition as a spreading pattern of activation among individually non-representative nodes in a computational network. Around the same time, another alternative to classical cognitivism emerged in the form of dynamical systems theory. The core idea of the latter is that cognition can be understood as a context-sensitive phenomenon continuously developing over time, the essence of which can be captured using differential equation modelling. In the following years, yet another conceptual framework gained traction in the form of the 4E – extended, embedded, embodied, enacted – cognition. Unified by the idea that cognition should not be understood as purely 'internal' information

processing leading to an output of acting on the world, the 4E approach proposed that cognition is a process that – depending on the exact variety – stands in close connection to acting in the world, is itself a bodily process not limited to neural activity, or ought to be understood as extending into the world, transcending the cranial barrier between brain and outside world. These frameworks, besides differing in their claims about the nature of cognition, also starkly differ in the *type of explanation* they deem suitable for the field of cognitive science, with some of them being committed to specific mathematical structures, such as differential equation modelling in the dynamical systems approach or artificial neural networks in the connectionist framework. Thus, their differences are not restricted to the content of theories and models, but also concern the form they should take.

Besides being a multiparadigmatic field, cognitive science is secondly also a multidisciplinary one. It aims to gain an understanding of cognition using the methods of such different disciplines as psychology, computer science, linguistics, neuroscience, and philosophy (Bermudez, 2014). Given both the multidisciplinary and multiparadigmatic nature of the field, its characterization given in the present examination cannot go far beyond the already stated vague and general notion that the aim of cognitive science is the study of cognition. What exactly this entails and excludes is itself subject to heated debate and arguably one of the core questions of the field, which is why a clear and definite demarcation of the field cannot be given at this point.

Third and last, cognitive science is fragmented in that many of its results lack a cohesive theoretical framework within which they can be accounted for; it accumulates statistical results which are often not unified and integrated to make theoretical advances (cf. Meehl, 1992). This practice, however, does not help the field to progress in devising concrete theories of cognition, i.e., it limits cognitive science to a collection of individual results as opposed to set of cohesive theories (or so critics in the field sometimes argue, cf. Meehl, 1992).

It is this multiply-fragmented state of the field of cognitive science in which the aforementioned trend of Bayesian approaches to cognition gained traction. Building upon work developed in the 1980s and 1990s, proponents of the Bayesian brain hypothesis aimed to explain the mind as operating essentially in accordance with the laws of probability theory, as inference machines working following the principles of Bayesian statistics (Chater & Oaksford, 2008). The concurrent development of more advanced Bayesian computational methods and the increasing availability of computational power in the fields of computer science and statistics contributed to this resurgence, enabling more complex models and analyses that align with Bayesian

principles (cf. Chater & Oaksford, 2008). This in turn allowed cognitive scientists to employ these newly developing technologies to conceive of computationally costly computer models and simulations of cognitive processes, a practice I shall refer to as ‘Bayesian cognitive modelling’, thereby contributing to and supplementing the core idea of the Bayesian mind (Chater & Oaksford, 2008).

Proponents of this emerging Bayesian framework claim that its initially simple-seeming idea has immense unificatory power for the fragmented field of cognitive science (especially in regard to unifying hitherto independent empirical results), with Bayesian inference serving as the underlying explanatory principle. For example, Clark (2013, p. 201), a vocal supporter of one particular flavor of Bayesianism, the predictive processing theory, states that

> one way to think about the primary “added value" of these models is that they bring perception, action, and attention into a single unifying framework. They thus constitute the perfect explanatory partner [...] for recent approaches that stress the embodied, environmentally embedded, dimensions of mind and reason.

In a similar vein, Hohwy (2013, p. 1) proposes that this approach to understanding the mind “has enormous unifying power and yet it can explain in detail too”, thus also ascribing to the dual claim about both unificatory and explanatory power of Bayesian approaches.

Why introduce the present state of cognitive science in such detail? The reason is simple. The fragmented state of cognitive science leads to the very notion of a cognitive explanation remaining contested in the field, with different research paradigms proposing different types of models or theory as explanatory. This disagreement is not simply a theoretical one between researchers adhering to different accounts of a given phenomenon, it is a metascientific disagreement about the nature of cognitive explanation, i.e., about what it means to explain a cognitive phenomenon. What does it mean for a model in cognitive science to be explanatory? What constitutes cognitive explanation? Which kind of entities must it allude to? How does it relate to explanation in other fields of scientific inquiry? There is no overarching standard of scientific explanation in cognitive science providing answers to these questions. The types of explanation accepted within the rival frameworks of cognitive science are often bound to the contents of their respective framework, with, for example, 4E theorists rejecting – due to their theoretical convictions – purely computational accounts as insufficient for explanation. It is often unclear in such discussions whether the validity of an explanation is contested, i.e., whether the explanation is rejected on grounds of its allegedly erroneous claims, or whether the very status of a proposal as a possible (albeit potentially wrong) explanation is denied.

This is where the present thesis aims to make its contribution. By exploring accounts of scientific explanation in the philosophy of science, it aims to investigate the nature of explanation in the field of cognitive science. In doing so, it does not merely pursue the goal of adding to the discussion about the character of scientific explanation for its own sake, although this still strikes me as a goal worth pursuing, but rather to endorse and expand upon a set of criteria by which one can assess the quality of any posited cognitive explanation, and thus ultimately assist theoretical advancement in cognitive science. To this end, the mechanistic framework of neuroscientific explanation proposed by Carl Craver (2006, 2007) shall be introduced and defended. My claim, in short, will be that explanation in cognitive science is ultimately mechanistic, and its specific character is captured by Craver's framework (despite it not being devised with the field of cognitive science in mind), as well as that supposedly non-mechanistic forms of explanation in the field are either reducible to mechanistic explanation or fail to provide genuine explanatory value.

The standards of explanation thus developed shall then be applied to (some of) the models proposed in the newly emerging paradigm of Bayesian cognitive science, not taking claims by their proponents about the explanatory power of such models at face value, but rather applying the measuring stick of mechanism developed by previous philosophical consideration to them. Wherein exactly lies the explanatory value of the Bayesian framework, if it possesses any? To what extent is it able to explain, as opposed to merely mathematically reframe, mental phenomena, and what is the difference between the latter two (assuming there is one)? Is the supposed explanatory power of Bayesianism sufficient for a unification of the field of cognitive science? Vice versa, to what extent can its unificatory potential be said to contribute to its explanatory value? These are the questions the present thesis aims to provide answers to.

To this end, it pursues three partially distinct argumentative goals, which also provide the rough outline of the thesis. First, it aims to introduce some of the recent developments in the field of cognitive science in some more detail, namely Bayesian approaches to cognition. This is what sections one and two will be devoted to, with section one giving a general introduction to the principles of Bayesianism and section two elucidating how these principles have been applied in the field of cognitive science.

Second, its goal is to endorse a particular philosophical stance regarding the nature of cognitive explanation, i.e., what type of explanation is acceptable within the field of cognitive science, namely new mechanism. To this end, section three shall give a general overview of the philosophical debate regarding scientific explanation and introduce a set of desiderata for a

proper philosophical account of it. Section four shall introduce the position of new mechanism, a philosophical stance on the explanation question, and more particularly, as previously stated, Craver's (2006, 2007) mechanistic framework. In section five, it will be argued that said account, originally developed with the field of neuroscience in mind, is suitable for assessing scientific explanations in the field of cognitive science.

Third and finally, the present thesis aims, employing the very stance established in sections three and four, to evaluate some of the scientific developments occurring in the Bayesian framework of cognitive science, assessing the explanatory virtues and shortcomings of this approach. This will mostly occur in section six, with section seven discussing and attempting to refute possible objections to the presented account.

Whereas the first and the second aims are largely independent of each other, and the respective sections devoted to them can be read independently, the third aim brings both of these strands together, hopefully providing a critique of the field grounded in both a comprehensive and nuanced grasp of the subject matter as well as a carefully deliberated philosophical framework of evaluation. As such, the current thesis aims to make a contribution to the field of philosophy *of* cognitive science, rather than philosophy *in* cognitive science. Let this remark be elucidated. "Philosophy of cognitive science", following Brook (2009), denotes any inquiry into questions in the philosophy of science pertaining to the field of cognitive science, for example the nature of theory, modelling, or explanation within this particular field, whereas "philosophy in cognitive science" denotes any effort to aid the goal of cognitive science, understanding the nature of cognition, using the methods of philosophy. Given the closely interwoven subject matter of these adjacent fields, no clear demarcation line can be drawn between them, nor is there, I believe, any good reason for attempting to do so. The categorization as one rather than the other merely serves as a rhetorical tool for adequately conveying the argumentative goals of the present thesis and thereby, hopefully, adequately shaping the reader's expectations of it.

Before commencing with the pursuit of the outlined argumentative goals, two brief remarks regarding the terminology employed in the present thesis are in order. First, the term "cognition" shall be reserved for the type of biological cognition studied by psychology and cognitive biology, i.e., cognition displayed by humans and non-human animals, thus excluding both the vast field of AI research and philosophical speculations regarding merely conceivable but not, to the best of our knowledge, actually instantiated kinds of minds (cf. Putnam, 1967, for a well-known example in the philosophical literature) from our current considerations. This is the case not because I wish to make some fundamental ontological distinction between biological and

artificial cognition, nor is it because I wish to deny the possibility of fundamentally different types of cognition from the ones found on the biological tree of life. Rather, the reason is that the current thesis focuses on cognitive science as a natural science, studying the natural phenomenon of cognition as it is displayed by humans and animals, and how it can be explained scientifically, and natural, biological cognition is the proper subject of this inquiry (of course cognitive science also investigates and aims to create forms of artificial intelligence, but it does not do so in its capacity as a natural science, rather in its capacity as a subfield of computer science and engineering. This distinction may be to some extent arbitrary, but it is nevertheless useful for explicating the scope of our investigation). Thus, when I speak of "cognitive systems" without giving further qualifications this term shall denote humans, chimpanzees, mice and other biological cognizers sharing our phylogenetic roots, but it shall exclude AI programs, cognitive systems made entirely of Swiss cheese (cf. Putnam, 1979), or other artificial or purely hypothetical cognizers.

The second specification I wish to make is one taking away as opposed to adding constraints on the usage of certain terms, namely 'model' and 'theory'. In using these terms, I do not wish to commit to any specific philosophical theory about the exact nature of either entity, nor to an account of their relation to each other. Thus, when I speak of a scientific model (or a scientific theory), this simply denotes the kind of things that scientists usually call a model (or a theory), without any further commitments to its exact defining characteristics. The reason for this lack of philosophical commitment is that the considerations given in the present thesis are largely orthogonal to questions regarding the exact character of theories and models, and making a commitment regarding the latter thus contributes terribly little to the enterprise of the former.

Having thus specified in some detail the argumentative goals of the present work as well as clarified the relevant terminology, let us commence in the ensuing section by providing a general introduction to the principles of Bayesianism.

Bayesian Inference

The Bayesian Theorem

Although the term "Bayesianism" refers to a different set of assumptions, theses, and practices depending on the context of its usage (e.g., statistics, cognitive science, epistemology, …), what connects these approaches across different domains is their endorsement and usage of Bayesian inference methods in some capacity. Before diving into paradigm of Bayesianism in cognitive science, the present section aims to give a more general, domain-independent introduction to the basics of Bayesian inference that lie at the heart of all varieties of Bayesianism.

Bayesian statistics is framework for inference under uncertainty, and thus a subfield of probability theory. Its core idea is to use probabilities to model the (un)certainty of a given hypothesis holding (Martin, 2016). "Probability", in a Bayesian context, ought to be understood as (subjective) degrees of belief in the hypothesis in question. In non-theoretical terms, what a Bayesian means by "the probability of x" is something along the lines of "the certainty of the judgement that x will occur" or "the certainty that x is the case" (depending on the exact hypothesis in question). This notably differs from the dominant frequentist interpretation which poses – informally put – that probability of x ought to be identified with the rate of x (rather than not-x) occurring in a (hypothetical) infinite series of events (Hajek, 2023). Much can be (and has been) said about the theoretical and philosophical qualities of both approaches, and the so called "statistics wars" continue to rage on in statistics and philosophy, but as this matter is not pertinent to the discussion at hand (which deals with the problem of explanation in cognitive science rather than the philosophy of mathematics), this issue will be glossed over for now (for an overview, see Hajek, 2023).

Having introduced the Bayesian understanding of the term probability, let us continue with the core of Bayesianism of all varieties. At the heart of Bayesian inference lies the Bayesian theorem, which relates the probability of a certain proposition H being true in light of a piece of evidence D, P(H|D), to a number of other probabilities, namely the probability of H being true independent of the evidence, P(H, the probability of D obtaining given the proposition H being true, P(D|H), and the probability of the evidence independent of other considerations, P(D). The Bayesian theorem relates these probabilities as follows:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

In epistemological terms this relation can be interpreted, as I have done when introducing the terms of the theorem (and will continue to do for the remaining section), as evaluating the probability of a hypothesis in light of a new piece of evidence, whereas in statistical data analysis the more technical notion of *"*conditioning [a] model on [the] data" (Martin, 2016, p. 9) is employed. The theorem above can be derived from the definition of conditional probability, the so called 'product rule' (Martin, 2016), which states that the joint probability of the two propositions *H* and *D* holding is equal to the probability of *D* holding given that *H*, multiplied by the probability of H (and thus also the probability of *H* given *D*, multiplied by the probability of *D*, as joint probability between two propositions is a symmetrical relation) (Chater and Oaksford, 2008).

$$P(H, D) = P(D|H)P(H)$$

$$P(H, D) = P(H|D)P(D)$$

From the product rule, it follows that

$$P(H|D)P(D) = P(D|H)P(H)$$

holds, which, by simply rearranging the terms of the expression, results in the initially defined statement in its usual form

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

The theorem is usually attributed to Rev. Thomas Bayes, an 18th century English mathematician and Presbyterian priest (Bellhouse, 2004), though some doubts have been raised on this notion (Stigler, 1983, also cf. Hacking, 2006). Independent of its historical origin, this theorem relating new evidence to prior information now forms the basis of an entire branch of statistical inference, which inherited its name from its supposed initial proponent, namely Bayesian statistics. The four probabilistic terms related to each other in the theorem are usually referred to in the literature as follows (my portrayal will follow, unless stated otherwise, Kruschke, 2014):

"$P(H \mid D)$" is termed the "posterior", which denotes the probability of an event after taking (the probability of) a given piece of evidence into consideration. "$P(D \mid H)$" is referred to as the "likelihood" and describes the plausibility of the data given the hypothesis under examination. Sometimes it is also termed the "sampling model" or "statistical model" in the literature on data analytics (Martin, 2016). "$P(H)$" represents the so-called "prior", that is, the probability of the

considered hypothesis independent of the evidence under examination (i.e., the probability prior to accounting for the evidence). This value can (in the absence of any knowledge about the hypothesis' probability) be a best guess which is then adjusted in light of the new evidence, or it can be the cumulative best guess based on prior evidence, i.e., the posterior of a previous Bayesian inference. It is common in the literature on Bayesian inference to state the relation of these three parameters as

$$P(H|D) \sim P(D|H)P(H)$$

thus, expressing them via proportionality rather than strict equality. This expression of the theorem leaves out the fourth term in the stricter formulation stated above, $P(D)$, which is referred to as the "marginal likelihood" or "marginalization" (Martin, 2016). The marginal likelihood ensures the probabilities sum to one, thus acting as a normalization factor. This parameter is often not of theoretical interest and merely ensures absolute comparability of results, as the proportions between relative values are usually what is relevant. (Martin, 2016). Mathematically, the marginal likelihood can be represented as

$$P(E) = \sum_H P(E|H) \times P(H)$$

This definition sums all possible hypotheses weighted by how probable they are (the prior) and how likely they would result in the observed evidence (the likelihood). The definition of the marginal likelihood somewhat differs for continuous variables, replacing the sum operator by an integral, but for our introductory purposes, the definition above shall suffice (the notion of a continuous hypothesis space will be introduced later on).

The shortened, proportional version of the theorem (leaving out the marginalization) captures the essence of Bayesian thinking, and the core property employed in many applications of Bayesianism in cognitive science, namely that – informally speaking – the posterior estimation of a hypothesis is a compromise between the initial estimation of said hypothesis and the plausibility of some new evidence under the hypothesis. Very strong priors will skew the posterior in their direction, just like very strong evidence will strongly influence the posterior estimation. In non-technical terms, when being unsure about a hypothesis one will have a greater degree of belief in its truth upon encountering strong evidence in its favor, whereas being very sure that a hypothesis is false, one will likely remain convinced of its falsity if only presented with weak evidence against it.

It follows that the role of the prior in Bayesian belief updating ("belief" here is not to be understood as denoting a psychological state, but merely an assigned probability in Bayesian inference) is substantial due to its the strong influence on the posterior estimation. There exists a vast array of literature on the problem of picking an adequate Bayesian prior (Gelman et al., 2020), but there is no single 'correct' solution to for this problem. Mathematically, 'anything goes' in Bayesian prior selection, and the choice one makes is determined more by practical considerations and epistemic values rather than mathematical necessity (cf. Gelman, 2020). Although this debate is philosophically rich and interesting, it is not of relevance for the topic under examination, which is why it shall not be explored any further

Hypothesis Spaces and Bayesian Models

So far, we have only considered the simple case of updating the probability of one particular hypothesis. What is usually under consideration in Bayesian analysis, however, is not a single hypothesis, but rather various hypotheses which may potentially account for the evidence at once (cf. Gelman et al., 2020). The entirety of hypotheses under consideration in such an analysis is termed the "hypothesis space" in Bayesian statistics (Colombo & Hartmann, 2017). It follows from Kolmogorov's (1933) axioms of probability theory (K1-K3) that the numerical degrees of belief assigned to the hypotheses in the hypothesis space must sum to one. These axioms state the following (Kolmogorov, 1933; DeGroot, 1975):

> *K1 Non-negativity*: For any event *A*, the probability of *A* is greater than or equal to zero $P(A) \geq 0$, i.e., probabilities are always non-negative.
>
> *K2 Unit measure*: The probability that (at least one) event in the sample space Ω (the space of all possible outcomes) will occur is 1, denoted as P(Ω)=1.
>
> *K3 Additivity*: For any sequence of mutually exclusive events *A1, A2, A3*, … the probability of their union is equal to the sum of their individual probabilities: $P(\bigcup_{i=1}^{\infty} A_i = \sum_{i=1}^{\infty} P(A_i)$.

As the hypotheses in hypothesis space cover every possible scenario and are mutually exclusive, it follows from K1-K3 that their joint probability must sum to 1.

The hypothesis space under consideration in a Bayesian analysis need not be finite, it can also contain infinitely many hypotheses. This is the case when estimating the parameter for a continuous variable given observed data, as a continuous variable can take on infinitely many values (Sprenger & Hartmann, 2019). When infinitely many hypotheses are examined, the probability distribution takes the form of a probability density function, i.e., a polynomial curve assigning a probability to intervals of possible values (Kruschke, 2014). As a probability density

function assigns probabilities to areas rather than singular values, it is the integral of the density function which must, following Kolmogorov's axioms, equal one. In this case, the key parameters of a Bayesian analysis, i.e., prior, posterior, and likelihood, do not take singular values, but are represented by a probability distribution over the hypothesis space (Gull, 1988).

As this usage of the term "hypothesis" ventures quite far from what the latter usually denotes in non-technical contexts, it is worth noting at this point, as pointed out by Jones & Love (2011, p. 173), that

> the term hypothesis should not be confused with its more traditional usage in psychology [or philosophy for that matter], connoting explicit testing of rules or other symbolically represented propositions. In the context of Bayesian modeling, hypotheses need have nothing to do with explicit reasoning, and indeed the Bayesian framework makes no commitment whatsoever on this issue.

Rather, a hypothesis in the Bayesian sense should be understood as an assumption about the value one or more parameters take on, based on which predictions about incoming additional datapoints can be made (cf. Kruschke, 2014). Although the 'traditional' sense of hypothesis can often be reconstructed or tested in this Bayesian framework (by representing hypotheses numerically and assigning them probability values), the two notions are not conceptually identical.

A Bayesian model, then, consists of the just described hypothesis space and an assignment of a prior (as defined above) to each hypothesis under consideration (Jones & Love, 2011). These two elements, in conjunction with a likelihood function that relates hypotheses to observed data, fully specify any given Bayesian model (Kruschke, 2014). Such a model can then be used to accommodate new evidence in accordance with the computational steps described above, thus generating a posterior estimate. Therefore,

> [a]ll Bayesian models share a common core structure. They capture a process that is assumed to generate some observed data by specifying: (i) the hypothesis space under consideration; (ii) the prior probability of each hypothesis in the hypothesis space; and (iii) the relationship between hypotheses and data. The prior of each hypothesis is updated in light of new data according to Bayesian conditionalization, which yields new probabilities for each hypothesis. (Colombo & Hartmann, 2017, p. 3)

An Illustrative Example

Let the foregoing, rather abstract characterization of the Bayesian inference process be illustrated by a simple example. Assume that we want to make a prediction about the probability of a randomly selected student of an academic philosophy department belonging to the continental tradition of philosophy rather than the analytical one (for the sake of illustration, we shall make the wildly inadequate assumptions that first, this distinction is clear and tenable, second that these two categories are exhaustive and third, that every philosophy student can clearly be classified as either one or the other, with no intermediary positions. It does not matter that these assumptions are almost certainly wrong, as the example merely serves the purpose of illustration). The prediction will be made on the basis of the student wearing a black turtleneck sweater, enjoying Sudokus, and smoking Parisienne cigarettes. First, we define the hypothesis space of our Bayesian model. In this case, only two hypotheses are considered, namely the randomly selected student belonging either to the analytical or to the continental tradition. Second, we must assign prior probabilities to all hypotheses under consideration. Let us assume that this hypothetical examination is carried out at the Sorbonne, an institution usually associated more with the continental school of philosophy, which may lead us to assigning a higher prior probability to the hypothesis that the student will belong to the continental rather than the analytical tradition. Let us (arbitrarily) quantify this probability as 0.7 for the hypothesis "x belongs to the continental tradition" and as 0.3 for the hypothesis "x belongs to the analytical tradition". Now, we must specify how the hypotheses in the hypothesis space relate to the observable data. Let us assume that we assign a probability of 0.7 to the conditional statement that the student smokes Parisienne cigarettes *if* they belong to the continental tradition.

$$P(P|C) = 0.7$$

Additionally, we can assign a probability of 0.3 that they like Sudokus and a probability of wearing a black turtleneck of 0.9 if they belong to the continental tradition.

$$P(S|C) = 0.3$$

$$P(T|C) = 0.9$$

Let us then assign the following probabilities to our available pieces of evidence conditioned on the student under investigation belonging to the analytical tradition:

$$P(P|A) = 0.2$$

$$P(S|A) = 0.8$$

$$P(T|A) = 0.3$$

This means that – assuming our student under investigation belongs to the analytical tradition – there is a 0.2 probability of him smoking Parisienne cigarettes, a 0.8 probability of him liking Sudokus, and a 0.3 probability of him wearing turtleneck sweaters. These conditional statements relate the evidence to the hypothesis space under consideration, as per Colombo & Hartmann (2017)'s third criterion stated above. Having these conditional probabilities in place, let us assume that we observe our student-under-investigation smoking Parisienne cigarettes. Using Bayesian inference, we relate the probability of the student belonging to the continental tradition given that they smoke Parisienne cigarettes to the likelihood $P(P|C)$, prior $P(C)$, and marginalization $P(P)$ using Bayes theorem as follows:

$$P(C|P) = P(P|C) \times P(C) / P(P)$$

Applying the definition of the marginalization term introduced earlier allows us to substitute $P(P)$ as follows

$$P(C|P) = P(P|C) \times P(C) / (P(P|C) \times P(C) + P(P|A) \times P(A))$$

Filling in the values we defined above, this results in the following calculation of $P(C|P)$:

$$P(C|P) = 0.7 \times 0.7/ (0.7 \times 0.7 + 0.2 \times 0.3) = 0.55 + 0.49 \approx 0.89$$

Thus, incorporating Bayesian belief updating, we arrive at a posterior probability estimation of our hypothetical student belonging to the continental tradition given our prior assumptions and observed evidence of about 0.89. We could repeat the same process for evaluating the posterior probability of the hypothesis that the student belongs to the analytical tradition, but based on the knowledge that the probabilities for all hypotheses in hypothesis space must sum to one, we can simply infer this from our previous result, leaving us with a posterior probability $P(A|P) \approx 0.11$. We have therefore updated our estimation for all the probabilities assigned to the hypothesis space in question. This estimation could be further refined by incorporating additional evidence, e.g. by observing the student in question enjoying themself while solving Sudokus or wearing turtleneck sweaters, as (as discussed before) by choosing posterior probabilities as priors for further inferences, one can account for accumulating evidence. By repeatedly incorporating new evidence into one's model, the influence of the original prior will wash out over time, and models with different priors incorporating the same evidence begin to converge in their posterior estimations. Incorporating the steps just laid out, we have set up an,

admittedly very simple, Bayesian model and updated the probabilities expressed therein in light of the available evidence.

Estimation Methods in Bayesian Inference

Although this illustrative example was very simple, more complex model estimations can become computationally quite costly or even intractable. This can be the case when the hypothesis space under consideration is not – as in the example above – discrete, but rather continuous. As laid out above, updating probability estimates for such continuous hypothesis spaces requires updating the probability density curve over the respective hypothesis space. The computational costs or intractability of updating complex Bayesian models stems from the difficulty of integrating these complex, multi-dimensional integrals (cf. Gelman et al., 2020). As Betancourt (2017, p. 4) puts it,

> for any nontrivial target distribution we will not be able to evaluate [the] integrals analytically, and we must instead resort to numerical methods which only approximate them. The accuracy of these approximations, and hence the utility of any given algorithm, however, is limited by our finite computational power.

A solution to this computational cost problem, as Betancourt states, is using estimation methods which are capable of approximating the relevant parameter values at a fraction of the computational cost (see also Gelman et al., 2020). Estimation methods include algorithms such as different variations of Monte-Carlo sampling, particle filters, or variational inference (Gelman et al., 2020). Understanding these approximators in detail is not required for the current examination. However, as the debate will touch upon such methods (or rather their interpretation in the field of Bayesian cognitive modelling) again in a later section, I shall briefly introduce the most commonly known and applied among them, the Markov Chain Monte Carlo method (henceforth MCMC), for the purpose of illustration.

MCMC is a method used for parameter estimation in Bayesian inference which uses Markov chains to sample from the posterior probability density function (Gelman et al., 2020). What does this mean? This estimation algorithm draws samples from the probability density function in order to approximate its parameter values (Gelman et al., 2020. This is the “Monte Carlo” part of the method. In MCMC, this sampling process focuses on regions of higher probability in order to use computational resources efficiently, as they are more informative for parameter estimation (Gelman et al., 2020). It does this by not generating samples at random (doing so would be referred to as Rejection-sampling, in which samples are generated at random and

independently of one another), which could be very computationally costly, but rather identifying samples within the distribution and using them as the basis for generating further samples (Betancourt, 2017). Thus, sample n+1 is not drawn independently of sample n, but rather depends on the latter. This relationship of dependence is mathematically captured using Markov chains, which are mathematical entities capturing a series of states and their transition probabilities, in which each state is dependent solely on its predecessor state (Betancourt, 2017). Betancourt (2017, p.10) once more captures this technical notion in rather simple terms when he states that a

> Markov chain is a sequence of points in parameter space generated by a Markov transition density […] that defines the probability of a new point given the current point […]. Sampling from that distribution yields a new state in the Markov chain and a new distribution from which to sample. […] Repeating this process generates a Markov chain that meanders through parameter space.

Therefore, MCMC uses Markov Chains to sample the posterior distribution, thus alleviating the need for solving the potentially complex integral of said distribution, and hence works as an estimator for the posterior parameters. It may be worth pointing out that – much like its supercategory, Monte Carlo approximation – MCMC does not refer to a single algorithm, but rather to a group of algorithms, containing methods such as for example the Metropolis-Hastings algorithm or Hamiltonian Monte Carlo, all of which follow the outlines of estimation just described (Betancourt, 2017).

Much more can be said about MCMC estimators, as well as other approximation methods which do not rely on Markov-Chain sampling, but for the purposes of the present discussion, the rudimentary understanding provided by the passage above shall suffice, as the details in advanced mathematics used for making these approximations are irrelevant for the argument presented within this thesis. The relevant takeaway from this cursory introduction should not be which specific mathematical techniques are best suited for approximating the posterior distribution in Bayesian inference, but rather *that* advanced mathematical algorithms are required for providing these approximations. This will be of relevance later when discussing the plausibility and epistemic merit (or lack thereof) of using Bayesian inference as a model for the operations carried out by the human mind.

Before delving into the applications of Bayesianism within cognitive science, let us briefly summarize our main takeaways about Bayesian inference in non-technical terms. Bayesian inference is a method for updating one's beliefs about the probability of a hypothesis (or series

of hypotheses, known as hypothesis space), by striking a compromise between one's prior expectations and the observed data, which is known as the posterior probability. This entails that the way the value of prior beliefs can have a strong influence on the resulting posteriors, and selecting faulty priors can lead to likewise faulty posterior estimations, despite choosing a probabilistically optimal mode of belief updating. As assessing the posteriors for complex hypothesis spaces can be very computationally costly or intractable, a series of approximation methods exist, which aim to estimate the true parameter values at a fraction of the computational cost.

Having laid out the fundamental tenets of Bayesian belief updating in sufficient detail for the purposes of our philosophical investigation, let us continue the discussion by introducing the implementations of this framework in contemporary cognitive science. This will be the topic of the following section.

Bayesianism in Cognitive Science

As stated in the introduction, applications of Bayesianism are manifold, the most obvious one being the application of Bayesian inference methods in statistics, data science, and natural or social science to the end of statistical data analysis, where they have been by now established as a genuine alternative to the traditional frequentist NHST approach (Faulkenberry et al., 2020). Under this relatively new Bayesian paradigm in inferential statistics, Bayesian alternatives have been developed to more traditional statistical methods, such as the Bayesian t-test (for comparing the means of two samples), the Bayesian ANOVA (for comparing the means of various samples) or the Bayesian regression (for fitting predictive data models) (Faulkenberry et al., 2020). Although these developments have greatly influenced cognitive science, this type of *inferential Bayesianism*, as I shall term it (for it deals with the manner in which inferences about scientific hypotheses are made), shall not be discussed in any detail in the present thesis, the reason being that, although pertaining to cognitive science and philosophy of science, this matter is different from and largely independent of the problem under examination therein. Inferential Bayesianism deals with the empirical assessment of scientific hypotheses, independent of the subject matter these hypotheses are about. Although one could use Bayesian statistics to test hypotheses about the nature of the human mind, e.g. in the field of psychology, the application of these statistical methods will not drastically differ from their application in the disciplines of epidemiology, economics, genomics or any other scientific discipline engaging in statistical hypotheses testing.

Rather than dealing with the Bayesian school of statistical hypothesis testing, the present thesis shall limit its scope to the discussion of Bayesian inference as a model of cognition, an approach I shall term *cognitive Bayesianism* for our present purposes. To avoid any possible confusion and clearly demarcate the scope of our investigation, let us explicitly spell out the difference between cognitive and inferential Bayesianism once again. Cognitive Bayesianism claims that cognition operates (in some way or another) by adhering to the principles of Bayesian inference (or at least in a similar enough fashion to make the latter useful as a model of it). It is thus a paradigm of cognitive science alongside classical cognitivism, connectionism, or dynamical systems theory, and does not, inherently, specify any means for evaluating its hypotheses. Inferential Bayesianism, on the other hand, is a commitment to a specific manner of evaluating scientific hypothesis independent of the content of these hypothesis. Thus, the two frameworks are – at least in principle – entirely independent of one another. One can endorse or reject either of them without being committed to the same stance in regard to the other. Having narrowed

the scope of our examination, let us continue by establishing the core tenets of cognitive Bayesianism.

Like previously stated, cognitive Bayesianism is a paradigm of contemporary cognitive science which claims that cognition can be understood as operating using (or at least in accordance with) Bayesian principles. There have been attempts to account for a wide variety of cognitive phenomena using a Bayesian framework, such as, as summarized by Colombo & Hartmann (2017), categorization, causal reasoning, judgement, decision making, reasoning about other agents' beliefs and desires, perception, illusions, psychosis, motor control, and language.

The usage of a collective term, cognitive Bayesianism, to refer to these developments in different subfields of cognitive science should not be misunderstood as implying that cognitive Bayesianism can be understood as a unitary movement or school of thought sharing theoretical or methodological assumptions beyond the usage of Bayesianism in some capacity. Rather, cognitive Bayesianism should be thought of as an umbrella term denoting a range of partially independent and possibly contradictory strains of scientific research which are united solely by their commitment to the idea that cognition operates (or can be modelled as operating) in a Bayesian manner to some extent. To which extent this is the case, e.g. whether Bayesian inference is the guiding computational principle underlying all of cognition or whether it is limited to certain cognitive domains, or the exact way in which the Bayesian models supposedly carried out by the mind are construed, can be – and are in fact – points of disagreement between individual researchers (cf. Fink & Zednik, 2017). Within this loosely connected research program of cognitive Bayesianism, I shall introduce the further division between what I shall term *descriptive* and *normative* Bayesian approaches.

I shall refer to any cognitive Bayesian thesis as a '*descriptive Bayesian'* approach if it aims to describe or theorize about a cognitive process (broadly construed) as operating according to Bayesian principles without normatively evaluating the validity or rationality of said process. This descriptive Bayesian stance shall be contrasted with the '*normative Bayesian'* approach, by which I shall denote any attempt to evaluate the rationality or validity of a cognitive process from a Bayesian perspective to establish its optimality and/or rationality. This stance is characteristic of many Bayesian approaches to human reasoning which aim to re-analyze seemingly faulty inferences commonly made by humans from a Bayesian point of view, often with the explicit goals of demonstrating their rationality if the inferential task in question is framed from a Bayesian perspective. This approach, therefore, differs from the purely descriptive one in that it explicitly aims to provide a normatively optimal solution to an

inferential problem which is then, in a subsequent step, related to observed behavior, instead of merely using the mathematics and/or concepts of Bayesian statistics to describe, theorize about, or model cognition.

The most influential and vocal proponent of the normative approach is the psychologist John Anderson (1991, 2013, 2014), who refers to it as “rational analysis”. Rational analysis operates by specifying the environment a cognitive system operates in and the inferential problems the system has to solve within it, and then attempts to demonstrate that the solutions produced by said system are optimal from a Bayesian viewpoint (a more extensive characterization of rational analysis will follow shortly). Whereas a purely descriptive approach, in the way I have introduced the term, would limit itself to propose a thesis regarding the nature of the cognitive process leading to an observed behavior, rational analysis, as performed by Anderson, adds the claim of optimality to this modelling goal.

Summed up in a slogan, one could characterize the difference between both approaches as follows: Whereas descriptive Bayesianism attempts to describe or model the process underlying cognition in Bayesian terms, normative Bayesianism aims to demonstrate that cognition is (close to) optimal if evaluated by Bayesian standards. It follows from this distinction that Bayesian accounts of maladaptive forms of cognition, mental disorders, will typically be descriptive, simply due to the primary argumentative goals of both approaches. Whereas it is intuitively interesting to see what process may lead to disordered forms of cognition, little of interest will be said by stating that they deviate from optimal standards of rationality, as, to some extent, this is what qualifies them as disorders in the first place. However, normative Bayesianism could add something of interest to the discussion of disordered cognition by demonstrating how forms of cognition typically classified as maladaptive and disordered do in fact, contra prima facie assumptions, adhere to normative standards of optimality and rationality. I am not aware of any literature devoted to demonstrating this (if it exists, it certainly constitutes a small fraction of the research performed in Bayesian cognitive science), but it strikes me as an intriguing possibility.

Before diving into some illustrative examples from both categories which will hopefully further illuminate and clarify the respective goals and methods of both approaches, it should be noted that the distinction between descriptive and normative Bayesianism should not be thought of as exclusive and jointly exhaustive. Rather, they should be considered opposite ends of a spectrum on which individual models and theories lie. The claim that some form of behavior is Bayes-optimal, given certain niche constraints, is normative based on the taxonomy just laid out.

However, it will inevitably still contain descriptive elements, as it still purports that Bayesian inference accurately captures the behavior in question, as normative accounts aim to demonstrate that human cognition is optimally suited for some niche (however that niche may be specified), necessarily, they must demonstrate that Bayesian inference is a fitting characterization of the cognitive process under investigation at least to a minimal extent, even if they do not commit to the notion that actual Bayesian inference was involved in bringing it about. Likewise, any descriptive account must, in some minimal way, also be normative, as it claims that Bayesian inference is part of a suitable model of or theory about a given cognitive process. As Bayesian inference is, mathematically speaking, the optimal way of incorporating new information into probability ratings, any cognitive process modelled by Bayesian inference must at least in this trivial sense be optimal – although this need not entail any further claims about optimal niche adaption or similar claims of this sort (the ensuing discussion of Bayesian models of mental disorders will demonstrate this clearly). To briefly illustrate this point, consider a cognitive system – for the purpose of our example, let us take a college student living in a shared flat – assigning a very high, bordering on absolute certainty, prior probability to the hypothesis that their roommate is secretly an alien from a distant galaxy. Even if incorporating a piece of counterevidence into this assessment in a Bayes-optimal fashion, the resulting posterior estimate will still assign a very high probability to the alien-hypothesis, simply due to the manner in which Bayesian inference incorporates prior assumptions into its estimates. As, mathematically, there are basically no constraints on how to set one's prior for a Bayesian inference, the resulting posterior probability estimation will be Bayes-optimal, but it will certainly not constitute a form of optimal niche adaptation or a thought anyone would feel comfortable describing as rational.

In light of the impossibility of strict separation between both frameworks, the categories *normative* and *descriptive* are best considered not as mutually exclusive research frameworks, but rather as aiming to answer slightly different questions, possibly using the exact same mathematical structure to model cognitive processes. The difference between them is the claim that is supported using this structure, i.e., a claim limited to capturing the inner workings of the minds versus a claim about its normative evaluation.

The reader may wonder at this point – given the strong constraints on the mutual exclusivity of the introduced subcategories of cognitive Bayesianism – what the benefit of establishing the distinction in the first place may be. Without going into too much detail yet, the reason is that, as both approaches follow slightly different argumentative goals, it will be necessary to give different evaluations in regard to their explanatory power. Thus, the distinction – shaky as it

may sometimes be in the classification of concrete Bayesian models – aims to allow for a more subtle and fine-grained evaluation of the field from a philosophy of science perspective, as opposed to falsely treating it as homogenous in its methods and assumptions. To this end, a compromise must be struck between generalizing over relevant differences between distinct research subfields and taking into account every difference between individual approaches and thus losing the generalizability of the proposed claims. My hope is that the introduced dichotomy is able to fulfill this desideratum.

Having now established this tentative classification of approaches within the Bayesian literature, let us discuss some examples from each category, while bearing in mind that the distinctions applied are not mutually exclusive in any absolute sense.

Descriptive Bayesianism

Possibly the most influential research program in the wider field of cognitive Bayesianism is that of predictive processing. Predictive processing is a set of assumptions about the fundamental working principles of the mind, only some of which are directly related to Bayesianism. Under the predictive processing framework, the mind is thought of as operating in a hierarchical and interactive (i.e., both top-down and bottom-up) fashion in which predictions are generated at each level of the hierarchy about the state of the respective next lower level of the hierarchy, and deviations from these predictions are fed back up the hierarchy to generate more accurate predictions the next time. At the basis of this predictive pyramid lie basic forms of sensory perception, whereas high-level, conceptual forms of cognition reside at the top (Clark, 2013; 2015). To use a simple example, expecting to visually perceive an elephant (a high-level concept) will generate a cascade of predictions descending the hierarchical pyramid, such as predictions about concrete arrangements of geometric bodies, which will, further down the pyramid, generate predictions about specific arrangements of detected edges, and so on. Discrepancies between the predicted states at each level and the encountered feedback from the lower levels then lead to the predictions about lower-level states being adjusted in light of the received feedback information. In this way, prediction error is backpropagated, ascending the hierarchy and influencing the predictions made at the highest, most abstract level. Although this example is somewhat simplified, it captures the essence of hierarchical predictive processing. Hohwy (2020) summarizes this hierarchical processing as follows:

> Predictions are messages that descend in the internal structure of the system, to be tested against the incoming, ascending sensory signal. Any discrepancy between prediction and

> sensory signal gives rise to prediction error messages that then ascend from the sensors and upward in the system [….} This basic notion of predictive coding provides an efficient message passing scheme because prediction errors carry information about the quality of the prediction and are used to update the model, leading to new predictions […]. (Hohwy, 2020, p. 210)

The fed-back information is termed "prediction error" in the predictive processing framework, with one of the core claims of the latter being that the minimization of prediction error (referred to as the "minimization of free energy") is the ultimate principle underlying cognition (Clark, 2013, 2015). This principle is termed the "free energy principle", an in-depth exploration of which would go beyond the scope of the present examination, due to its notorious complexity and breadth of scope (Clark, 2013). Thus, to linger on the example of visual processing a bit longer, as opposed to e.g. Marr's (1982) influential theory of visual processing, in which increasingly complex visual patterns are fed up a perceptual hierarchy in a bottom-up fashion, the predictive processing account of visual perception starts at the top of the hierarchy, where predictions about incoming sensory stimuli are generated and fed down the perceptual hierarchy, with the deviances from these predictions (the prediction errors) being propagated back up the hierarchy.

Notice the stark parallel of each level's prediction-and-updating-cycle to Bayesian inference, which likewise starts with a prediction generating model which is then updated in light of the newly accumulated evidence, thus incorporating the newly acquired evidence. This parallel is not coincidental. Rather, the interactions between individual levels of the hierarchy are explicitly construed as following Bayesian principles in the predictive processing framework (Clark, 2013; 2015). The core proposal – shared between Predictive processing and other Bayesian accounts of the mind – is that the information available to a cognitive system (be it sensory, interoceptive, etc.) for generating internal models of the world is uncertain, and hence only probabilistic inferences about its distal causes possible. Bayesianism provides the proper framework to account for this integration of uncertain information while also taking into account its degree of uncertainty. This is achieved by operating not on discrete probabilities, but rather on probability distributions (see the introductory section on Bayesian inference), which can represent the degree of uncertainty by the shape the distribution takes, especially its dispersion from the central tendency. As Knill and Pouget (2004, p. 713) put it,

> The fundamental concept behind the Bayesian approach to perceptual computations is that the information provided by a set of sensory data about the world is represented by a

> conditional probability density function over the set of unknown variables – the posterior density function.

For the concrete example of depth perception, they specify this type of uncertainty representation as follows:

> A Bayesian perceptual system […] would represent the perceived depth of an object, for example, not as a single number Z but as a conditional probability density function p(Z/I), where I is the available image information (e.g. stereo disparities). Loosely speaking, p(Z/I) would specify the relative probability that the object is at different depths Z, given the available sensory information […]. [A] Bayes' optimal system maintains, at each stage of local computation, a representation of all possible values of the parameters being computed along with associated probabilities. (Knill & Pouget, 2004, p. 713)

What makes the predictive processing account descriptive, following the classification scheme introduced at the beginning of the present section, is that it is used indiscriminately to account for both typical and atypical cognition, employing the mathematical framework of Bayesian computation to provide an abstract model of cognition irrespective of its optimality or niche adaptiveness. Pellicano and Burr (2012), for example, have proposed that the perceptual atypicalities associated with autism spectrum disorder can be accounted for by a Bayesian account within the predictive processing framework. What they propose is that overly unspecific priors (so-called *hypopriors*) in autistic individuals lead to a relative over-weighting of bottom-up components of perception and cognition in shaping their world model. The symptoms this core idea is supposed to explain include hypersensitivity to certain stimuli (typical examples include brightly colored objects or patterns of fractioned light), at the same time as seeming obliviousness (i.e., *hypo*sensitivity) to other sensations. Pellicano and Burr (2012) claim that these typical perceptual symptoms of autism can be accounted for by thinking about priors of sensory perception as enhancing or attenuating biases about the perceptual characteristics of sensory information. Thus, because prior expectations about sensory stimuli are very unspecific, the relative influence of stimulus-inherent characteristics contributes more – relative to neurotypical populations – to the process of perception. In the previous chapter introducing Bayesian inference, it was stated that – informally speaking – the posterior in a Bayesian inference can be thought of as a compromise between the prior expectations (in the context of the Bayesian mind, think "mental world model") and the likelihood (think "incoming stimuli"). By attenuating the specificity of the priors, that is, by increasing the dispersion of the probability distribution over the hypothesis space, the relative influence of the likelihood

increases. This is the case because this unspecific prior assigns a higher probability to a wider range of values in hypothesis space, as opposed to being concentrated to a relatively small area to which a high probability is assigned, thus making less specific predictions about which values to expect, whereas a very specific prior expectation (a very narrow probability distribution) constitutes a strong bias exerting a relatively larger influence on the resulting posterior probability distribution. In short, they "propose that altered autistic perception results from atypicalities at the level of the prior – either in its construction or in combining appropriately with sensory information – yielding unusually attenuated priors or hypo-priors" (Pellicano & Burr, 2012, p. 9).

Notice how this account uses the framework of Bayesian model updating to capture the characteristics of a set of cognitive processes without committing to a normative evaluation regarding the optimality of these processes. Quite the contrary, the authors display how the mathematical framework of Bayesianism can be used to capture markedly distinct processes – sensory perception in autistic and neurotypical populations - by adjusting the values of certain parameters in the mathematical framework employed for modelling these processes, in this case priors, using the framework to account for cognitive processes that can be considered at least partially maladaptive (for example in the case of so called 'meltdowns' due to sensory overload in autistic individuals, cf. Yalim & Mohamed, 2013).

Besides permanently atypical modes of cognition such as autism, there also exist predictive processing accounts of seemingly faulty types of perception situated within otherwise typical cognitive systems, namely, perceptual illusions. Consider, as an example, the rubber hand illusion (cf. Ramakonar et al., 2011). In this well-known perceptual illusion, participants develop a sensation of body ownership pertaining to a rubber hand placed in front of them. This state is achieved by hiding the participant's real hand out of sight (e.g. by placing it behind a screen) and positioning a rubber hand in the position the real hand would usually be in front of the participant. The hidden real hand and the visible rubber hand are then stroked in synchrony, e.g. using a brush, by the experimenter. The visual perception of the hand being rhythmically stroked and the sensation of the rhythmic stroking of one's own hand can then lead to a sense of body-ownership over the artificial hand. How does predictive processing account for this effect? Predictive processing, by its very nature, maintains that the integration of the rubber hand into the body schema must be the result of prediction error minimization. Matamala-Gomez et al. (2020, p. 4) summarize how this may happen:

> In the case of body ownership illusions [such as the rubber hand illusion], altered body perception results from a self-representation that is updated dynamically by the brain to minimize sensory conflicts (i.e., the differences between the predictions about sensory data and the real sensory data at any level of the hierarchical model). [In] body ownership illusions our brain tries to minimize the "surprise" through the predictive coding scheme, when encountering a signal that was not predicted, it will generate prediction errors and will update the model in order to minimize the differences between the predictions about sensory data and the real sensory data at any level of the hierarchical model (e.g., synchronous visuo-tactile feedback in the RHI study…). Therefore, the subjects will update their internal body representation.

Applied explicitly to the rubber hand, the core idea is that there is a high prior probability that visual and tactile stimulation in perfect rhythmic synchrony are caused by the same external source of stimulation, and – on the flipside – a very low prior probability that such synchrony between two modalities is unrelated, leading to Bayesian "surprise" (a technical term describing the deviance between model predictions and observed evidence) upon encountering perfect synchrony between one's visual sensation of a seemingly external, hand-like object being stroked and one's tactile perception. To minimize this surprise, the object is incorporated into the body model, thus minimizing the degree of surprise in accordance with the free energy principle. It may be interesting to note at this point that Shaun Gallagher and colleagues (Gallagher et al., 2022) have critiqued this account as failing to explain why high-level knowledge, in this case knowledge that the rubber hand is not part of one's own body, fails to make the illusion disappear, as high-level predictions should – following the predictive processing account – have a strong influence on the outcome of a cognitive process. This ties into a common critique of predictive processing, namely that no clear predictions can be derived from it, and that it is seemingly capable of accounting for phenomena of any kind, thus making it hard, if not impossible, to collect evidence against it (cf. Hohwy, 2020). While Gallagher et al.'s critique is an interesting one, it shall only be mentioned in passing, as the aim of the current section is merely to introduce predictive processing, not (yet) to critique it. However, the more general criticism of predictive processing failing to establish clear empirical consequences will, in a different form, come up again at a later point.

Just like in the case of autism, the rubber hand illusion is an example of a cognitive process that can be regarded, in a certain sense, as not rational and arguably not optimal in its niche adaptability (due to the simple fact that it is, by its very nature as an illusion, non-veridical),

hence making the Bayesian account of it descriptive in nature. These two examples were chosen – besides serving as a general introduction to predictive processing – to display this paradigm's ability to also incorporate types of atypical or abnormal cognition (though I shall at this point refrain from judging its ability to actually explain them). It should nonetheless be pointed out once more that predictive processing is not a framework of atypical cognition, but rather of cognition in general, and that the selection of these two examples was to serve the argumentative narrative of the present paper, rather than serving as a pure introduction to predictive processing, for which more general and influential models, such as that of visual or language perception, may be better suited.

A great many more cognitive phenomena could be listed in addition to one's discussed so far, as proponents of predictive processing have made great efforts of re-contextualizing various mental processes within this framework (Clark 2015), but for now, this shall conclude our introduction to predictive processing and the descriptive approach to cognitive Bayesianism in general. The ensuing section shall continue the discussion of Bayesian models of cognition by treating what has been termed "normative Bayesianism" for the purposes of our present investigation.

Normative Bayesianism

Normative Bayesian approaches, in the sense I have introduced the term, evaluate human behavior with respect to a standard of optimality/rationality. John Anderson (1990, 1991), as stated previously, the perhaps most influential proponent of this approach, termed it "rational analysis", and developed it into a distinct research framework within contemporary cognitive science. The aim of this framework, as aptly summarized by Fink & Zednik (2017, p. 2), is "to formally characterize a cognitive system's behavior as an optimal solution to a particular probabilistic inference task in the environment". The core idea of this framework, similar to that of predictive processing, is that a cognitive system's environment is probabilistically structured, i.e., the inflow of stimulation is ambiguous and compatible with different interpretations, and the computational aim of the system is to make probabilistic inferences about the external world based on this ambiguous stimulation. What is distinct about the rational analysis approach to cognition, and the normative approach more generally, is that it puts its focus more on this probabilistic structure of the environment itself, as well as optimal solutions for dealing with it, than on the nature of the cognitive system faced with this cognitive task (Anderson, 1990; Fink & Zednik, 2017). The idea is to precisely specify the environment's probabilistic structure in mathematic terms and then devise a normatively optimal solution to

the problem thus posed to the system making inferences about said environment (Anderson, 1990; Fink & Zednik, 2017). This normatively optimal solution is then compared with actual (typically human) behavior and used as a model of cognition (Anderson, 1990; Fink & Zednik, 2017). Anderson (1990) states this core idea underlying rational analysis as follows:

> We can understand a lot about human cognition without considering in detail what is inside the human head. Rather, we can look in detail at what is outside the human head and try to determine what would be optimal behavior given the structure of the environment and the goals of the human. (Anderson, 1990, p. 3)

In a subsequent publication, Anderson (1991) explicitly spells out a set of six distinct steps for the formulation of a rational analysis for a given cognitive task, which are:

> 1. Specify precisely the goals of the cognitive system.
>
> 2. Develop a formal model of the environment to which the system is adapted.
>
> 3. Make the minimal assumptions about computational limitations.
>
> 4. Derive the optimal behavioral function given 1-3 above.
>
> 5. Examine the empirical literature to see whether the predictions of the behavioral function are confirmed.
>
> 6. Repeat, iteratively refining the theory.
>
> (Anderson, 1991, p. 473)

Although not every Bayesian model of cognition which evaluates human behavior under a normative perspective strictly follows this iterative process laid out by Anderson, it does reveal some characteristic features of the normative approach, which will become evident when discussing some examples from the literature. What is common to all normative approaches is that they start by specifying an optimal solution and then – in a subsequent step – compare this optimal solution to actual observed behavior, as opposed to specifying a model of the operating principles underlying cognition and then deriving how a system following these principles would interact with its environment.

In short, by focusing on the environment, the rationalist program aims to predict a system's behavior "from the structure of the environment rather than the structure of the mind" (Anderson, 1991, p. 474), by identifying the optimal solution for navigating this environment and using this optimal solution as a model of human performance. This strict focus on the tuple

of environment and optimal behavior explicitly and deliberately (cf. Anderson, 1991) omits any commitments regarding the inner structure of the mind, whereas the descriptive approach aims to identify operating principles of cognition, albeit, at least in the case of predictive processing, at a rather abstract level.

To further illuminate the normative Bayesian approach, let us discuss some concrete examples from the literature. The first among these shall be taken directly from Anderson's (1991) previously cited seminal contribution to the rational analysis discussion and consists in a rational analysis of human memory. Following this account, the primary aim of memory is to retrieve the information most relevant for one's current context. Anderson (1991) gives the trivial example of needing to memorize where one has parked one's car. He frames this retrieval process as a cost-benefit analysis, where an information processing system needs to balance the cost of retrieving a particular memory with the benefit of having access to the thus retrieved information. In doing so, he makes the (plausible, although not further justified) assumption that there is a cost for the processing system attached to retrieving a memory, with the benefit being – trivially – the usefulness of the retrieved information for the current context. In his model, for any given memory A, $p(A)$ is the probability of the memory being relevant for the current context. Therefore, "[a] rationally designed Information-retrieval system would retrieve memory structures ordered by their probabilities $p(A)$ and would stop retrieving when $p(A)G < C$ (Equation 1)" (Anderson, 1991, p. 474), with G being the value of achieving the goal associated with the memory retrieval and C being the retrieval cost. Following this basic setup, Anderson argues that a rational analysis of memory needs to elucidate how to optimally estimate the factor $p(A)$, which is done based on the two (sets of) cues *context independent history of retrieving the memory in question*, formalized as H_A and *set of cues given by the current context*, formalized as Q, with the members of the set being denoted by a series of indices i, an analysis he bases on computer information retrieval processes. H_A represents the frequency at which a memory is retrieved, as well as how recent the last retrieval was. The underlying assumption is that more frequently and more recently retrieved memories have a higher probability of being relevant for the current context. Anderson does not specify in much detail the explicit content of the set Q, other than stating the individual cues i denote the "associative strengths between cues and memories" (Anderson, 1991, p. 475). These associative strengths are informed by the system's processing history, i.e., how often a memory was relevant in the past when cue i or a cue similar to i was present.

Following these assumptions, "estimation can be characterized as a process of inferring a Bayesian posterior probability of being needed, conditional on these two sources of

information" (Anderson, 1991, p. 475). Based on the formalizations just introduced, Anderson specifies the way of arriving at *p(A)* given *Q* and H_A, i.e., $p(A|Q \,\&\, H_A)$ as Bayesian odds:

$$\frac{\mathrm{p}(\mathrm{A}|\mathrm{Q}\,\&\,H_A)}{\mathrm{p}(\bar{\mathrm{A}}|\mathrm{Q}\,\&\,H_A)} = \frac{\mathrm{p}(\mathrm{A}|H_A)}{\mathrm{p}(\bar{\mathrm{A}}|H_A)}\; x \prod_{i \in Q} \frac{P(i|A)}{P(i|\bar{A})}$$

Bayesian odds refer to *relative likelihoods* of two (contradictory and thus mutually exclusive) hypotheses given the observed data, which is computed by dividing the posterior probability of one hypothesis by the posterior probability of the other (cf. Gelman et al., 2020). This, so Anderson's line of reasoning, is the optimal solution to the inferential problem a cognitive system is faced with in regard to memory retrieval, which can serve as a mathematical model of human performance. His account ends with this formal specification, and does not, importantly, contain any additional elements linking this mathematically optimal solution to any cognitive processes bringing it about in human cognizers. In a similar vein, Anderson (1991) introduces rational analyses for causal inference, problem solving, and categorization, but for the purpose of illustration, the example of memory shall suffice.

Not all accounts classifiable within the category "normative Bayesian" explicitly follow Anderson's six step scheme of rational analysis. What is commonly found in the literature are papers approximately adhering to the following structure (which, other than Anderson's six-step scheme, is merely an informal characterization of loosely connected research approaches not necessarily connected by a determinate shared methodology):

1.) Identify supposedly faulty human modes of inference (e.g. intuitive logical or statistical reasoning) described in the previous literature
2.) Re-analyze the inferential problems described therein from a Bayesian perspective (i.e., provide a normative Bayesian solution to the problem)
3.) Attempt to show that human inference is (close to) optimal under this Bayesian re-interpretation of optimality

Such work is, as becomes immediately apparent when contrasting this schema with the previously introduced rational approach, not entirely distinct from Anderson's rational analysis approach. Quite the contrary, Anderson himself states as one of the goals of rational analysis to demonstrate that human performance in certain tasks is often wrongfully characterized as suboptimal or irrational:

> Many characterizations of human cognition as irrational make the error of treating the environment as being much more certain than it is. The worst and most common of these errors is to assume that the subject has a basis for knowing the information processing

> demands of an experiment in as precise a way as the experimenter does. What is optimal in the microworld created in the laboratory can be far from optimal in the world at large. (Anderson, 1991, p. 473)

As such, even normative Bayesian models not explicitly following the six-step scheme of rational analysis are closely related to it and share some of its core assumptions and methods. What makes these approaches normative is their commitment to what Jones and Love (2011, p. 170) have called "the primary goal of much Bayesian cognitive modeling", namely, to "demonstrate that human behavior in some task is rational with respect to a particular choice of Bayesian model". Typically, models in this framework respond to and build upon an extensive body of work in cognitive psychology characterizing human inference as faulty in the sense that it deviates from a normative standard of optimality as dictated by formal logic or probability theory (see for example Tversky & Kahneman, 1974; Tversky & Kahneman, 1981), followed by attempts to account for this seeming deficiency by a psychological theory. Many examples discussed include alleged failures of human participants to correctly reason deductively, e.g. by failing to validly make inferences using the logical conditional (cf. Oaksford & Chater, 2008). Supposedly, day-to-day human reasoning deviates from the norms of formal logic, e.g., by failing to recognize (normatively correct) modus tollens arguments, that is, arguments of the following structure

$$P \rightarrow Q$$

$$\neg Q$$

$$\therefore \neg P$$

as valid or by accepting invalid inferences of the type "denying the antecedence", i.e.,

$$P \rightarrow Q$$

$$\neg P$$

$$\therefore \neg Q$$

or "accepting the consequence", formalized as

$$P \rightarrow Q$$

$$Q$$

$$\therefore P$$

(Oaksford & Chater, 2008).

Likewise, traditional psychological research indicates that humans use erroneous probabilistic reasoning. One example of this are so-called framing effects, in which the context and manner in which a problem is presented influences the probabilistic inference that is drawn about it (Sher & McKenzie, 2008; Tversky & Kahneman, 1981). In one of their influential studies, Tversky and Kahneman (1981) confronted participants with the following hypothetical problem (the "Asian disease" problem, as it is usually referred to, cf. Sher & KcKenzie, 2008): A rare disease is expected to kill 600 people if no intervention is taken. Two interventions are possible, which differ in their outcomes. To some of the participants, these outcomes were described as follows:

> *Intervention 1: This intervention will save 400 out of the 600 people.*
>
> *Intervention 2: With a 1/3 probability, this intervention will save all 600 people, but with a 2/3 probability, this intervention will save no-one.*

The other participants were given the following description of the two possible outcomes:

> *Intervention 1: This intervention will kill 200 out of the 600 people.*
>
> *Intervention 2: With a 1/3 probability, this intervention will kill nobody, but with 2/3 probability, this intervention will kill all 600 people.*

Despite both manners of describing the two interventions being mathematically equivalent, participants reliably preferred intervention 1 over intervention 2 in the first framing, but intervention 2 over intervention 1 in the second framing, thus indicating that the way in which the information is presented influences their decision making rather than the probabilistic character (a sure loss versus a possible loss of greater magnitude) of each intervention.

Many more examples of seemingly faulty reasoning have been discussed in the psychological literature over the years, of which these two merely serve as exemplary illustrations. Traditional psychological accounts of these findings start with the (intuitively plausible) assumption that these cases are examples of human reasoning being in some way deficient, irrational, or deviating from optimality, and that this deviance from optimality needs theoretical accounting for by a psychological theory (Tversky & Kahneman, 1981 being among the most prominent ones).

Following the normative Bayesian goal of demonstrating that seemingly irrational or suboptimal decisions do often (despite claims to the contrary) in fact adhere to normative standards of rationality if the latter are understood properly, many researchers working under

the Bayesian framework have rejected this traditional approach. In place of it, they attempt to propose an alternative "optimal" solution from a Bayesian point of view, and to demonstrate that humans do, in fact, perform (near) optimal from this perspective (Chater & Oaksford, 2008). Thus, they argue that the results that human reasoning (broadly construed) is suboptimal stems from applying the wrong normative standards, and that properly construed standards of rationality will yield the result that human behavior is, in fact, closely approximated by this normative standard. Once more, it should become apparent that this approach closely resembles Anderson's rational analysis framework, despite not strictly following its methodological plan. For the sake of illustration, let us take a look at a Bayesian re-interpretation of a seemingly faulty mode of human inference/decision making, namely that of framing effects in probabilistic reasoning.

Sher & McKenzie (2008, cf. also Sher & McKenzie, 2006) argue that the core problem leading to the erroneous classification of human probabilistic reasoning as irrational stems from the researchers making such claims not taking into account the actual amount of information available to test subjects in their decision process. If subjects use subtle information not accounted for by the researchers to inform their probabilistic assessments, this will – trivially – lead to divergences between the actual assessments and the assessment deemed rational by the researchers. This will be the case even if both the researchers devising the optimal solution to the problem they pose and their participants attempting to solve said problem adhere to the same standards of rationality, simply due to the optimal solution in both cases making use of different information to arrive at a decision. In the disease problem outlined above, this leads – so the argument – to the normative model applied in traditional research (which, to reiterate, states that both descriptions of the possible outcomes are equivalent) being inadequate, as it does not account for the full information available to the participants.

> [I]t turns out that, even in the simplified situations experimenters have specially contrived, the normative model used in their analysis has been inadequate. Even in this simple case [i.e., the disease problem], the experimental situation makes subtle information available which should matter to the normative analysis, but which has not been considered in the interpretation of experiments. (Sher & McKenzie, 2008, p. 80)

Wherein consists this allegedly unaccounted for information? Sher and McKenzie argue that it consists in the kind of additional pragmatic information a speaker normally wishes to convey by selecting a particular framing (e.g., ¼ full, 75% empty, ¾ empty, 25% full, etc.) in day-to-day conversation. Thus, so Sher's and McKenzies assumption, the frame chosen to express a

given proposition will usually convey additional information about a speaker's attitudes towards said proposition. They formalize this assumption by considering a simple hypothetical case with two potential frames of a given proposition, which they term *A* and *B*. Consider a background condition *C* which makes it more probable for frame *A* to be chosen than it not being chosen, so that $P(A|C) > P(\sim A|C)$ holds. Assume further that the receiver of the conveyed information knows about the probabilistic relevance of condition *C*. Following Bayesian reasoning, the receiver would therefore be justified in assigning a higher probability to *C* holding based on the proposition being expressed via frame *A* as opposed to *B*, as $P(C|A)$ is greater than $P(C|B)$.

> Therefore, a listener, aware of the regularity that relates the background condition C to the speaker's choice of frame, may rationally infer a higher probability of C when the speaker says 'A' than when the speaker says 'B'. If C is choice-relevant, we would expect a rational actor to use this information, and therefore potentially to respond differently depending on the speaker's choice of frame. *When no choice-relevant background condition C meeting the above description exists, two frames are information equivalent. Otherwise, they are information non equivalent.* [emph. added] (Sher & McKenzie, 2008, p. 83)

In view of this probabilistic structure of expressing a proposition in a given frame, Sher and McKenzie's hypothesis is that, ceteris paribus, people are more likely to choose a negatively-valenced frame for expressing a proportion (e.g. mortality rate instead of survivor rate of a treatment) when the proportion is atypically high, i.e., lies above some reference point of normality. For example, following this conjecture, "35% of attendees did not enjoy the performance" is more likely to be chosen as a frame for expressing the proportion of satisfied to unsatisfied attendees of a hypothetical performance compared to "65% of attendees did enjoy the performance" if 35% of dissatisfied attendees lies above some threshold the conveyer of the proposition deems relevant. If, following this line of reasoning, the threshold of normality in regard to dissatisfied attendees were at 45%, the latter frame would be more likely to be chosen, as the negatively valenced proportion (in this case, the proportion of dissatisfied viewers) of the total number of attendees would be within normal bounds. Sher & McKenzie (2008) term this conjecture regarding the choice of frame and its relation to thresholds of normality the "reference point hypothesis".

Following the reference point hypothesis, the statement *"This intervention will kill 200 out of the 600 people"* and the statement *"This intervention will save 400 out of the 600 people"* are not informationally equivalent, as a hypothetical interlocutor would be more probable to utter

the first of the two given the background condition of the mortality rate being exceptionally high. Participants confronted with the 'Asian disease problem' therefore correctly apply probabilistic reasoning – contrary to the traditional interpretations of their behavior – in that they consider the conditional probability of a statement being made given an informationally relevant background condition and then make their judgement of which treatment to choose based on their inference of the background condition – in this case, the background condition that a rate of 200 deaths is exceptionally high. They thus adhere to Bayesian (or rather probabilistic, as making use of the Bayesian theorem itself is not required for performing this computation[1]) standards of rational inference in their decision process. By reframing the probabilistic inference required for making a decision when choosing one of the two treatments, Sher and McKenzie (2008) thus attempt to save the optimality of the inferences participants typically draw from the traditional critique of cognitive scientists. Much like in Anderson's rational analysis, the normatively justified, optimal solution to a problem thus can serve as a model for actual human behavior.

Both of these examples – Anderson's rational analysis and Sher and McKenzie's reference point hypothesis – aim to capture processes of higher-level cognition, such as memory retrieval or explicit inference, from a Bayesian standpoint. However, (normative) Bayesian models of the mind are not limited to such high-level, explicitly inferential processes. As stated above, a wide range of mental faculties have been recast in Bayesian terms, ranging from high-level mental capabilities to low-level processes such as visual edge detection. For the latter, Geisler and colleagues (2001) have proposed a Bayesian model. "Edge-detection" refers to the identification of object contours, which has been proposed to be one of the first steps in the hierarchy of visual processes (cf. Hubel & Wiesel, 1979). Much like the previous account studying high-level cognition, Geisler et al. (2001) devise an 'optimal' (within a certain normative framework) solution for edge detection using a probabilistic Bayesian algorithm and comparing its performance to the performance of human raters. Said algorithm operates by computing object contours based on environmental statistics. The complexity of their methods goes beyond the technical notions introduced in the present paper and discussing them in detail

[1] Despite itself not making explicit use of Bayesian inference, the overall structure of Sher and McKenzie's (2008) argument so adequately fits the characterization of normative Bayesian cognitive science that it serves as an excellent example for illustrative purposes nevertheless, hence leading to its inclusion. In choosing to thereby implicitly treat the terms 'Bayesian' and 'probabilistic' as somewhat interchangeable, I follow a general trend in cognitive science, which often uses the term "Bayesian" when "probabilistic" may, strictly speaking, be more adequate (Sher and McKenzie's, 2008, paper was itself published as part of a compendium subtitled "Prospects for Bayesian cognitive science" in Chater and Oaksford, 2008). The reader is asked to forgive this slight imprecision.

will not be necessary for its argumentative aims. Rather, what the mention of this example is intended to demonstrate is that normative Bayesian approaches are not limited to "high-level" forms of inferential cognition but have also been proposed for simpler processes. The basic operating principles remain the same: cast a cognitive task (memory, explicit probabilistic reasoning, edge detection, etc.) as a problem of probabilistic inference and provide a (Bayesian) optimal solution to said problem. In the next step, demonstrate how this Bayes-optimal solution corresponds to human behavior when faced with a similar task. Whether it is Anderson's rational analysis, attempts to recast alleged examples of human irrationality in Bayesian terms to demonstrate their optimality, or devising computational visual algorithms, this is the basic pattern shared by many normative Bayesian approaches to cognition. What these approaches likewise have in common, as previously pointed out, is their lack of commitment regarding any explicit structural constraints on the respective cognitive system they are modelling. As Anderson stated in the quotation above, the goal is rather to accurately capture the probabilistic structure of the environment the cognitive system interacts with. The sole aspect of the system itself that needs to be captured by this modelling process is its behavior when dealing with its environment and its optimality in doing so with respect to the proposed Bayes-optimal computational model (cf. Jones and Love, 2011).

This shall conclude our introduction to and overview of Bayesian models of the mind in contemporary cognitive science. To what extent are these models capable of providing genuine explanations of the phenomena they capture? The remainder of the present thesis will be devoted to discussing this question. To this end, the ensuing section shall introduce and motivate the philosophical stance I shall take in doing so, namely that of (new) mechanism.

Explanation in the Sciences

The previous two sections were concerned with introducing Bayesian inference and providing an overview of its application in cognitive science to account for a plethora of phenomena discussed in this field. However, the aim of the present investigation is not merely to review the current state of cognitive science, but rather to evaluate some of its developments from a philosophical perspective, more precisely, to assess their explanatory character. To this end, it will first be required to establish an account of what it means for a scientific theory to explain a given phenomenon, i.e., an account of scientific explanation. Such an account should adequately specify what it means for a scientific theory to elucidate *why* or *how* a certain phenomenon occurs, though of course the details of what an adequate account of scientific explanation ought to achieve is itself, to a certain degree, a matter of philosophical controversy. The aim of the current section is to elucidate what kind of answers are acceptable to this type of how-and-why-questions in the context of cognitive science. Although the investigation of this matter will include a general discussion of the philosophy of scientific explanation, including a few historical remarks, its goal is not to settle definitively for an account of scientific explanation valid across all disciplinary boundaries. The reason for this limitation in scope is that it seems doubtful that any simple abstract scheme will ever be able to account for the full spectrum of scientific explanation, as scientific disciplines and the theories and models they bring forth are far too heterogenous in nature to allow for this simple unification (cf. Salmon, 1989, for a similar point). What counts as an adequate explanation in cognitive psychology may be entirely distinct from what counts as such in quantum mechanics, evolutionary biology, or sociology. I see no a priori reason for assuming that any trans-disciplinary scheme of explanation exists, nor do we need one to achieve our current philosophical objectives. Quite the contrary, the first philosophical framework to be covered in the ensuing discussion, the covering-law approach to scientific explanation, serves as an example, I believe, that an attempt to capture the entirety of scientific explanation with a simple abstract scheme can open up more problems than it solves. The notion that no single account can adequately cover the entirety of scientific explanation, however, is little more than mere intuition, and shall not be defended further in the current discussion, as none of the ensuing arguments hinge on this position being correct. Their relation to this question is entirely orthogonal, aiming merely to demonstrate that mechanism is the proper stance towards scientific explanation in the context of cognitive and neuroscience, while remaining entirely agnostic towards the question whether it is an adequate model for other, or even all scientific disciplines.

In addition to this restriction, some more preliminary remarks regarding the intended scope of the ensuing discussion are in order before the philosophical details of the covering-law approach to scientific explanation can be dissected in more detail: Necessarily, the present discussion cannot do justice to the full breadth and depth of the subject matter; much of 20th century philosophy of science has been devoted to the discussion of this topic, and the present overview can merely do so much as scratch its surface. The goal of the present discussion of the covering-law model is therefore not to provide an in-depth overview of this framework, but rather to demonstrate why it failed and how new mechanism emerged from the criticism targeted at it. Thus, the following overview should be read more as a historical motivation for a mechanistic approach to explanation, rather than a self-contained treatise in the history of philosophy. Paralleling the historical development in the field of philosophy of science, in which new mechanism emerged as the dominant approach once it became apparent that the "received view" (i.e., the covering-law model) is untenable (cf. Salmon, 1989), my aim is to initially motivate my commitment (within the previously discussed scope) to new mechanism by displaying how it avoids the failures of the covering-law approach. Avoiding the pitfalls of the covering-law model can thus serve as desideratum (as has been proposed by Salmon, 1984 or Craver, 2007) by which a satisfactory account of scientific explanation (which I claim new mechanism to be, at least for our present purposes) can be measured. Of course, avoiding the errors of an alternative model does not – in itself – validate any philosophical account, as it says nothing about the problems said account itself may be confronted with. However, it can serve as an initial motivation for giving it serious consideration. What, then, does the covering-law approach to scientific explanation state?

The Covering-law approach

Under "covering-law" approach, I shall – following convention (cf. Salmon, 1984, 1989) - understand any model of scientific explanation which proposes that to explain a phenomenon is to infer it from a general law using auxiliary premises. The most prominent exemplar from this range of philosophical models is undoubtedly Hempel and Oppenheim's (1948) deductive-nomological model (henceforth D-N model). This account maintains that scientific explanations have the form of logical arguments, with the phenomenon to be explained, the explanandum, being formally derivable from a set of premises which serve as the explanans. This is the deductive aspect of the model. The nomological aspect is that among the set of premises from which the explanandum is derived, there must be at least one general law, so that the explanandum can be derived from said law in conjunction with other premises about the state of the world (Hempel & Oppenheim, 1948). In his excellent historical overview of the

developments in the philosophy of scientific explanation in the decades following the publication of the by-now classic Hempel and Oppenheim paper of 1948, Salmon (1989, p. 8) summarizes the D-N account as follows:

> According to […] [the D-N] account, a D-N explanation of a particular event is a valid deductive argument whose conclusion states that the event to be explained did occur. This conclusion is known as the explanandum-statement. Its premises- known collectively as the explanans - must include a statement of at least one general law that is essential to the validity of the argument - that is, if that premise were deleted and no other change were made in the argument, it would no longer be valid. The explanation is said to subsume the fact to be explained under these laws; hence, it is often called "the covering-law model." An argument fulfilling the foregoing conditions qualifies as a potential explanation. If, in addition, the statements constituting the explanans are true, the argument qualifies as a true explanation or simply an explanation (of the D-N type).

Quite soon after the publication of the original Hempel and Oppenheim (1948) paper, it became clear that a purely deductive approach is unable, even as a regulatory ideal, to account for (the entirety of) actual scientific practice due to the ubiquity of statistical approaches in all scientific fields, most prominently of course in quantum mechanics (the notion of statistical explanation was already brought up in the original Hempel and Oppenheim (1948) paper, but not treated in any detail there). This shortcoming led to the proposal of an additional model covering statistical explanation, namely the inductive-statistical I-S model (Hempel, 1962). The I-S model strongly resembles the D-N model, with the key difference being that the explanans phenomena is not deductively derived from a general law, but rather inductively derived from a statistical law, i.e., an explanation under the I-S model aims to demonstrate that a phenomenon is – statistically – to be expected given the covering statistical law. Following Salmon (1989, p. 8f.),

> the inductive- statistical model (l-S), explains particular occurrences by subsuming them under statistical laws, much as D-N explanations subsume particular events under universal laws. There is, however, a crucial difference: D-N explanations subsume the events to be explained deductively, while I-S explanations subsume them inductively. An explanation of either kind can be described as an argument to the effect that the event to be explained was to be expected by virtue of certain explanatory facts. In a D-N explanation, the event to be explained is deductively certain, given the explanatory facts (including the laws); in an I-S

explanation the event to be explained has high inductive probability relative to the explanatory facts (including the laws).

Thus, both the D-N model and the I-S model identify scientific explanation with inferring the explanandum from a series of premises containing at least one general law (either deductively or inductively in the case of the statistical laws). Salmon (1989) summarizes the central philosophical commitments of the covering-law approach (both in the deductive and in the inductive flavor), which are the following:

i. All scientific explanations have the form of arguments, deriving a conclusion from a set of premises (referred to by Salmon (1977) in an earlier publication as the "third dogma of empiricism")
ii. All scientific explanations must allude to at least one general law
iii. Scientific explanation and scientific prediction are essentially the same and differ – if at all – by pragmatic considerations. Both share the same logical form under the covering-law approach (referred to by Salmon (1989, p. 24) as the "explanation/prediction symmetry thesis")
iv. Causal explanation is – under the covering-law model – only valid insofar as it follows the scheme of deriving the explanans from at least one general law. There is no special explanatory role that causality has to play beyond this (as explicitly pointed out by Hempel, 1965).

(Salmon discusses more commitments of the covering-law approach in his essay, but for our current purposes, the selection above is sufficient).

In the decades following the initial proposals of the D-N and the I-S models, all four of these assumptions have come under severe scrutiny, casting serious doubts on the tenability of the covering-law approach. As the problems of the covering-law model are well known and have been much discussed in academic philosophy, I will not treat them in too much depth in the present thesis, but rather give only a cursory glance over the most blatant problems of this stance.

First, there exist several lawful generalizations which enable successful predictions but which – intuitively – provide nothing in terms of explanation of a phenomenon. Consider the famous example of the barometer (Reichenbach, 1956). If the barometer falls, one can predict that the weather will change. There seems to be a lawlike relationship holding between the falling of the barometer and the change in weather, yet hardly anyone will be willing to accept that the falling of the barometer serves as an explanation for the weather change.

Another commonly cited argument against the D-N model is Bromberger's (1965) flagpole example. The example goes as follows: Given knowledge about the illumination angle, one can predict the length of a flagpole's shadow from the length of the flagpole itself (in conjunction which general laws of geometrical optics and trigonometry). Following the covering-law approach's symmetry thesis, the flagpole's length thus also serves as an explanation of its shadow's length. This case of explanation/prediction symmetry seems intuitively unproblematic; few would object that the length of the flagpole in conjunction with the laws of geometrical optics and trigonometry as well as knowledge of the illumination angle can adequately explain the length of the shadow. However, the inverse prediction also holds. Having knowledge of the illumination angle, the relevant general laws, and the length of the shadow, the length of the flagpole can be predicted. In this case, one would intuitively object that no explanation has been given. The flagpole's length is causally efficacious in bringing about the length of the shadow, but the inverse does not hold. It seems therefore that the former can explain the latter, but not vice versa. Such causal concerns, however, are irrelevant under the covering-law approach, following assumption iv.

In the flagpole example, the explanandum is derived deductively. However, a further counterexample has been presented by Salmon (1989) specifically for the case of statistical laws:

> Syphilis and paresis. Paresis is one form of tertiary syphilis, and it can occur only in individuals who go through the primary, secondary, and latent stages of the disease without treatment with penicillin. If a subject falls victim to paresis, the explanation is that it was due to latent untreated syphilis. However, only a relatively small percentage-about 25%-of victims of latent untreated syphilis develop paresis. Hence, if a person has latent untreated syphilis, the correct prediction is that he or she will not develop paresis. (Salmon, 1989, p. 49)

Although the covering-law approach takes prediction and explanation to be symmetrical, the correct prediction made on the basis of the relevant statistical laws differs from the phenomenon to be explained in this example. The prediction that a person will not develop paresis is of no explanatory value when it comes to explaining why a person *has*, as a matter of fact, developed this condition.

Examples such as these cast serious doubt on assumption iii, the symmetry thesis, i.e., the notion that there is no fundamental difference between scientific explanation and prediction. Not only assumptions iii and iv, however, have undergone severe criticism. Consider assumption i., the

notion that all scientific explanations have the form of arguments. As Craver (2007, p. 37) points out following Salmon (1984),

> [t]he strength of an argument is not diminished one bit by the addition of any number of irrelevant premises. The strength of an explanation, on the other hand, depends crucially upon whether the factors described in the explanatory text are relevant.

This idea is illuminated by Kyburg's (1965) example of salt upon which a spell was cast, which is also taken up and discussed by Salmon (1989) and Craver (2007):

> This sample of table salt dissolves in water, for it has had a dissolving spell cast on it, and all samples of table salt that have had dissolving spells cast on them dissolve in water (Kyburg, 1965, p. 147)

Undoubtedly, the generalization that salt upon which such a spell is cast will dissolve in water will hold, and this generalization can serve as a premise in an argument. Such an argument, insofar as its mode of inference holds, is completely valid, independent of any number of additional, irrelevant additions to its premises. The same, however, cannot be said of scientific explanations. A scientific explanation alluding to a spell cast on salt to account for its dissolving in water will, rightly, be rejected. Much like in the flagpole example, the reason is that the salt being hexed has no causal influence at all on its dissolving in water.

Beyond the problems they generate for assumption i., i.e., that explanations are formal arguments, two of the counter-examples discussed so far, the flagpole and the hexed salt, furthermore cast serious doubt on assumption iv., namely, that causality has no special role to play in scientific explanation, for the simple reason that both examples seem to fail as explanations due to the manner in which explanans and explanandum are causally connected (or rather, fail to be causally connected in the right manner). Salmon (1989, p. 59) argues this point as follows:

> What is crucial for […] explanation […] is not how probable the explanans renders the explanandum, but rather, whether the facts cited in the explanans make a difference to the probability of the explanandum.

In addition to the previous criticisms, which were directed against the suitability of the covering-law model independent of any concrete scientific discipline, one can argue that this approach fails to account for explanation in a wide variety of special sciences in particular, as they, contrary to fundamental physics, often do not involve the formulation of general laws. Scriven (1959) has made this point in regard to evolutionary biology: Although evolutionary

biology is capable of providing a wide variety of explanations, it does not invoke any universal laws; rather, it describes a unitary, idiosyncratic process, namely evolution by natural selection. Similar points could be made for other special sciences. It is doubtful, for example, that psychology uncovered more than a handful of (usually rather simple) general laws, if any, and this small selection is certainly incapable of accounting for all psychological explanation (cf. Teigen, 2002). Many of the special sciences seem to deal with localized, idiosyncratic phenomena, which do not easily lend themselves to lawlike generalizations. Nonetheless, we intuitively see no problem is ascribing to these disciplines genuine explanatory power. Considerations like these cast doubt on the last of the remaining core assumptions of the covering-law approach, namely assumption ii., which, to reiterate, states that all scientific explanations must involve at least one general law.

Although attempts were made to salvage the covering-law model by modifying it in such a manner as to be able to deal with some of the problems listed above, e.g. by listing further constraints (for example about the nature of scientific laws) or dropping some of the assumptions made in the original versions of the D-N and I-S models (Hempel, 1968, Hempel, 2012 [1976]), the extent of the problems above, in addition with the emergence of alternative models of scientific explanation, gradually led to the downfall of the covering-law model, rendering it, at this point, little more than a historical artefact of logical empiricism.

Given that the D-N and I-S models described above have few – if any – contemporary proponents, what is the takeaway from discussing them? The failure of the covering-law approach is illuminating in showcasing the pitfalls a satisfactory account of scientific explanation ought to avoid. Motivated by the problems this approach has, Craver (2007) proposes, based on considerations by Salmon (1984), five desiderata a satisfactory philosophical theory of explanation should fulfill (and which the covering-law approach fails to satisfy). Each desideratum reflects the rejection of one pattern of explanation one would intuitively be inclined not to accept, but which constitutes a valid explanatory argument (if properly constructed) under either the I-S or the D-N model (as has been attempted to show using the foregoing examples). These desiderata are the following:

Desideratum 1: "*Mere temporal sequences are not explanatory* [emph. added]" (Craver, 2007, p. 26). A weather change follows the barometer falling in a lawful manner, but the former does not explain the latter.

Desideratum 2: "Causes explain effects and not vice versa (explanatory asymmetry) [emph. added]" (Craver, 2007, p. 26). The flagpole's length explains the length of its shadow, but not the other way around.

Desideratum 3: "*Causally independent effects of common causes do not explain one another* [emph. added]" (Craver, 2007, p. 26). Both the barometer falling and the weather changing are explained by a change in atmospheric pressure, but they do not explain each other.

Desideratum 4: "*Causally irrelevant phenomena are not explanatory* [emph. added]" (Craver, 2007, p. 26). The hexed salt is not explanatory of the salt's solubility in water.

Desideratum 5: "*Causes need not make effects probable to explain them* [emph. added]" (Craver, 2007, p. 26). The syphilis is explanatory of the paresis, even though the latter remains improbable given the former.

What Craver attempts to demonstrate by these desiderata is that adequately tracking the causal chain bringing forth an explanans phenomenon can avoid the pitfalls that the covering-law model succumbs to. He thus uses these desiderata to argue for a *mechanistic* account of scientific explanation. Together with the *unificationist* account (Friedman, 1974; Kitcher, 1989), this mechanistic approach constitutes the most influential alternative to the covering-law model (Craver, 2007). Before delving into the mechanistic account of scientific explanation, let us explore its most prominent contemporary alternative, the unificationist account of explanation. Although it provides some genuine insights into how science progresses (and ought to progress), this model, as I will attempt to show, cannot completely avoid some of the pitfalls that proved problematic for the covering-law approach. It fails to fulfill Craver's list of desiderata and commits itself to similar implausible cases of pseudo-explanation as the covering-law model. This shall be the topic of the ensuing subsection.

The U-Model

The U-model states that the defining characteristic of scientific explanation is that it unifies a range of different phenomena under a single theoretical framework (Friedman, 1974; Kitcher, 1989; Woodward & Ross, 2021). Consider the case of the Newtonian synthesis following the introduction of Newtonian mechanics (cf. Woodward & Ross, 2021), i.e., the unification of the hitherto deemed separate phenomena of celestial motion and earthly motion under a single unifying framework (JRank Science Encyclopedia, n.d.). The notion that the Newtonian explanation is superior to the peripatetic explanation, deeming celestial and earthly mechanics distinct domains, because it is able to account for a wider variety of phenomena under a single

framework is prima facie quite appealing. The mere idea that explanation in science hinges on unification is itself of course not yet a genuine philosophical account of scientific explanation. Kitcher (1989) specifies his account in more detail by arguing that an explanation is better if it allows to derive more phenomena from fewer *argument patterns*. In his own terms:

> Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same pattern of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts we have to accept as ultimate. (Kitcher, 1989, p. 432)

Let us elucidate Kitcher's notion of unification and the role argument patterns play in it in more detail (the following summary follows the portrayal by Woodward and Ross, 2021). In the U-model, a *schematic sentence* is a sentence in which some terms have been replaced by variables. A sequence of such sentences is referred to as a *schematic argument*. In such a schematic argument, the *classification* assigns argumentative roles (premise, conclusion) to the individual schematic sentences and specifies the used inferential pattern. *Filling instructions* specify how these variables can be filled, e.g. by stating that a given variable can be replaced by reference to a discrete mass, a carcinogenic substance, a cognitive bias, etc. A triple of a schematic argument, filling instructions for each of the schematic sentences in said argument, and a classification specifying their role and argumentative pattern is referred to by Kitcher as an *argument pattern.* As summarized by Barnes (1992),

> […] an argument pattern consists of a sequence of schematic sentences which can be instantiated so as to generate a derivation of some explanandum from some scientific theory. (Barnes, 1992, p. 559)

As stated before, an explanation is good – following the U-model – if it is able to derive various different phenomena from as few argument patterns as possible, with the restriction that a schematic argument ought to be as *stringent* as possible. This means that it ought to pose restrictions on the sentences that fill it, in order to avoid non-explanatory arguments of the type "X is the case because god wills it that X is the case" or similar, which pose no restrictions whatsoever on the instantiating schematic sentences. Explanations which fulfill the above criteria and are sufficiently restrictive jointly constitute what Kitcher calls an "explanatory story […] which collectively provide[s] the best systematization of our beliefs" (Kitcher, 1989, p. 430).

Although the above is but a brief summary of the core tenets of the U-model, it becomes immediately apparent that the U-model resembles the DN-model in one key aspect, namely in its proposal that scientific explanations take the form of formal arguments. Kitcher (1989, p. 448) himself explicitly commits to this position when he states that *"in a certain sense, all explanation is deductive"*. His account is explicitly anti-mechanistic, rejecting the notion of alluding to causal chains in formulating scientific explanations. One of the strengths of the U-model, he states, is its ability to derive notions of causation from notions of unification and explanation (the reversal of the derivative relation mechanism proposes), thus not depending on causal notions to account for explanation. In his own words,

> the 'because' of causation is always derivative from the 'because' of explanation. In learning to talk about causes or counterfactuals we are absorbing earlier generations' views of the structure of nature, where those views arise from their attempts to achieve a unified account of the phenomena. (Kitcher, 1989, p. 477)

The unificationist account not only shares some of the assumptions of the covering-law model but also (as hinted at before) some of its shortcomings. What are said shortcomings? As Craver (2007) argues, the U-model, like the covering-law model, fails to fulfill the desiderata for an account of scientific explanation listed above. Specifically, desideratum ii., the explanatory asymmetry of cause and effect, is not met by the U-model. To reiterate, this desideratum states that while a cause can be invoked in explaining its effects, an effect provides no explanation for its causes. This point is first argued by Barnes (1992), who states

> that Kitcher's unificationist theory of explanation is flawed beyond repair insofar as this theory is intrinsically unable to respect a well-known feature of empirical explanation: its asymmetric structure. (Barnes, 1992, p. 558)

Why is this the case? Consider once more the afore discussed case of the flagpole and its shadow. Knowing the height of a flagpole and the angle from which light is falling on it, one can deduce the length of its shadow. Likewise, knowing the length of the shadow and the angle of illumination, one can infer the height of the flagpole. However, whereas the height of the flagpole explains the length of the shadow, the inverse does not hold. This example has proven problematic for the covering-law model. Is the same the case for the U-model? Kitcher (1989) argues that it is not. His reasoning is that the shadow explaining the length of the flagpole is an *unprojectable* explanation pattern because not all objects cast shadows from which their length could be inferred, whereas for every shadow there is an object which can – in conjunction with

the source of illumination – explain its length. Hence, explaining the length of the flagpole from the length of its shadow is not a proper case of unification.

Barnes (1992) aims to counter this objection by attempting to provide examples of asymmetric explanations that are not unprojectable and whose asymmetry cannot be accounted for by the U-model. For the present discussion, let us focus on only one of his various examples, namely that of a temporally symmetric closed system, i.e., a physical system which does not exchange energy with its surroundings and is governed by laws of nature working the same way independent of whether time is moving forward or backward. The latter is, for example, the case for Newtonian mechanics. In such a closed system, following Newtonian laws, one could derive predictions of an object's position in space and velocity at any given point in time t_1 based on a complete description of the system at a prior point in time t_0. So far, this seems unproblematic. However, one could likewise retrodict the object's position and velocity from a complete description of the system at a later point in time t_2. For inferring position and velocity at t_1 from a complete description of the closed system at t_n, it does not matter whether t_n comes prior to or after t_1 in the temporal sequence of the system's states. Newton's laws can be employed in both cases, either to predict future states, or to retrodict previous states, with no difference in their application (Barnes, 1992). This circumstance, however, is at odds with the asymmetry desideratum. Barnes states the problem as follows:

> The Newtonian Predictive Pattern is intuitively explanatory […]. A perfectly legitimate explanation of the fact that some object has a particular position and velocity at some time would consist of a citation of the fact that the system containing the object had a particular state at some earlier time, together with the deterministic Newtonian laws that led inexorably from the latter to the former. However, the Newtonian Retrodictive Pattern is utterly nonexplanatory […] [E]xplananda cannot in general be explained by citing facts that occurred subsequent to the explananda themselves. These two patterns, it seems to me, are completely identical with respect to their ability to supply a unified account of the class of explananda [in question]*:* Every prediction […] fitting the Predictive Pattern will be matched by a corresponding retrodiction of the same explanandum fitting the Retrodictive Pattern. (Barnes, 1992, p. 565)

In short, the state of the system at t_0 causes the state of the system t_1 and can thus be invoked in an explanation of the latter. The state of the system at t_1, however, cannot be explained by the state of the system at t_2, as the latter is a causal effect of the former. The unificationist model (much like the covering-law model) is not capable, however, of accounting for this asymmetry.

As stated before, Barnes (1992) provides further examples in support of the claim that the U-model, much like the DN-model, is unable to account for the asymmetric nature of explanation. However, as this point has already been established, I shall forego reconstructing these additional examples.

What, then, are the implications of this failure to account for explanatory asymmetry? The implication is not that scientific unification is not a valuable goal to aim for in scientific theorizing, nor that unification and explanation do not often go hand in hand. It does, however, suggest that – although perhaps correlated – explanation and unification are conceptually distinct, in that the former cannot be reduced to the latter. Halonen and Hintikka (1999, p. 28) argue this point in their aptly titled paper "Unification: It's magnificent but is it explanation?", where they propose that "[u]nification is an important desideratum in theory formation, but it does not play any role in the actual process of explanation in the sense of explaining a given explanandum".

Despite its shortcomings, the unificationist approach points out a central and relevant aim of scientific theory construction. The problems listed above are thus perhaps best read not as an outright rebuttal of this framework, but rather as a demonstration of its incompleteness: to provide explanation, it does not suffice to provide unification. A theory being incomplete, however, is distinct from it having nothing valuable or insightful to add to the discussion. Salmon (1989, p. 183), himself a proponent of the mechanistic account, hints at this idea when he states the following:

> Whereas the unification approach is "top-down," the causal/mechanical is "bottom-up." When we pause to consider and compare these two ways of looking at scientific explanation, an astonishing point emerges. These two ways of regarding explanation are *not incompatible* with one another; each one offers a reasonable way of construing explanations. Indeed, they may be taken as representing two different, but compatible, aspects of scientific explanation. Scientific understanding is, after all, a complex matter; there is every reason to suppose that it has various different facets.

Consider, in this context, the example which Friedman (1974) chooses to open his essay defending a version of the unificationist account of explanation, namely the question why water turns to steam upon being heated. He answers this question as follows:

> Water is made of tiny molecules in a state of constant motion. Between these molecules are intermolecular forces, which, at normal temperatures, are sufficient to hold them together. If the water is heated, however, the energy, and consequently the motion, of the molecules

> increases. If the water is heated sufficiently the molecules acquire enough energy to overcome the intermolecular forces - they fly apart and escape into the atmosphere. Thus, the water gives off steam. This account answers our question. (Friedman, 1974, p. 5)

In his paper, Friedman considers different, competing accounts of scientific explanation in order to evaluate which can best capture the type of explanation just described. As previously stated, he arrives at the conclusion that the unificationist approach is best suited to this end. It is worth noting, however, that what Friedman describes in the passage cited above is not a story about reducing the number of independent phenomena in the world, at least not in itself and without further elaboration. Rather, it is a story about physical entities and how they interact in time and space to give rise to the explanandum phenomenon. In short, it is a causal/mechanistic explanation. Mechanism itself is not among the theories Friedman considers in his paper, but it is telling that the example he chooses as a paradigmatic example of a good scientific explanation has this form. It goes to show, as previously established, that unificationism is not a thesis in strict opposition to mechanism. It is a theory about how explanations ought to fit into the wider nexus of established scientific knowledge. It is a global account of explanation, rather than a local one. However, these purely global aspects of explanation, as has been attempted to show in the foregoing paragraphs, are by themselves insufficient to account for scientific explanation without falling prey to some of the same problems plaguing the covering-law model.

This assessment concludes our evaluation of the DM-model and the U-model. I have attempted to demonstrate that neither, by themselves, offer a satisfying account of scientific explanation. While rebutting the most prominent alternatives does not necessitate accepting the mechanistic stance, the preceding sections may provide sufficient motivation for giving it serious consideration. The ensuing section shall, after a brief general introduction to new mechanism, introduce Craver's (2007) mechanistic account of explanation in the neurosciences, as well as the reasons for adopting it. It will be further attempted to demonstrate that this account can be – contra prima facie objections – extended to purely cognitive theories, the type of theory typically provided in cognitive science.

New Mechanism

Much like the unificationist account, the new mechanistic conception of scientific explanation has been developed in the final decades of the 20th century as an alternative to the hitherto dominant covering-law model. As such, it was intended to avoid the pitfalls of the latter and provide an account of scientific explanation which does not necessitate the allusion to general laws (Ross & Woodward, 2023). The "new" in "new mechanism" serves to distinguish this approach from the old mechanism of Descartes and his disciples, which commits to a metaphysical view of a clockwork-like universe which proponents of new mechanism usually do not wish to subscribe to (cf. Craver & Tabery, 2023). As the Cartesian approach will not be further discussed, I shall use the terms "mechanism" and "new mechanism" (as well as the respective derivative adjectives) interchangeably.

The core idea of new mechanism is that scientific explanations work not by inferring an explanandum from general laws (either deductively or inductively), but rather by explicating the underlying causal mechanism giving rise to it. As noted earlier when discussing the barometer example, the sharply falling barometric reading is a satisfactory basis for predicting a storm but contributes in no way to its explanation. From a mechanistic point of view, the failure to explain the weather change in this example stems from a lack of a direct causal connection between the barometer and the change in weather. Rather than the former being part of the causal chain leading to the latter, both stem from the same cause, a change in atmospheric pressure, while they themselves are not causally connected. The lawful connection between both events is thus – so the mechanist – not sufficient to achieve explanation.

Although still working under the „received view" of the covering-law model, Railton (1978, p. 208) succinctly captures the core idea underlying this conception when he writes that

> [t]he goal of understanding the world is a theoretical goal, and if the world is a machine - a vast arrangement of nomic connections - then our theory ought to give us some insight into the structure and workings of the mechanism, above and beyond the capability of predicting and controlling its outcomes […] Knowing enough to subsume an event under the right kind of laws is not, therefore, tantamount to knowing the *how* and *why* of it. As the explanatory inadequacies of successful practical disciplines remind us: explanation must be more than potentially-predictive inferences or law-invoking recipes.

Railton's (1978) proposal to incorporate causal-mechanistic considerations into scientific explanations involves adding additional causal premises to a D-N style nomic explanation. His

proposed solution to the problematic barometer example is to add a premise specifying the connection between the barometer falling and the drop in atmospheric pressure. Thus, his explanation has the form of a deductive argument (though his view is also open to inductive arguments), which specifies the causal mechanism giving rise to the explanans. Additionally, Railton gives up the notion that any scientific explanation must appeal to at least one general law, stating that it is "difficult to dispute the claim that many proffered explanations succeed in doing some genuine explaining without either using laws explicitly or (somehow) tacitly asserting their existence." (Railton, 1981, p. 249). Railton's account makes for an interesting case study, as it seems to occupy an intermediary position between the received view of the covering-law model and the emerging mechanistic account, and thus illustrates how the problems of the former, historically, gave rise to the latter. Despite perhaps not being adequately characterized as a fully fletched New Mechanist yet, his theses provide a strong motivation and point of departure for the latter. His argument is one taking into consideration actual scientific practice and aiming to philosophically account for it:

> If one inspects the best-developed explanations in physics or chemistry textbooks and monographs, one will observe that these accounts typically include not only derivations of lower-level laws and generalizations from higher-level theory and facts, but also attempts to *elucidate the mechanisms* at work. […] It seems to me implausible to follow the old empiricist line and treat all these remarks on mechanisms, models, and so on as mere *marginalia,* incidental to the "real explanation", the law-based inference to the explanandum. […] [I]t seems clear enough that an account of scientific explanation seeking fidelity to scientific explanatory practice should recognize that part of scientific ideals of explanation and understanding is a description of the mechanisms at work, where this includes, but is not merely, an invocation of the relevant laws (Railton, 1981, p. 242).

The mechanistic thesis has undergone numerous amendments and refinements since the days of its early conception. In a more recent publication, Bechtel and Abrahamsen (2005, p. 423), giving up allusions to general laws or formal inferability, define 'mechanism' as "a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena". As pointed out by Craver & Tabery (2023), this definition hinges on four key concepts, namely an explanandum (i.e., the phenomenon to be explained), constitutive parts, causal connection, and (structural) organization, four components shared – so Craver and Tabery (2023) – with other influential accounts of (explanatory) mechanisms (cf. Machamer, Darden, & Craver, 2000; Glennan, 2002). Details of what mechanisms are, however, and how

they relate to target phenomena, as well as how they achieve explanation all may vary between individual accounts of mechanistic explanation. Instead of discussing each of these conceptions in detail, in the ensuing section, I shall focus on specifying one account of mechanistic explanation in detail, namely that of Craver (2006, 2007). This account, proposed by Craver originally as an account of mechanistic explanation in the neurosciences, is capable, so I will argue, of providing a singular evaluative framework for assessing theories within both the cognitive- and neurosciences, as well as accounting for actual scientific practice within these fields – something that the received view of the covering-law model, as argued in the previous sections, was incapable of.

Craver on mechanism

Craver (2006, 2007) distinguishes two forms of mechanistic explanation. The first specifies the causal history leading up to an event, in which case the relevant causal chain bringing forth the phenomenon in question is the explanandum. He refers to this type of mechanistic explanation as *etiological* explanation. This category is contrasted with what Craver (2007) terms *constitutive explanation*, in which the explanans consists in the specification of the organization and interaction of constitutive components bringing forth the explanandum phenomenon. Thus, to explain a given phenomenon X constitutively, one needs to show how the parts constituting X are organized and interacting so that by their joint interaction the target phenomenon comes to be. "Constituents" in this context is to be understood simply as mereological parts of the target system. Although not all mereological parts of the system are necessarily mechanistically relevant components, all components are – following Craver – mereological parts of the system. Thus, the term "constituent" shall be reserved for those mereological parts of a system which are explanatorily relevant for the aspect of the system to be explained. Whether or not a mereological part of a system ought to be considered a constituent thus hinges on what capacity of the system is the explanandum in a given explanation and can differ depending on what aspect of the system is to be explained. Therefore, which entities serve as components in an explanation is restricted by the causal and mereological structure of the world as well as the specific phenomenon of interest. A mechanistic explanation of visual perception will not involve the metatarsal bones as a constitutive component simply because they lack the relevant constitutive relationship with the target phenomenon. Besides these ontic considerations, however, pragmatic considerations also play a role in deciding which components to include in a given scientific explanation. This decision is partially a matter of the context in which the explanation is given, and to whom. As such, there is no single, definite mechanistic explanation of a phenomenon, there are a great many. Any system has either an infinite number of

mereological parts or a umber so vast it borders on infinity for all practical purposes (depending on one's metaphysical views on mereological questions), and therefore an either vast or infinite number of accounts – varying greatly in granularity and detail – could be given of how these parts give – through their organization and interaction – rise to the explanandum phenomenon. Some of these will be too coarse to provide much insight into the network of component interactions giving rise to a phenomenon, others will be too fine grained to be tractable. A mechanistic explanation of mathematical reasoning in humans invoking subatomic particles and their interaction may, conceptually, be possible, but as a matter of intellectual and computational limitations not feasible in practice. We therefore require additional specifications regarding the level of detail a mechanistic account should provide.

While this issue resists simplistic or formulaic solutions, Craver (2007) provides some further specifications regarding this matter. He refers to (hypothetical) explanations invoking every constituent of an explanans system that is relevant in bringing forth its behavior of interest as *ideally complete descriptions of a mechanism*. Based on the manner in which his account conceptualizes constituents, it seems implausible that only one such ideally complete description exists for a given mechanism. Constituent components are defined as mereological parts, and – as already touched upon – any given system has either an infinite number of mereological parts or as many parts as it has fundamental atomic parts and mereological sums thereof. Therefore, it would seem arbitrary to declare any one of the myriads of possible mechanism descriptions invoking mereological parts the sole ideally complete one, especially considering that Craver (2007) explicitly rejects the notion that ideally complete mechanism descriptions must allude to fundamental parts. It would therefore seem more sensible to consider the set of ideally complete mechanism descriptions a subset of the set of all mechanism descriptions, with the defining feature of this subset being that its members take every system-constituent relevant in bringing forth the explanandum behavior into consideration. This is not a consideration Craver himself gives, but it is a natural consequence of the stance he takes. This implication is neither problematic nor especially surprising, given the nature of mereological parthood. It is not problematic because these different ideal mechanism descriptions would merely subdivide the same spatio-temporal process in different ways. The underling causal structure of the world remains unchanged by this multitude of possible mechanism descriptions, and choosing an adequate one is once more a matter of pragmatic considerations rather than anything to do with the phenomenon itself.

As stated before, ideally complete descriptions of a mechanism are nothing more than a hypothetical gold standard of mechanistic explanation. No actual scientific explanation will

invoke every causally relevant constitutive component due to the intractable complexity of such a model and the practical impossibility of devising it (Borges' (1975 [1935]) famous short story "On exactitude in science" comes to mind, which describes a kingdom becoming so obsessed with cartography that its king orders a map with scale one-to-one to be created, which then covers the entire land). Real mechanistic explanations do not (and cannot) display this level of granularity. Often, they are little more than rough outlines of the relevant mechanism underlying a phenomenon in which the exact details of the causal story are never specified. We would nonetheless ascribe at least some amount of explanatory power to such models. To account for this circumstance, Craver (2006, 2007) introduces the notions of *mechanism sketches* and *mechanism schemas.* He defines the former as follows:

> A mechanism sketch is an incomplete model of a mechanism. It characterizes some parts, activities, or features of the mechanism's organization, but it has gaps. Sometimes gaps are marked in visual diagrams by black boxes or question marks. More problematically, sometimes they are masked by filler terms […] [that] indicate a kind of activity in a mechanism without providing any detail about exactly what activity fills that role. Black boxes, question marks, and acknowledged filler terms are innocuous when they stand as place-holders for future work […] In contrast, filler terms are barriers to progress when they veil failures of understanding. If the term "encode" is used to stand for "some- process-we-know-not-what," and if the provisional status of that term is forgotten, then one has only an illusion of understanding (Craver, 2006, p.5).

Those mechanistic models falling in between mere sketches and complete mechanism descriptions on the spectrum of completeness are what Craver calls *model schemata*:

> Between sketches and complete descriptions lies a continuum of mechanism schemata that abstract away to a greater or lesser extent from the gory details (as Philip Kitcher calls them) of any particular mechanism. What counts as an appropriate degree of abstraction depends, again, upon the uses to which the model is to be put. (Craver, 2006, p. 6)

Craver further expands this taxonomy of mechanistic models by adding phenomenological models to it. They sit at opposite of ideally complete model descriptions on the completeness spectrum in that they do not specify any details of the underlying mechanism, but merely serve as an input-output function, describing which input into a system produces which output, while remaining completely agnostic as to how this transformation occurs. Such models

> […] sit at the farmost limit of sketchiness. They are complete black boxes; they reveal nothing about the underlying mechanisms and so merely "save the phenomenon" to be explained. (Craver, 2006, p. 6)

In addition to the specified degree of completeness and abstraction, models can vary – so Craver – in mechanistic plausibility. By that he means that even highly detailed and specific models can fail to be explanatory (beyond merely "saving the phenomenon" as a phenomenological model) if the entities and their interactions postulated in the model are very unlikely to correspond to the actual entities and interactions bringing about the explanandum phenomenon in the real world. Consider the example of using an artificial intelligence system programmed to play chess as a model for human chess playing abilities. (Some of) these artificial intelligence systems operate by calculating a vast number of possible outcomes of different moves and picking the most promising one on that basis. Such models are – despite being specified in very fine detail by their underlying algorithms – not a mechanistically plausible model of human chess performance due to the computational limitations and processing speed of the human brain. In other words, even though a series of chess moves may be brought about both by a human chess player and a chess-AI, which means that the latter is capable of emulating the performance of the former (in some instances at least) and is thus phenomenon-saving when viewed as a mechanistic model, it does not serve as a plausible mechanistic model, because the way the human plays is mechanistically entirely distinct from the way the AI plays. The algorithm of the AI does not correspond in any meaningful way to the causal-mechanistic process underling the human's chess playing abilities. Craver (2007) refers to such models as "how-possibly" models. They specify a mechanism that is *potentially capable* of bringing forth a phenomenon, but not the mechanism *actually underlying* said phenomenon in the real world.

So far, we have, in passing, touched upon the notion of explanatory value a few times. It has already been established that explanatory value is, under Craver's framework, not a binary property that a given model either possesses or lacks. Rather, the Craver treats explanatory value as a dimensional property, which can exist in degrees. As such, two models can have explanatory value for explaining the same phenomenon, but one can achieve this to a higher degree than the other. Following this line of reasoning, explanatory sketches have more explanatory value than phenomenological models, as the former provide some restrictions on the space of possible mechanisms and thus begin to fill-in the picture of the components and interactions underlying the explanandum phenomenon, whereas phenomenological models merely capture the input-output mappings of their target phenomenon, leaving the space of

possible mechanisms accounting for this transformation completely unrestricted. Model schemas, in turn, have more explanatory value than mere sketches, as they provide a richer picture about how a target phenomenon is brought about by its constituents and their interactions. Ideally complete mechanism descriptions provide the maximum explanatory value, as they, per definition, capture every mechanistically relevant aspect of a phenomenon.

Given this conceptualization of explanatory value, one could ask whether phenomenological models provide any explanatory value at all, as they fail to establish any (sketch of a) mechanistic intermediary between inputs and outputs of an explanandum system. Although Craver does not explicitly address this issue, I shall argue that they do, i.e., that purely phenomenological model do have some (albeit strongly limited) explanatory value. The reason why this is the case will become apparent once we consider Craver's criteria for assessing explanatory value in more detail. So far, it was merely established that providing more causally relevant aspects of a mechanism in a model adds to its explanatory value (although at some point the costs of additional complexity in the model must surely outweigh the benefits of an increase in explanatory power, at which point, once more, pragmatic considerations come into play to find the proper balance). Craver (2006, 2007) adds to this notion by providing a range of criteria which allow to assess a given model in terms of its explanatory power in more detail. These criteria are as follows:

1. A mechanistic explanation must specify the phenomenon it aims to explain.

An explanation is a conjunct of an explanandum, i.e., the phenomenon to be explained, and an explanans, which accounts for, i.e., explains the explanandum. Although theories of scientific explanation usually focus on the explanans (as became apparent in the previous section), following Craver's (2006, 2007) mechanistic account, a proper scientific explanation ought to begin by exactly specifying the explanandum, i.e., the phenomenon in question. The phenomenon – prior to its explanation – can be thought of as a black box connecting inputs to outputs. The aim of a scientific explanation – following the mechanistic viewpoint – is situating the phenomenon in the wider context of causal forces operating in the world, i.e., opening the black-box and specifying (some of the) entities and interactions that bring forth these input-output mappings. In order to do so, one must first describe the phenomenon with sufficient precision. Only by doing so can the mechanism bringing about the input-output mapping in question be properly understood and specified. Craver (2007) names five kinds of conditions one ought to specify in order to adequately capture a phenomenon before one can aim to explain

it. These should not be considered a list of necessary criteria, but rather as exemplars of the type of specification that ought to be undertaken in order to adequately situate a phenomenon in the causal structure of the world. What aspects of and in what detail a phenomenon ought to be described depends on the particular case one aims to construe an explanation of. Given the heterogeneity of phenomena different sciences deal with, it seems unlikely that devising a list of necessary and sufficient criteria for adequately specifying a phenomenon across disciplines is feasible. Once again, pragmatic considerations are in order in this context. The exemplars of the kinds of conditions which ought to be specified according to Craver (2007) are:

1. precipitating conditions
2. inhibiting conditions
3. modulating conditions
4. non-standard conditions.
5. byproducts or side-effects of the phenomenon

Thus, in order to adequately specify the phenomenon of interest, one should not merely specify what leads to its input-output mappings under standard conditions (precipitating conditions), but also specify under which conditions it fails to occur (inhibiting conditions), how different initial conditions effect the outcome of the phenomenon (modulating and non-standard conditions), and the different consequences ("outputs" in the blackbox analogy used above) it brings about (byproducts of the phenomenon).

Following a counterfactual understanding of causality, knowledge about a causal connection allows inferences about a phenomenon under different counterfactual conditions (Menzies & Beebee, 2024; for an example of a counterfactual theory of causation, see Lewis, 1973). In the present case, this reasoning is somewhat reversed: knowledge about the behavior of the phenomenon under different conditions (although these are of course not genuine counterfactuals) is used to infer the causal structure underlying the phenomenon. Knowing how a kind of phenomenon typically behaves under different conditions does not equate knowing how a particular token of the phenomenon would behave under counterfactual conditions, as every instance of the phenomenon is unitary and thus only knowledge of different instances of phenomena of the same kind is available. However, the latter serves as the best epistemically accessible substitute for the former available to be used in scientific model creation.

2. A mechanistic model must specify the constituent parts of the mechanism in question

As laid out above, the current discussion focuses on what Craver (2007) terms *constitutive explanation (*as opposed to *etiological explanation*), which aims to explain a target phenomenon by elucidating how said phenomenon is brought about by its constituent parts and their organization and interaction. Constituent parts were previously described as mereological parts which are causally relevant for bringing about the target phenomenon. Although this is implied by the notion of constituents as mereological parts, it is worth highlighting – and indeed Craver himself does so – that the constituents picked out by the model must be real constituents of the mechanism 'out there in the world', and not merely useful fictitious entities postulated by the model to account for the phenomenon's input-output mapping. The latter would merely constitute a how-possibly model, which – to reiterate from a previous section – Craver explicitly does not consider explanatory. He makes this notion explicit in his discussion of box-and-arrow diagrams, a type of model commonly found in cognitive psychology, which perform input-output mapping by postulating functionally defined intermediary processing modules (such models and their explanatory status will be considered in more detail later on). Regarding such models, Craver (2006, p. 15) states that

> box- and-arrow diagrams can depict a program that transforms relevant inputs onto relevant outputs, but if the boxes and arrows do not correspond to real component entities and activities, one is providing a redescription of the phenomenon […] or a how-possibly model […] not a mechanistic explanation.

As box-and-arrow models – so Craver (2006) – can be genuinely mechanistic if they manage to pick out non-fictional constitutive parts whose organization and interaction brings about the explanandum phenomenon, they are able to provide scientific explanation. However, especially with abstract components such as the functional modules postulated in box-and-arrow models which do not pick out concrete physical entities but rather functionally defined abstract entities, it can be difficult to assess whether a postulated constituent is real or fictional. Craver recommends using the following heuristics to determine the fictitious status of a given entity postulated in a model. First, parts should "exhibit a stable cluster of properties" (Craver, 2006, p. 370.) and be detectable with independent research methods. This – so the underlying line of reasoning – increases the likelihood of the component in question not being an ad-hoc postulate to fill the gap in one's model, but rather to correspond to actual real-world entities. This recommendation is on-par with best practice recommendations for conducting research in e.g. the cognitive neurosciences, in which multiple, independent sources (for example functional imaging, lesion studies, single cell recordings, etc.) converging on a neural cluster's

involvement in a cognitive function is usually considered the strongest kind of evidence in favor of the functional relevance of the region (cf. Ward, 2019).

Furthermore, proposed constituents should, following Craver, be plausible in the context they are postulated to operate in and respect other theoretical constraints. For example, a mechanistic model of human cognition assuming information transmission at a greater rate than is realizable by the (relatively slow) electrochemical transmission of the human nervous system fails this plausibility constraint. This point does not, it seems to me, require any further elucidation.

Last, individual components should – at least in theory – be able to be manipulated in such a way that they modify the behavior of other components or the mechanism as a whole. To give an example of this, a mechanistic component of reward learning in rats, dopamine receptors, can be targeted for manipulation (e.g. by administering dopaminergic substances such as cocaine) in order to thus intervene in other, coupled constituents and the phenomenon – reward learning behavior – as a whole (cf. Hyman et al., 2006). Although this constraint is useful and often relatively easily applicable in identifying mechanistic components in neurocognitive models (models linking behavior and cognition to neural activity patterns), there are practical problems in the application to models postulating abstract entities such as purely functionally defined box-and-arrow models. Even if the processing modules postulated by such models can be said to constitute genuine, non-fictional components of the mechanism, it is unclear how they could be directly manipulated without having established a further theoretical link to underlying neural activity. Thus, in practice, this criterion fails as a benchmark for establishing the mechanistic validity of purely abstract modules in cognitive models.

3. A mechanistic model should specify the activities and interactions of the constituent components

“Activities” in this context can be simply understood as the “things that entities do” (Craver, 2006, p. 371). As stated previously, the core claim of new mechanism is that scientific explanations work by “filling in” the input-output mapping of a phenomenon by specifying their underlying causal interactions. Thus, the activities specified in a mechanistic explanation are the (causal) processes happening between the constituent entities that lead from the inputs to the outputs. Craver (2006) proposes a manipulationist understanding of mechanism activities. This manipulationist thesis states that

> the activities in mechanisms should be understood partly in terms of the ability to manipulate the value of one variable in the description of a mechanism by manipulating another. (Craver, 2006, p. 372)

More could be said about this view, but for now, it shall suffice to point out that the identification of constituent activity via targeted manipulation is problematic for purely cognitive models, the constituent entities of which are not specified in terms of concrete, manipulable physical entities. This point has been labored before when discussing component identification, and I shall refer to the reader to the ensuing discussion of the status of purely abstract model components, which will dissect their explanatory status under a mechanistic framework in more detail.

4. A mechanistic model should specify the organization of its constituent components

Craver's final criterion for assessing the adequacy of a mechanistic model is the degree to which it specifies the organization of its constituents. As he puts it, "mechanistic explanations are not merely aggregative; the phenomenon is not merely a summation of properties of the component parts." (Craver, 2006, p. 17). Components can be organized in more than one way. Craver highlights spatial and temporal organization in his 2006 proposal of his account. Of these, especially the latter seems not clearly disambiguated from component activities, though Craver does not seem to acknowledge this or evaluate whether this can be potentially problematic. In mechanistic explanation, so Craver, the spatio-temporal organization of the mechanism components ought to enable the component activities leading to the input-output mappings of the phenomenon under examination. It has been emphasized twice so far that it can prove difficult to apply Craver's criteria to purely cognitive models. While the same is true for their spatial organization – locating an abstract processing unit in space is anything but a trivial task – such models can exhibit very fine-grained temporal organization, due to them oftentimes being developed to account for experimental reaction time findings (De Boeck & Jeon, 2019). As such, they can be considered incomplete models in that they specify the temporal organization of the phenomenon components in detail, but only provide a rough sketch of their other organizational properties. Once more, I shall refer the reader to the ensuing discussion of the mechanistic value of purely cognitive models for more detailed considerations on this matter.

The criteria established so far remain largely unchanged between Craver (2006) and Craver (2007). The latter account, however, additionally introduces the notion of constitutive relevance

(Craver, 2007). This addition is motivated by his earlier account leaving it somewhat unclear – by his own admission – what ought to be considered a constituent of a given mechanism and what not. The additional criterion proposed by Craver (2007) is that of *mutual manipulability.* This means that something ought to be considered a relevant component of a mechanism if, by changing the component, one can change the mechanism as a whole, and likewise, by intervening on the mechanism as whole, one can manipulate individual components. More formally, he expresses his conditions as follows (Craver, 2007, p. 153):

(i) X is part of S;
(ii) In the conditions relevant to the request for explanation there is some change to X's Φ-ing that changes S's Ψ-ing; and
(iii) In the conditions relevant to the request for explanation there is some change of S's Φ-ing that changes X's Ψ-ing.

As Craver points out, other than the relationship of causality, which is unidirectional, the component relationship is bidirectional, so that one can change a part by manipulating the whole and change the whole by manipulating one of its parts, hence the notion of mutual manipulability. Once more, it is unclear to what extent this criterion is applicable to purely cognitive components of a model. The reoccurring theme seems to be that – according to Craver – in order to come to conclusions about a mechanism, we should be able to perform targeted interventions on its components, and it is difficult to envision how this could be done for highly abstract functional models which are not specified in terms of their underlying physical structure. One could conclude, based on this difficulty, that functional cognitive models should not be evaluated (purely) based on a mechanistic framework. Rather, one should understand them as adhering to a different form of explanatory standard. Alternatively, one could argue that (Craver's) mechanism provides the proper benchmark to apply to cognitive science, and that cognitive science simply often fails to live up to its standard. Thus, one could either argue for a failure in model-to-epistemic-norm-fit or for a failure in epistemic-norm-to-model-fit. I reject both of these positions. In what follows, I shall attempt to demonstrate that 1.) new mechanism is the proper standard to apply to cognitive science models and that 2.) cognitive science does not fail to provide explanations when scrutinized using this standard (although the latter is able to uncover certain blind-spots and shortcomings in the field).

Before delving into this discussion, let us briefly summarize what has been stated so far. Mechanistic explanations should specify the phenomenon in question under different

conditions, and state the constituent components of its underlying mechanism, as well as how they give rise to the phenomenon through their spatio-temporal organization and joint activities and interaction. Craver (2007) himself summarizes his view as follows:

> The model of a mechanism does not describe capacities of the mechanism as a whole; it describes the activities of the mechanism's components. How-possibly models can be composed of fictional components, but how-actually models describe real components that have multiple properties, that are detectable with multiple techniques, that are utilizable for the purposes of intervention, and that are physiologically relevant. The model of the mechanism also describes the causal relations (activities) that compose the mechanism. These are not mere input-output relationships or *laws in situ* but relationships of manipulability […]. Finally, mechanistic explanatory texts do more than exhibit box-and-arrow diagrams; they reveal the active, spatial, and temporal organization of a mechanism. (Craver, 2007, p. 139)

Mechanism as a Benchmark for Cognitive Science

It was remarked initially that Craver's (2007) position is developed specifically with the field of neuroscience in mind. Although neuroscience and cognitive science are not neatly separable fields, and the former is sometimes viewed as a contributory discipline to the interdisciplinary project of the latter, the types of models proposed in both fields can vary to a significant degree. Whereas cognitive science is typically concerned with cognitive theories, that is, theories of cognitive capacities which aim to explain the latter not by allusion to how they are brought about by their physical substrate, but rather in terms of abstract processing steps, neuroscientific theories tend to account for phenomena in biological terms, which explicitly specify how neuronal activity gives rise to phenomena under investigation, or at least which neuronal activity patterns correlate with which types of cognition and behavior. To what extent is Craver's account, originally developed with the latter in mind, applicable to the former type of model? And are there genuine alternatives? In short, my argument will be that cognitive models provide mechanism sketches and schemas, thus providing genuine, although limited, explanatory value. This account ought to be adopted because, upon closer inspection, alternative models of explanation in cognitive science can only account for cognitive processes to the extent that they make mechanistic considerations. This suggests that the core characteristic to be embraced in explaining such processes is the concept of mechanism. Moreover, Craver's (2006, 2007) theory of mechanistic explanation adequately captures what makes these models explanatory, thus rendering it a suitable framework for evaluating explanation in cognitive science. Let us examine this claim in more detail.

Rational explanation

One possible a priori objection to adopting Craver's framework for evaluating theories in cognitive science may be that the goal of cognitive science is not – at least not exclusively – to uncover mechanisms. Chater (2014) for example argues that cognitive science operates at the intersection of two broad approaches to constructing scientific explanation, the mechanistic approach and the rational approach. His argument for this proposal is that cognitive science understands cognition as a form of computation and information processing, which can be analyzed both from a rational and a mechanistic perspective. As he puts it,

> [i]n rational terms, a computation solves some information processing problem (e.g., mapping sensory information into a description of the external world; parsing a sentence; selecting among a set of possible actions). In mechanistic terms, a computation corresponds

> to a causal chain of events in a physical device (in engineering contexts, a silicon chip; in biological contexts, the nervous system). (Chater, 2014, p. 331)

Simply speaking, a *rational explanation* attempts to demonstrate that a given cognitive process is normatively justified in relation to some normatively rational system, such as mathematical statistics or formal logic. It attempts to show that the input-output mappings generated by a cognitive system make sense given the limited resources available to the system (e.g. in terms of processing capacity, available information, etc.) as well as its ecological niche. What Chater (2014) refers to as *rational explanation* is essentially equivalent to what has been referred to in section two as the "normative Bayesian" approach to cognitive science (and is in terms of its nomenclature quite close to the most prominent exemplar from this category, Anderson's rational analysis approach).

Recall, to give an example of this purported type of explanation, the biases and alleged deficits in logical and probabilistic thinking previously outlined in section 2. A rational explanation would not aim to demonstrate how these seemingly flawed outputs come about as a consequence of the interactions of constitutive parts underlying the mechanism bringing forth cognition, but rather why they may be normatively justified. Sher and McKenzie's (2008) reference-point account of probabilistic reasoning in the 'Asian disease' task falls neatly into this category. As such, rational explanations are useful in demonstrating how "apparently boundedly rational, or downright irrational, behavior may sometimes result from a misspecification of the problem faced by the agent" (Chater, 2014, p. 333). This type of theorizing seems to operate in an entirely distinct domain of scientific modeling and theorizing compared to the mechanistic approach.

Based on Chater's perspective on theory construction in cognitive science as operating in two distinct explanatory domains, the conjunct of two related claims may be raised as an objection to the application of Craver's mechanistic account to cognitive science: First, there are non-mechanical types of modelling/theory construction in cognitive science. Second, this type of non-mechanical modelling is valid by itself as a source of scientific explanation and does not hinge on mechanistic considerations. Put slightly differently and employing the nomenclature introduced in the previous sections, this objection states that normative Bayesian models are explanatory independent of any mechanistic account, and therefore offer an alternative to mechanistic explanation in cognitive science.

If cognitive science, as Chater claims, operates using these distinct types of explanation, a failure to adhere to mechanistic standards of scientific explanation need not be problematic for a model in cognitive science at all. That is, if a cognitive system can be explained in a purely rational manner, using a mechanistic framework to criticize said explanation is a misguided endeavor; one would commit an epistemic category error. Is this the case?

I shall argue that to a certain extent, rational models can provide some degree of explanation of a cognitive system. However, this degree is very limited, and rational explanation does not constitute a fully fletched, genuine explanation of a cognitive phenomenon. Moreover, to the extent that rational explanation is explanatory, it is mechanistic. The reason for this is that explanation and justification are conceptually distinct. The goal of explanation in cognitive science is to show how a cognitive system operates and why it operates in the manner it does. The goal of justification (in the context of cognitive science) is to show that the operations performed by a cognitive system are legitimate in relation to a system of norms. There is, however, no apparent conceptual link or overlap between these two notions.

The notion of a cognitive system emerging which fails to adhere to standards of a normative evaluation based on a reference system of rationality is entirely coherent and seemingly unproblematic, even if the operating principles of the system are fully transparent and understood. The system, in this case, would be fully explainable, but lack normative justification. Likewise, it seems completely coherent to imagine a cognitive system (for the sake of illustration, let us use an alien artefact as an example), the *how*s and *why*s of which are completely opaque despite the system being normatively justified. To continue the alien artefact example, let us imagine said artefact, which someday mysteriously appears on earth, takes as input any natural number, and returns as output the prime factorization of said number (the exact shape and mode of input and output of the artefact are left to the reader's imagination). The operating principles of this artefact would be justified in relation to the system of norms provided by the fundamental theorem of arithmetic. However, intuitively, it seems as though, despite this justification, the *how* and *why* of this system remain unaccounted for. Put differently, the system would not be explained, but it would be justified. This aims to show that justification and explanation are not inherently conceptually linked, for we can imagine one without the other.

However, it could well be the case that the justification of a cognitive system and its explanation are linked not as a matter of a priori conceptual connection, but as a matter of empirical fact.

This would be the case if, usually, the normative justification of a cognitive system carried information about the operating principles of said system, thus helping in its explanation. In this case, the justification, although itself not explanatory, would indicate how the system could be explained. For example, if every system operating in accordance with the principles of first order predicate calculus shared a distinct physical substrate with distinct operating principles, then the mere fact that the system operates in this manner would carry information about how the system is explainable. This particular example seems far-fetched and at first glance implausible, but the core idea does have some merit.

Making the (seemingly sensible) assumption that normatively justified modes of cognition are adaptive in biological cognizers and are thus – ceteris paribus – selected for in the course of evolution, rational explanation can answer why a system displays a certain behavior by explaining why the behavior in question was selected for by showing how it is adaptive for its ecological niche. In that sense, it contributes to the *why*-aspect of scientific explanation. It is part of the explanation of how, phylogenetically, a cognitive system came to be the way it is, namely by providing a selective advantage due to its niche adaptiveness. In this sense, a rational explanation of an evolved cognitive system provides (part of) an etiological mechanistic explanation of the system, as the rationality of the system played a role in the causal process shaping the system. To reiterate, etiological explanation, according to Craver (2007), explains a phenomenon by reconstructing its causal history. Therefore, the mere fact that a cognitive system is rational can, in this strictly limited sense, provide information about the etiological explanation of the system. This does not mean, however, that rational explanation can provide anything over and above mechanistic explanation or provide an alternative to it. What evaluating the normative validity, i.e., the rationality of a system given its niche constraints, does is provide a story about why the system was selected for, how it came to be the way it is, namely, by adapting to its environment. This phylogenetic story, however, is a mechanistic one; it uncovers the causal mechanism by which a cognitive system was shaped by evolution. The normative rationality of the system in regard to its specific niche is a consequence of this causal process, and a valuable tool for shedding light on the latter. It does, however, not provide additional explanatory value. If the rationality of the cognitive system – gradually emerging via natural selection – did not play a causal role in shaping the system's current form, the rationality would lack any explanatory value.

Granted, this type of causal-mechanistic explanation does not fall under the type of *constitutive* mechanistic explanation Craver (2006, 2007) develops in detail in his account, but is rather a

kind of *etiological* mechanistic explanation. Etiological explanation is, however, still very much part of the mechanistic conception of explanation, rather than a different and distinct form of non-mechanistic explanation. Rational analyses are explanatory insofar as they shed light on the (mechanistic) causal story of how a cognitive system was shaped. The mere fact that it is rational adds, in isolation, nothing to this explanation. By showing how a cognitive system is ideally adapted to its environment we shed light on its causal history and thus add to our understanding of it, not by demonstrating its relation to a set of norms of rationality.

Functional analysis

A second objection one could raise to the application of mechanism to cognitive science is that although rational analyses are not an explanatory alternative to mechanistic explanation, functional analyses are. Purely cognitive models usually specify the interactions between highly abstract entities, such as certain functionally specified modules or increasingly simple homunculi (cf. Dennett, 1993). To illustrate this common character of cognitive models, consider, as an example, the pandemonium model of visual image processing (cf. Sternberg & Sternberg, 2016). This model describes a hierarchical system for visual object recognition, which consists of a number of metaphorical "demons", each of which fulfill a specific task, the result of which is passed on to the demons at the next level in the hierarchy: lower-level demons detect simple features in a visual scene, such as edges or points. They pass these patterns on to higher-level demons which look for complex patterns having those simpler features as constituents, such as shapes or, at an even higher level, objects. The pandemonium model is a typical exemplar of a purely cognitive model in that it aims to demonstrate how seemingly high-level processes such as object recognition can be implemented by the cooperation of simple and by themselves unintelligent information processing modules (Sternberg & Sternberg, 2016). These modules are functionally defined in that they remain agnostic about the actual physical substrate that brings about the input-output mapping of the explanandum phenomenon. As such, they do not specify any concrete physical entities or their causal interactions, which, so the objection may go, is required to provide a mechanistic explanation.

Initially, this objection may seem plausible given Craver's understanding of mechanism. To reiterate, a mechanistic explanation should – following Craver – allude to the parts constituting the mechanism that brings forth the phenomenon. These parts should have a stable cluster of properties and constitute the mechanism in question by their spatial, temporal, and organizational properties. To what extent can this be said to be the case for purely cognitive

models? Such models do not specify any concrete physical entities and their interactions, but rather the relations holding between abstract processing units and their joint behavior. As such, purely cognitive models completely fail to specify any spatial aspects of a mechanism. The question thus arises (and has been brought up repeatedly in the foregoing introduction of Craver's criterions for mechanistic explanation) to what extent it makes sense to apply Craver's mechanistic view to such purely cognitive models.

One could of course bite the bullet and argue that Craver's view of scientific explanation is correct and applicable to cognitive science, but that cognitive models fail to adhere to its standards of explanation due to the aforementioned issues. However, I shall not adopt this line of reasoning. I take it that purely cognitive models are – at least potentially – capable of providing explanations of phenomena (at least to a certain degree), and that Craver's account is capable of accounting for why this is the case.

Initially, it may seem like the idiosyncrasies of cognitive models are better captured by an approach along the lines of Cummins (1975) functional analysis model. This approach states, as aptly summarized by Von Eckardt and Poland (2004, p. 975), that

> a system's or organism's capacity to do something can be explained by a *functional analysis* of that capacity. [A] functional analysis of some capacity *C* [is] a breakdown of the complex capacity into a set of constituent capacities [and] a specification of the sequence in […] which those constituent capacities must be exercised for the result to be a manifestation of the complex capacity.

As such, rather than specifying the spatial, temporal, and organizational details of a set of mechanism-constituents displaying stable properties, Cummins' explanation via functional analysis merely requires the specification of a set of constituent capacities (what I have previously referred to as "modules" or "homunculi") and their temporal order. Although this epistemically lower bar to clear is certainly fulfilled by many models in cognitive science (such as the Pandemonium model), Cummins' approach is problematic in that it is too permissive on what constitutes a scientific explanation. As Von Eckart (1977) points out, for any given (cognitive) capacity, various different functional analyses capable of reproducing their input-output mapping are possible. Any one of these analyses has, following Von Eckardt, merely the status of a how-possibly model, providing a possible story how a cognitive system may transform its inputs to outputs, without giving any additional reasons for assuming that this functional mapping is in fact the one carried out by the system in question. She thus proposed

the further criterion of such an analysis being *structurally adequate*, to reconcile Cummins' approach with the intuition that mere how-possibly stories are not genuinely explanatory. She defines this criterion as follows:

> A functional analysis A of how an organism O has capacity C is structurally adequate if there exists a structural decomposition of O into component (physical) parts such that
> (1) Operation of those parts results in O's exercising C.
> (2) For each constituent capacity of A, there exists at least one of those structural components that has that capacity.
> (3) The order in which those component parts operate when O is exercising C mirrors the algorithm specified in A.
> Condition (1) guarantees the existence of a physical mechanism underlying the capacity C. Condition (2) ensures that some part of the mechanism corresponds to each constituent capacity specified by the functional analysis. Condition (3) guarantees further that the sequence of operation of these physical parts corresponds to that specified by the functional analysis. (Von Eckardt, 1977, p. 41)

Thus, Von Eckardt argues that a functional analysis is structurally adequate if and only if it is realized by a physical mechanism. This approach is strikingly similar to Craver's who, to reiterate, argues that purely functional stand-ins in mechanism sketches ought to correspond to actual physical entities bringing them about. If there are no such corresponding mechanism-constituents, or in Von Eckardt's terminology, if the functional analysis is not structurally adequate, the mechanism sketch is a mere how-possibly model and thus unable to provide genuine explanation.

In a subsequent publication, Von Eckardt and Poland (2004) explicitly endorse this understanding of purely functional analyses as mechanism sketches.

> [A purely functional analysis] is only a mechanism sketch because constituent capacities are hypothesized without reference to the constituent entities that (presumably) have those capacities. However, a functional analysis combined with a functional component model is equivalent to a complete constitutive mechanistic explanation. (Von Eckardt and Poland, 2004, p. 976).

A similar position is adopted by Piccinini & Craver (2011 p. 2), who

> argue that such decompositional, constitutive explanations gain their explanatory force by describing mechanisms (even approximately and with idealization) and, conversely, that they lack explanatory force to the extent that they fail to describe mechanisms.

In formulating their argument, Piccinini and Craver go against what may be termed the „received view“, which considers functional analysis to be distinct and independent from mechanistic analysis. This view maintains that while mechanistic explanations allude to structural features of (the components) of a system, such as size, shape, mass, etc., functional analyses allude to functionally specified components, which are individuated merely by their functional features, i.e., input-output relationships, as opposed to any structural features.

Their argument against this distinction is that mechanistic explanations often refer to functional properties and are not limited to structural ones. Furthermore, the strict distinction between these two sets of properties does not hold up under closer scrutiny. The functional role of a mechanistic component constrains which type of structural properties can bring it about. Vice versa, the structural properties of a component constrain the functional role it can play. In their terms,

> the functional properties of a black box constrain the range of structural components that can exhibit those functional properties. On the other hand, a set of structural components can only exhibit certain functional properties and not others. (Piccinini & Craver, 2011, p. 19)

Thus, contrary to what proponents of pure functional analyses may claim, functional characterizations of components are not independent of structural properties. An analysis proposing functional components which are not realizable in a system due to its structural constraints cannot provide a genuine explanation of the system’s capacities, but rather a mere how-possibly story. For example, a functional analysis of pain conduction in the human nervous system which requires information flow at a speed greater than what is permitted by the relatively slow conduction speed of the central nervous system cannot provide a genuine explanation of the phenomenon. Based on these considerations, Piccinini and Craver propose to eliminate the functional-mechanistic dichotomy, and instead treat functional analyses as mechanism sketches:

> Any attempt to specify the nature of a component purely functionally, in terms of what it is for, runs into the fact that the component’s function is to interact with other

> components to exhibit certain physical capacities, and the specification of the other components and activities is inevitably infected by structural considerations (Piccinini and Craver, 2011, p. 20)

Thus, although it is true that purely cognitive models need not correspond one-to-one to any readily identifiable physical mechanism, such models make constraints on organizational and temporal aspects of components, the sequence of processing steps, the required degrees of freedom in the system, etc. This is precisely what a mechanism sketch under Craver's mechanistic framework does, which provides an "incomplete model of a mechanism" (Craver, 2006, p. 5). Such a sketch traces the outlines of the mechanism bringing forth the phenomenon under investigation, while not yet making strong commitments about (all of) its specific implementational details.

A mechanism sketch, as stated previously, is not itself a complete explanation. It does not have the same kind of explanatory value that a more detailed mechanistic account has, an account which is able to tell a more detailed, fine-grained story about how a phenomenon is brought about by the interactions of its components. Applied to purely cognitive models in cognitive science, this would entail that, although such models provide explanatory value, they are not the end of the story, and benefit from being expanded by adding implementational details specifying how exactly the purely abstract processing units proposed in the model are brought about.

This way of thinking about cognitive models is congruent with actual scientific practice. Cognitive neuroscientists go through great lengths to identify the neural correlates of cognitive capacities, i.e., which kinds of neural activation patterns bring forth which kind of cognitive capacities. This is nothing more or less than trying to fill in the details of the rough cognitive outline of the mechanism underlying said capacities. Hardly anyone, I imagine, would be inclined to deny that if the aim of identifying the mechanistic details bringing about a given cognitive capacity *Y* was successful, then explanatory value would be added to our model of *Y*. If cognitive neuroscientists were able to demonstrate how the interacting activation patterns of neuronal populations bring about *Y*, we would, I conjecture, be inclined to say that they enabled us to better explain *Y*, as opposed to insisting that *Y* was already an explainable phenomenon before this breakthrough. Filling in these mechanistic details expands our understanding of the phenomenon under investigation, it expands our capacity to explain these phenomena and thus adds explanatory value.

So far, it has been argued that neither rational nor functional analyses constitute genuine alternatives to mechanistic explanations in cognitive science. Although both these approaches to scientific explanation can have valid applications, they achieve their explanatory power through the mechanistic model constraints they provide, rather than as an alternative to it. Besides these two frameworks, there is another influential account of explanation in cognitive science, namely that of Marr's (1982) levels of explanation. I shall argue that, like in the case of functional or rational explanation, this account is not a genuine alternative to mechanistic explanation, but rather works because it provides mechanistic constraints.

Marr's Levels of explanation

In his seminal work on visual perception, Marr (1982) specifies different levels at which a cognitive system can be described and explained. These levels are, as summarized by Marr (1982, p. 25) himself:

i. Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

ii. Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?

iii. Hardware implementation: How can the representation and algorithm be realized physically?

The idea of Marr's levels is that genuine explanations of cognitive phenomena can be given at each of these levels (semi-)independently. Although all describe the same system, an explanation at the computational level need not evoke representational or implementational considerations to be valid. This does not entail, however, that the accounts can be contradictory, e.g., by specifying a computational task at the highest level which is not carried out by the algorithmic structure given at the representational level. Marr's idea is that the representational account describes by which algorithmic manner the computation specified at the highest level are carried out, much like the hardware description demonstrates how these algorithms are implemented in a physical system. "Independence of the levels" is hence to be understood merely as implying that a higher-level explanation need not invoke considerations at the lower levels, not that they are not ontologically dependent upon each other. Furthermore, following

Marr, to fully understand a cognitive system, one must specify explanations at all three levels of explanation.

These aspects of Marr's levels-of-explanation account, the ontological dependence of higher on lower levels and the need for a specification at all three levels, is the reason that, just like in the case of rational or functional analyses, the tripartite account of explanation in cognitive science is not a genuine alternative to mechanism. The very same argument previously directed against the view that the functional analysis account provides an alternative to mechanistic explanation applies in this case as well. To reiterate, this argument stated that supposedly independent higher-level, abstract characterizations of a cognitive system are constrained by mechanistic considerations, and likewise provide constraints about the type of mechanism capable of bringing forth these higher-level processes. A model providing constraints on the exact mechanism implementing a given phenomenon, however, is nothing but a mechanistic model, or at least a sketch of one.

Just like in the case functional analyses, computational and algorithmic accounts are merely mechanistic models sketches which leave out the nitty, gritty details of implementation, tracing a more general and abstract description of the relevant mechanism instead. Piccinini and Craver (2011) point this out for computational explanations when they write that

> such an explanation [a computational one] is still mechanistic: it specifies the type of vehicle being processed (digital, analog, or what have you) as well as the structural components that do the processing, their organization, and the functions they compute. So computational explanations are mechanistic too (Piccinini and Craver, 2011, p. 21)

The same holds for algorithmic explanations, which are, in regard to their explanatory value, roughly equivalent to functional analyses: they provide an abstract characterization of the causal processes underlying a certain cognitive capacity. Instead of explicitly referring to individual physical entities and their interactions, they abstract away from such details to provide characterizations of these causal processes in more abstract terms.

Using the terms of Craver's mechanistic framework, computational descriptions of a system are essentially equivalent to phenomenological mechanistic models. Much like phenomenological models, system descriptions at Marr's computational level "sit at farmost limit of sketchiness" (Craver, 2006, p. 6). They specify the computational goal of a system without revealing any insights into how this goal is brought about. They specify which input-

output relations a system brings forth, while remaining agnostic about the actual transformations and interactions going on 'beneath the hood' which mechanistically account for these input-output mappings. Marr's algorithmic level explanations correspond to what Craver termed "model sketches". These sketches abstract away from a range of details of the underlying causal processes bringing about the input-output mapping in question and provide an overview over how some (abstract) entities interact to bring forth this mapping, without being concerned about the precise spatio-temporal causal interactions underlying this process. Last, the hardware/implementational level lies closest to Craver's *ideally complete descriptions of a mechanism* end of the modelling spectrum. At the hardware level, spatio-temporal interactions of physical entities which realize the algorithmic computations are specified. Whereas an algorithmic description of a computer program would specify the code underlying the program's performance, the implementational level specifies the electrical states in the computer hardware underlying the program execution.

It seems unlikely that Marr himself would have many objections to this view. Following his original formulation of the tripartite model, all three levels of explanation are needed to fully grasp a cognitive phenomenon. This idea is very much in line with Craver's model, as it essentially states that a phenomenon requires a fully specified (at different levels) mechanism in order to be explained. Purely computational or algorithmic accounts thus ought to be extended by specifying their implementational details. This requirement is virtually identical with Craver's requirement of filling in the details of an abstract model sketch in order to achieve a more complete explanation. The line of reasoning given above is thus not so much directed at Marr's tripartite division (which, as we have seen, is very much in line with Craver's account), but rather against a hypothetical objection based on this account.

This concludes our discussion of the applicability of Craver's model to explanation in cognitive science. It has been attempted to show that the types of explanation often proposed as alternatives to mechanistic explanation in cognitive science either are mechanistic at their core and can thus be integrated into and identified with types of explanations in Craver's framework, or do not really provide explanation at all. Purely cognitive models, insofar as they are adequate and capture the real structure of the phenomena they aim to represent, are best understood as mechanism sketches from the perspective I have endorsed. They narrow down and constrain the space of possible mechanisms underlying a given phenomenon, and thus provide genuine explanatory value even under a purely mechanistic framework. Having thus established the adequacy of assessing cognitive science using such a framework, let us proceed by critically

assessing the types of cognitive models introduced in the first two sections, Bayesian cognitive models, from this viewpoint.

The Explanatory Power of Bayesian Cognitive Models

Despite much ado about the unificational potential of the Bayesian approach to cognition (cf. Griffiths et al., 2010), the foregoing discussion in section 2 attempted to demonstrate that the field of Bayesian cognitive science is far from unified. There exist stark differences both in the methodological approach as well as in the ontological commitments of different Bayesian accounts. As such, there can be no unitary, generalized evaluation of the field of Bayesian cognitive science, for there is no such thing as a unified field to begin with. As such, different Bayesian approaches require different assessments in terms of their epistemic qualities, for they differ drastically in the kinds of models they propose. The aim of the present section is to provide such a nuanced assessment. To achieve this, methodologically speaking, it is necessary to strike a balance between generalizing over key differences among various approaches, which risks reducing the adequacy of the assessment in capturing individual differences, and delving too deeply into individual details, which may obscure general trends and shared structures among the approaches. To account for this necessity, I shall give separate assessments of trends in what has been termed normative and descriptive Bayesian cognitive science and the two most important and influential frameworks within these subfields, namely Anderson's rational analysis approach and the predictive processing framework. Although not all Bayesian approaches to the mind can be subsumed under these latter two frameworks, they exemplify the most important and influential general trends which can be found in Bayesian cognitive science. To what extent are these approaches capable of providing insight into the mechanisms underlying the respective phenomena they aim to capture?

Descriptive Bayesian Models

I have termed those Bayesian approaches descriptive which do not commence with an analysis of optimality in a certain niche or for a certain inferential problem, but which instead begin with a characterization of the cognitive process under consideration itself, using the mathematics of Bayesianism to capture some aspect of it. The most prominent theoretical framework aligned with this method is, as stated before, predictive processing. To reiterate, predictive processing posits that the mind operates as a hierarchy of increasingly abstract information processing units. These units make predictions about the activity at the next lower level and propagate back any deviations from these predictions. The generative models underlying the predictions made at each level of the hierarchy are then updated in light of the error signal reported from the lower level in line with the principles of Bayesian inference. Thus, each level in the hierarchy

essentially performs Bayesian model updating based on the information from the level directly beneath it.

Fink and Zednik (2017) argue that this framework captures both computational and algorithmic aspects of cognition under the tripartite model of explanation in cognitive science. The computational aspect pertains to the overall goal of cognition under this framework, namely, to minimize prediction error. The algorithmic aspect is that the predictive coding framework offers a glimpse into the representational transformations enabling this computational goal to be achieved, namely, an interactive, increasingly abstract hierarchy of Bayesian units reporting back error signals. They evaluate the explanatory potential of this framework as follows:

> the algorithms developed within the HPC [hierarchical predictive coding, read "predictive processing"] modeling framework can be viewed as potential answers to questions about *how* a particular cognitive system does what it does, i.e. as descriptions of the functional processes that contribute to that system's behavior. Although much work has yet to be done to determine which (if any) of these algorithms actually constitute a *correct answer*—i.e. a true description of functional processes in our brains—the fact that answers to how-questions are being developed is often considered the central explanatory contribution of HPC (Fink & Zednik, 2017, p. 5)

It was previously argued that Marr's levels of explanation can be understood as three discrete points on Craver's (2006, 2007) spectrum of granularity of mechanistic model specification, and that the former account is thus compatible with the latter. What does this entail for the explanatory status of predictive processing from a mechanistic viewpoint?

Predictive processing, given the operating principles of cognition it proposes are accurate, seems to have genuine explanatory potential, in that it proposes a sketch of the mechanism underlying biological cognition. As was argued previously, such model sketches contribute to explanation by tracing the outlines of the mechanism giving rise to a phenomenon.

The sketch proposed by predictive processing is a hierarchy of Bayesian nodes, each connected to the ones above and below it by a relation of error and prediction feedback. This paints a picture of the parts of the mechanism in question, namely Bayesian nodes, their interactions, prediction and error feedback, as well as their organization, a hierarchical network. As such, the Predictive processing framework seems to be genuinely mechanistic in its explanation of target phenomena.

However, the mechanistic framework defended in the present thesis not only requires a mechanistically plausible just-so story to achieve genuine explanatory power, but also an adequately robust and independently informed mechanism model. 'Robust and independently informed,' in this context, entails that the proposed components of the model be empirically assessable, among other things via direct manipulation.

It is this regard in which the predictive processing framework is lacking, for the abstract mechanistic structure it proposes seems difficult, if not impossible, to empirically judge without further assumptions about its physical substrate and concrete implementational details. In a sense, the predictive processing framework is holistic in its approach: it provides a general picture of how certain cognitive capacities are achieved by proposing a large and complex mechanistic network composed of a number of functionally specified constituents. In this core framework of the theory, however, the empirical adequacy of these constituents cannot be, it seems, independently assessed. There is no immediately available method for either manipulating the whole mechanism so as to change its constituents in an observable manner, nor does there seem to be a way to manipulate the constituents so as to thereby change the behavior of the entire mechanism, at least not without first establishing further theoretical links to more fundamental physical constituents. This constitutes a breach of what Carver (2007) termed the *mutual manipulability* criterion, which was laid out in the previous sections. To reiterate, this criterion states that a mechanism and one of its constituents are mutually manipulatable if

(i) X is part of S;
(ii) In the conditions relevant to the request for explanation there is some change to X's Φ-ing that changes S's Ψ-ing; and
(iii) In the conditions relevant to the request for explanation there is some change o S's Ψ-ing that changes X's Φ-ing.
(Craver, 2007, p. 153)

As a matter of ontological structure, this criterion does hold for predictive processing. By changing a constituent of a Bayesian hierarchical network, for example the generative model in one of the nodes, one would thereby change the behavior of the overall mechanism, in the given example by changing the type of error feedback that is received, how the generative model is thus adjusted, and how this influences the other, connected predictive nodes at higher and lower ranks in the hierarchy. Likewise, by manipulating the mechanism, e.g., by placing it in a radically different environment, one would thereby change the individual component, in the

example of the predictive node by rendering its predictions wildly inaccurate and hence leading to a drastic change in its generative predictive model.

However, while this relation holds at an ontological level, it remains elusive from an epistemic standpoint. In other words, while the overall mechanism and its parts may be mutually manipulatable as a matter of nomological possibility, this mutual manipulability cannot inform the modelling process. The reason is the lack of implemental specifications of the mechanism, which, to reiterate, is only specified in purely functional terms, lacking – besides very coarse constraints – any identification of the constituents' physical properties. These very coarse constraints dictate that the lowest level of the predictive hierarchy must be somehow directly connected to sensory receptors, such as photosensitive cells in the eye or mechanosensitive cells in the skin, as the functional designation of these low-level nodes in the hierarchy is to predict their inputs. As – by the functional specification of the predictive hierarchy – the only information transaction between levels in the hierarchy is via prediction and error feedback, the lowest levels in the hierarchy must therefore be directly connected to sensors in order to directly predict their activation (for a more detailed discussion of the mechanistic constraints of Predictive processing, cf. Miłkowski & Litwin, 2022). This highlights, once more, the previously discussed notion of the inseparability of functional and mechanistic considerations. However, due to the very rough and unspecific mechanistic constraints this provides, it does not suffice to allow for the manipulability of individual components called for by Craver's (2006, 2007) mechanistic theory.

In the absence of more detailed mechanism specifications, and more importantly, ways to assess it, the mechanism sketch provided by predictive processing remains a mere how-possibly model. This point is argued by Miłkowski & Litwin (2022), who state that in the absence of more robust empirically assessable mechanistic commitments, predictive processing is little more than a computational modelling tool:

> It can hardly be empirically disproved as it may be flexibly adapted to provide any explanation, also for phenomena which actually do not (or cannot) exist. This is because hierarchical PP lacks essential detail regarding the functioning of cognitive systems, and particular PP models can be easily customized through ad hoc assumptions and parameter adjustments. While this is common malpractice in computational modeling in cognitive (neuro)science, *we urge PP defenders to fill in the important details of their grand theory* [emphasis added]. Otherwise, it may turn out to be just a series of uninformative platitudes,

> which, even if falsifiable, would not offer satisfactory new understanding in cognitive science. (Miłkowski & Litwin, 2022, p. 461f.)

However, despite this genuine problem of the framework, some mitigating remarks on the explanatory power of predictive processing are nevertheless called for. Using Lakatossian terminology (cf. Lakatos, 1978), predictive processing can be said to have a "hard core" of assumptions about a hierarchical predictive structure of the mind, surrounded by a "protective belt" of additional assumptions, concrete mathematical and computational implementations of this general structure, and auxiliary hypotheses, all of which significantly vary between individual proposals within the general framework (cf. Keller & Mrsic-Flogel, 2018). As such, despite not having arrived at a unitary consensus regarding the mechanistic details of predictive processing, there have been various attempts to flesh out the said details. For example, Keller and Mrsic-Flogel (2018) propose a neural-circuit level implementation mechanism of hierarchical predictive processing in the cortex, explicitly invoking mechanistic considerations in doing so. They

> argue that existing data about neural activity and neural circuit organization of the (sensory) cortex can be understood in the context of a predictive processing framework, and we highlight recent direct evidence in support. We then discuss how computations required for predictive processing might be implemented at the circuit level and propose experiments that would *provide a mechanistic corroboration* [emphasis added]. (Keller & Mrsic-Flogel, 2018, p. 425)

The exact biological details of their account do not matter for our present purposes, what is rather of interest is the philosophical stance towards scientific explanation implied by their proposal. By speaking of "mechanistic corroboration", the authors implicitly acknowledge the problem of a purely abstract how-possibly model and attempt to fill in the gaps left by such an account. In doing so, these authors are far from alone. Schultz et al. (1997) propose that the dopaminergic reward system implements the type of prediction-error updating proposed by predictive processing. Wolpert et al. (1998) propose certain neural circuits in the cerebellum to be a neural substrate of internal predictive motor models. Last, Zou et al. (2022) attempted to identify cell types and specific neuronal circuits possibly involved in bringing about the predictive structure proposed by the predictive processing framework.

This short and unsystematic overview is but a glimpse into the vast array of literature published on the potential mechanistic implementation of predictive processing, but it highlights on key

aspect: mechanistic considerations are taken seriously in the field, and – despite not yet having arrived at a consensus regarding the mechanistic details of Predictive processing – active efforts are undertaken to provide a mechanistically complete picture.

Let us now explicate the degree to which mechanistic considerations influence research in the second major branch of Bayesian cognitive science, normative modelling, and its most influential theoretical framework, rational analysis.

Normative Bayesian Models

The second dominant Bayesian framework in contemporary cognitive science is that of Bayesian rational analysis, following and inspired by the work of Andersen (1991, 2013, 2014). To reiterate, Bayesian rational analysis starts with analyzing the probabilistic structure of an environment and the corresponding computational task a cognitive system must solve in navigating it. It then proposes an optimal solution following the normative guidelines set by Bayesian statistics. In a last step, it assesses whether actual (typically human) behavior resembles the optimal Bayesian solution. Due to this primacy of normative considerations, I have termed Bayesian rational analyses (together with similar approaches) a *normative Bayesian* framework. What is the explanatory value such normative models can provide? Fink and Zednik (2017) argue the following:

> The characterization of behavior and cognition as a form of optimal probabilistic inference allows investigators to answer questions at Marr's computational level of analysis […]. Specifically, it allows them to answer questions about *what* a cognitive system is doing, and *why.* In general, whereas an answer to a what-question is delivered by describing a system's behavior, an answer to a why-question is delivered by demonstrating this behavior's "appropriateness" with respect to the "task at hand". (Fink and Zednik, 2017, p. 3)

Thus, they propose that a purely computational analysis aims to answer what a cognitive system does, and, importantly, *why* it does it. They continue their assessment by claiming that Bayesian rational analysis is, or at least can be, successful in providing answers to both questions:

> Regarding questions about what a system is doing, whenever an optimal solution is closely approximated by the behavioral data, the former provides an empirically adequate description of the latter. As for questions about why a system does what it does*, there is a clear sense in which the system can be thought to behave as it does because that way of*

> *behaving is optimal in the sense prescribed by probability theory* [emphasis added]. (Fink and Zednik, 2017, p. 3)

I have no objections against the first conjunct of their two-part claim. Computational specifications, if they provide an adequate approximation of the behavior a system displays, can be understood as describing what the system does, i.e., which computations (in case of a cognitive system) it carries out. In that sense, Bayesian rational analyses can serve as phenomenological models, in that they accurately map which behavioral outputs a system will generate based on a specified set of probabilistic environmental inputs. Fink and Zednik (2017, p. 3f.) themselves aptly summarize this contribution when they state that

> [b]y providing formal answers to questions about what a cognitive system is doing, proponents of this modeling framework can attain a heightened understanding of the nature of cognition and behavior itself, including its mathematical structure.

However, I shall argue that their error lies in the claim that rational analysis — barring some exceptions — is capable of explaining why a cognitive system displays a certain behavior. The mere optimality of a cognitive process, as I have previously argued, does normatively justify the process in relation to a certain set of epistemic norms, but it does not explain why the process takes place. Cognitive systems such as humans are computationally suboptimal in a variety of ways. The reason for this can be mechanistic limitations, such as energy consumption of more optimal cognitive solutions, or computational tractability. Humans cannot solve complex multidimensional integrals in their head, even though computationally, such a solution would be optimal. Given that human cognition is, at least following Bayesian rational analysis, optimal in regard to some computational problems, but not optimal in all regards, the mere optimality of a computational solution is not sufficient grounds for declaring the capacity that produces it as explained. For that to be the case, as was argued earlier, one would need to demonstrate that a solution's optimality is in fact the reason it is produced. This connection is not trivial and cannot be assumed to hold without further argument. Danks (2008, p. 62) aptly summarizes why optimality itself is not explanatory when he states that

> [r]ational analyses aim to provide optimality-based explanations […]. In practice, however, rational analyses almost always consist solely in optimality analyses that show that a particular behavior is optimal for an environment, agent, and problem, followed by experiments to confirm that the optimal behavior occurs. An optimality analysis alone, though, is insufficient for an optimality-based explanation, as there are many other reasons

why X might occur. People might act optimally because of historical accident, or because there are no other options, or a number of other reasons.

Therefore, merely showing that a behavior is optimal is not sufficient for explaining why the behavior is taking place, because the behavior's optimality need not be involved in bringing about the behavior in question. A similar point has been argued previously in regard to the explanatory nature of rational "explanation" independent of its allegiance to Bayesianism. In this discussion, it was concluded that justification and explanation are conceptually distinct, and justification/optimality is only explanatory insofar as it is relevant in bringing about the behavior under consideration. This would be the case, as was argued, if a certain behavioral trait was selected for because it proved adaptive in a certain environment because of its optimality, thus providing a selective advantage.

This line of reasoning was employed in arguing that rational explanation is, despite some claims to the contrary, not a distinct and independent kind of explanation, but rather that it's explanatory power stems from mechanistic considerations. However, in the context of the present discussion, it also sheds some light on the explanatory contributions of rational analysis: Assuming that a certain cognitive process is an adaptive trait for a certain environment can account for the selection of that trait, and thus for its persistence in a population. This type of explanation, explanation via evolutionary adaption, does provide a genuinely explanatory causal story about how a trait came to be, and hence an etiological mechanistic explanation. In this sense, it can answer why a cognitive system performs a certain cognitive process.

While this is a valuable and interesting contribution to our understanding of the mind, the goal of cognitive science as it is usually understood is to study cognitive systems themselves, i.e., how they operate, rather than their evolutionary history. In other words, to some extent, rational analysis can answer the why-question for some cognitive systems (given they evolved due to optimal niche adaptation), but what can they contribute to answering the how-question?

At first glance, a mechanistic answer to this question must seem rather pessimistic. As (constitutive) mechanism is the position that to explain a phenomenon is to elucidate its underlying mechanism, and since rational analysis explicitly (as per its conception by Anderson 1990, 1991) remains agnostic about any constitutive mechanistic considerations, it seems as though rational analysis is entirely non-explanatory in regard to the how-question. As Reijula (2017, p. 4) summarizes this point,

> […] mathematical models in systems- and cognitive neuroscience [for our purposes, read "cognitive science"] can be explanatory only if there is a mapping between elements in the model and elements in the mechanism for the phenomenon. […] [R]ational analysis models provide no such mapping. They are agnostic about algorithmic and implementation level details, and intentionally so. They clearly do not track constitutive dependencies.

As mathematical models specifying the input-output relations of a given cognitive system, Bayesian rational analyses are best characterized as phenomenological models in Craver's taxonomy. As such, these models have some minimal explanatory power, as I have argued earlier, as they contribute to narrowing down the range of possible mechanisms due to the mathematically precise input-output mapping they provide. Although this is a genuine explanatory contribution, their explanatory status should not be overemphasized. As Craver (2006, p. 6) has, as previously cited, put it, merely phenomenological models "sit at the farmost limit of sketchiness". Thus, out of Craver's four criteria of mechanistic explanation, they merely contribute to the first, namely, specifying in detail the phenomenon to be explained. They make no specifications whatsoever in terms of the constituent components of the cognitive mechanism in question, nor their activities, interactions, or organization. To reiterate, the specification of the phenomenon itself involves modelling aspects such as precipitating, inhibiting, or modulating conditions. By specifying the probabilistic structure of the environment, this is precisely what Bayesian rational analysis does. It specifies in mathematical detail the computational problem a cognitive system is confronted with and provides a normative solution, although it does nothing in terms of addressing how this solution is achieved. Reijula (2017), who is not as pessimistic about the explanatory potential of rational analysis as the quotation above may indicate, argues that they thus contribute to specifying *environmental affordances*, stating that "well conducted modeling of environmental affordances can complement mechanistic theorizing by providing resources for understanding the possible space of behavior of agents" (Reijula, 2017, p. 2). This view is well in line with Craver's account and the explanatory potential merely phenomenological models can have under it. Despite insisting that rational analysis is a fully-fletched, independent research program in its own right, Anderson himself makes some concessions in this regard when he states that

> [t]he rational approach helps rather directly with the induction problem. If we know that behavior is optimized to the structure of the environment and we also know what that

> optimal relationship is, then a constraint on mental mechanisms is that they must implement that optimal relationship. This helps suggest mechanisms. (Anderson, 1991, p. 471)

How exactly does rational analysis achieve this? Among the aspects relevant for specifying the explanans phenomenon are, following Craver (2007), counterfactuals, i.e., considerations about how the phenomenon would differ given different, counterfactual input conditions (following the metaphor of a cognitive process as an input-output mapping device). Reijula (2017) argues that this is exactly what rational analysis models are able to provide. He refers to these counterfactual cognitive system inputs as "environment-behavior what-ifs" (Reijula, 2017, p.4) and argues that they are capable of answering questions of the type

> *"How would the behavior of the cognizer change when the cognitive task changes in some particular way?"* That is, the model could uncover objective dependencies between properties of the environment and the behavior of cognizers. (Reijula, 2017, p. 4)

In this context, "objective dependencies" should be understood as "charting the possible cognitive space of action for an organism" (Reijula, 2017, p. 6). The type of work `left to be done by such an analysis is, following a mechanistic understanding of cognitive explanation, to identify the actual mechanism a cognitive system employs to navigate this cognitive space of action. Thus, the metaphorical meat and potatoes of explaining a given cognitive phenomenon is, although supported and scaffolded by rational analysis, largely left untouched by it.

Enlightenment versus fundamentalism?

The foregoing discussion concluded that, by themselves, both major theoretical frameworks in Bayesian cognitive science – predictive processing and Bayesian Rational Analysis – suffer from similar shortcomings, though to different degrees. Whereas both lack a full specification of constituent mechanisms, predictive processing traces the outlines of such a mechanism to a higher degree of detail, but lacks an independent verification of its proposed constituents, whereas Bayesian rational analysis merely provides phenomenological models capturing the phenomena under consideration with mathematical precision, but without elucidating how they are brought about. This evaluation, by itself, is not the full story. Beyond differing degrees to which Bayesian models actually specify mechanical details, the literature on the matter contains vastly differing assessments of the necessity of providing such a mechanistic account, i.e., about the completeness and explanatory value of probabilistic models.

In their critical review of the state of the field, Jones and Love (2011) identify two contradictory trends therein. They refer to these trends as “Bayesian enlightenment” and “Bayesian fundamentalism”, respectively. Bayesian fundamentalism is content with modelling cognitive phenomena and assessing their normative adequacy. It thus makes no mechanistic commitments, insisting rather that showcasing the optimal niche adaptability of a cognitive system is explanatory in itself. This claim was already assessed in detail in the foregoing discussion.

Jones and Love contrast this approach with what they refer to as “Bayesian enlightenment”. Whereas Bayesian fundamentalism is content with prediction and mathematically adequate modelling, Bayesian enlightenment seeks to uncover the mechanism bringing forth the type of Bayes-optimal behavior captured by Bayesian modelling. Researchers working within this framework consider cognitive modelling not merely a mathematical convenience, but rather as making commitments to a certain structure of the mind, and investigating which entities and processes give rise to it.

> A critical distinction between Bayesian Fundamentalism and Bayesian Enlightenment is that the latter considers the elements of a Bayesian model as claims regarding psychological process and representation, rather than mathematical conveniences made by the modeler for the purpose of deriving computational-level predictions. (Jones & Love, 2011, p. 168)

Based solely on the nomenclature used for characterizing both sides of this debate, one should be able to infer where Jones and Love’s sympathies lie. And indeed, the mechanistic commitments made by “Bayesian enlightenment” are necessary, as has been tried to argue in the preceding sections, for providing genuine explanatory value. As has been demonstrated, such mechanistic considerations are lacking in many Bayesian models of cognitive capacities.

Despite failing to provide explanatory power, a further problem with omitting mechanical considerations is that post-hoc model fitting can be camouflaged as genuine insight into a phenomenon. Given known results about human behavior, it is possible to construct Bayesian models fitting these patterns using assumptions about prior distributions not as independently researched claims about human psychology, but merely as mathematical tools for arriving at the desired results. Even if the physical implementation of a cognitive entity (such as priors in a Bayesian belief update) by concrete neural structures is unknown, mechanical considerations such as those given by Craver (2006, 2007) can guide and inform model construction invoking

such entities. Howhy (2013), a prominent proponent of predictive processing, sums up this idea as follows:

> The challenge […] is to avoid just-so stories. That requires avoiding priors and likelihoods that are posited only in order to make them fit an observed phenomenon. To avoid just-so stories any particular ordering of priors and likelihoods should be supported by independent evidence, which would suggest that this ordering holds across domains. (Hohwy 2013, p. 94)

It is worth bearing in mind in this context that, following Craver's approach, functional modules proposed in purely cognitive models are not alternatives to mechanistic descriptions of causal interactions, but rather abstract stand-ins for the latter, and as such liable to the same types of consideration. Providing a mechanistic explanation of a cognitive phenomenon need not entail specifying in detail how biochemical reactions at the cellular level bring forth a phenomenon. Craver's mechanism allows for mechanistic models with various degrees of abstraction, as long as genuine mechanical considerations are applied in proposing them. This means, as Howhy proposes, avoiding just-so stories and how-possibly explanations, and instead trying to ensure the robustness and independently verified validity of postulated model components.

This proposal is not merely a philosopher's fiction. There have been serious attempts to provide mechanistic interpretations of what are often considered purely instrumental aspects of Bayesian cognitive models, namely the mathematical parameters and inferential algorithms used for devising and updating them. Recall from section one the Monte Carlo sampling approach to Bayesian inference, the most popular of which being the Markov Chain Monte Carlo algorithm (or rather algorithms, as there are various different kinds, cf. Betancourt, 2017). To reiterate, this group of algorithms aims to make Bayesian model inference less computationally costly by approximating the posterior model by drawing from it a series of samples dependent upon each other via Markov chains (cf. Betancourt, 2017).

Purely instrumentalist approaches to Bayesian cognitive modelling (such as the afore discussed approach of rational analysis) do not give any cognitive interpretation to these algorithmic structures. They merely serve as a handy shortcut for computational inference which makes models computationally tractable.

However, there are also researchers who take the cognitive interpretation of Monte Carlo sampling serious (Griffiths et al., 2012; Shi et al., 2010). They propose that Monte Carlo sampling is a good approximation of the actual inferential computations carried out by the brain.

Whether this empirical conjecture is accurate remains to be established by research in cognitive science and neuroscience. Independent of whether this assumption turns out to be fruitful in accounting for observations in future research, it is an enticing proposal for the field of Bayesian cognitive science, for it takes a step in the direction of adding to its abundance of mechanism sketches and phenomenological models of different cognitive capacities by beginning to fill in the explanatory gaps these models leave open. Let us illustrate one of these attempts.

Shi et al. (2010) propose that importance sampling, a form of Monte Carlo inference, is performed in the mind by so-called exemplar models, which

> assume that people store many instances (exemplars) of events in memory and evaluate new events by activating stored exemplars that are similar to those events. […] [These models] can be interpreted as a sophisticated form of Monte Carlo approximation known as importance sampling. This result illustrates how at least some cases of Bayesian inference can be performed using a simple mechanism that is a common part of psychological process models. (Shi et al., 2010 p. 444)

Without delving into the mathematical details, importance sampling is a type of Monte Carlo inference in which each sample is assigned a weight based on its likelihood under a target distribution (Shi et al., 2010). This weight determines the contribution of the sample to the estimation. The core idea of understanding Bayesian inference via exemplar models is that stored exemplars are viewed as being sampled from a distribution of possible exemplars the system may encounter (Shi et al., 2010). To reiterate from previous sections, the Bayesian conception of the mind understands cognition primarily as inference under uncertainty. Thus, when perceiving a new stimulus, one is tasked with making inferences about the stimulus based on noisy or incomplete sensory data. The exemplar model conception of Bayesian inference maintains that when a stimulus is observed, its similarity to stored exemplars is evaluated. The similarity measure between stimulus and exemplar reflects how likely the (noisy) sensory stimulus would be if it was caused by the exemplar, thus acting as a likelihood function in Bayesian terms. The contribution each stored exemplar makes to the final inference is weighted based on how similar it is to the data, with higher similarity resulting in a greater weight. This is, so the idea, the relevant parallel to importance sampling, in which each sample is weighted based on its relevance to the estimation at hand. In the illustration just given, the example chosen was that of perceiving an ambiguous stimulus. Shi et al. (2010), however, maintain that similar frameworks can be given for other types of cognition, i.e., that the exemplar modelling framework of Bayesian cognition is not limited to perception. More specifically, they

> expect that importance sampling with a relatively small number of samples drawn from the prior should produce an accurate approximation of Bayesian inference in cases in which prior and posterior share a reasonable amount of probability mass. This can occur in cases in which the data are relatively uninformative, as a result of either small samples or high levels of noise. Despite this constraint, we anticipate that there will be a variety of applications in which exemplar models provide a good enough approximation to Bayesian inference to account for existing behavioral data. (Shi et al., 2010, p. 448)

This summary skipped over the mathematical details of both exemplar modelling and importance sampling to focus on the link between two approaches proposed by Shi et al. (2010). The core idea of their approach is, to summarize, to ground the purely computational (i.e., lacking mechanistic implementation details) task of performing Bayesian inference in an established framework of psychological theory, namely exemplar modelling (Shi et al., 2010).

Although this account is still but a schema of a fully-fledged explanatory model under the mechanistic framework proposed by Craver, it is a step in the right direction, and – assuming this account of cognition is correct – an illustrative case study of mechanistic modelling in the field of Bayesian cognitive science: The modelling process starts with a mathematically specified phenomenological model of the phenomenon in question, in this case various forms of perception and cognition. This model specifies in detail the explanans phenomenon, as required by Craver's (2006, 2007) framework. This input-output model is then expanded by specifying a schema of how the computation in question is performed, namely – so the proposal at hand for the particular case just considered – by exemplar sampling.

Although is example constitutes a step towards a fully fletched explanatory model, it is still lacking in that it fails to give any account as to how the exact interaction and organization of components gives rise to the phenomenon at hand. Although exemplar sampling can be seen – to a certain extent – as a specifying the interaction of components, this interaction is characterized at a very abstract level, and a far cry from the precise spatio-temporal characterization set by Craver as the ideal of mechanistic modelling.

There are, however, attempts to ground Bayesian aspects of cognition in more fine-grained components, namely neuronal activity patterns (or at least abstract models thereof). Karratzadeh and Schulz (2016), for example, propose a framework of how Bayesian computation can be performed by a population of deterministic neurons, which consists in an artificial neural network model of Bayesian inference. Explicitly motivated by the lack of mechanistic grounding of many Bayesian models, they aim to provide "a neurally–based model of

probabilistic learning and inference that can be used to simulate and explain a variety of psychological phenomena" (Karratzadeh and Schulz, 2016, p. 100). The aim of this model, so the authors, is to show that "randomness and probabilistic representations can emerge as a property of a population of deterministic units rather than a built-in property of individual stochastic units" (Karratzadeh and Schulz, 2016, p. 100). A similar, albeit more biology-oriented, approach is taken by Köver and Bao (2010), who propose that the number of neurons devoted to representing a prior hypothesis in a Bayesian model indicates the strength of the prior.

From a Craverian perspective, what is intriguing about these approaches – should they prove successful – is that they open the door for a fully fletched mechanistic account of how individual components – deterministic neurons – can give rise to a phenomenon – Bayesian modes of cognition – via their interactions and organizations, i.e., their activation patterns. By bridging the gap between neuronal firing patterns and Bayesian high level inferential processes, these authors and their peers engaged in similar work may have made significant contributions to the regulatory ideal of an ideally complete mechanistic model in the field of Bayesian cognitive science.

Once again, it remains to be determined whether or not these or similar accounts prove empirically adequate and fruitful for empirical science and whether they prove compatible with higher level accounts such as the psychological account by Shi et al. (2010). These are questions beyond the scope of the current thesis. However, even if the summarized accounts turn out to be a dead end, they display exactly the type of theorizing required (and often lacking) in the field of Bayesian cognitive science. If cognitive scientists are able to ground their hitherto mostly abstract or purely phenomenological models in lower-level processes, they add explanatory power to their account that is presently, at least in many instances, missing. Thus, the mechanistic stance does not merely serve as a tool for evaluating existing models, it is able to identify explanatory gaps in them, as well as giving a blueprint for what kind of theorizing is required to fill them.

Let us summarize the main conclusions of the present section. Different approaches within the field of Bayesian cognitive science take different stances on the explanatory value of purely abstract modelling of psychological capacities and the necessity of mechanistic considerations in this field. This difference is best exemplified by the contrast between the predictive processing and the rational analysis framework. The former attempts to characterize the computations carried out by the mind as well as a sketch of the mechanistic structure giving

rise to them. As such, should the framework be able to prove its empirical adequacy, it can be considered as genuinely explanatory, for it provides mechanism sketches of cognition. Its main weakness, from a mechanistic perspective, is that it lacks independently manipulatable components, proposing an entire hierarchical structure at once and hence making it difficult to assess the mechanistic claims it makes. This is, so to speak, a technical problem, requiring ways of independently testing and manipulating the proposed model constituents. Solving this problem would not be at odds with the core idea or methodology of predictive processing but rather add to it.

In contrast, the rational analysis approach often fails to take mechanistic aspects of explanation into consideration at all. This is not merely a technical problem. It is a more fundamental problem with the core methodology of rational analysis. As the mere justification of a behavior or type of cognition is not by itself explanatory (as was argued), rational analysis often fails to provide a genuine explanation of its target phenomenon, providing instead little more than a phenomenological model (leaving, for now, aside the etiological aspects it can shed light on).

A possible solution to this lack of mechanistic grounding was discussed in the form of research trying to explicate how the interactions of either abstract psychological or lower-level neuronal constituents may fill the black box of such purely phenomenological models. Such approaches, if taken seriously and expanded upon in the field of Bayesian cognitive science, are capable of solving the problems just laid out (i.e., the lack of mechanistic grounding many models have), and thus expanding the explanatory potential of models in the field. In the words of Jones and Love (2011), this would entail going down the path of Bayesian enlightenment rather than Bayesian fundamentalism.

This concludes our discussion of common Bayesian modelling practices and their evaluation from a philosophy of science perspective. The reader may at this point be satisfied with the account that was given or feel the need to make some strong objections to is. The ensuing and final section aims to address some of these potential objections.

Objections

The main line of reasoning pursued in the present thesis was the following: Models in cognitive science can provide explanation of a given phenomenon insofar as they are capable of elucidating the causal mechanism underlying it. This identification of a causal mechanism is often missing or developed only in very rudimentary outlines in many Bayesian approaches to cognition. Such models can have some explanatory value, but to provide genuine explanations, it is necessary to supplement them with providing mechanistic details of how component interaction and organization enable a system to carry out the computations specified in purely abstract Bayesian models. Finally, some attempts to achieve this mechanistic underpinning of Bayesian cognitive models were discussed.

Explanatory power of mathematical modelling

One could object to the proposed position by pointing out that the account I have given hinges on the acceptance of mechanism, and that this acceptance is not ubiquitous. It is sometimes argued, e.g. by proponents of dynamical systems approaches in cognitive science, that mathematical models of a phenomenon are by themselves sufficient for explaining the phenomenon due to their explanatory and unifying power (Chemero & Silberstein, 2008; Chemero, 2009; cf. Gervais, 2015). Proponents of this position can invoke a philosophical tradition going back to Bertrand Russell (1912), who proposed, based on the observation that physics does not employ notions of causal explanation but rather mathematical modelling, that ideally, science should do without notions of causality and focus solely on providing mathematical descriptions of its phenomena.

It is true that the presented account critically hinges on the acceptance of mechanism. It is worth noting, however, that merely insisting on the explanatory power of mathematically modelling a phenomenon does not provide an alternative account of what it means to achieve scientific explanation. In other terms, it is unclear by the power of what property a mathematical model in cognitive science would be able to explain cognitive phenomena, if not through mechanistic considerations.

Proponents of the notion that mathematical modelling is itself explanatory sometimes invoke their predictive power. Port and Van Gelder (1995), influential proponents of dynamical systems theory in cognitive science, propose that the mathematical modelling provided by the latter "yields not only precise descriptions […] but also predictions which can be used in evaluating the model" (Port & Van Gelder, 1995, p. 5), thus lending it explanatory power. Chemero and

Silberstein (2008) make explicit this alleged conceptual connection between explanation and prediction when they state that "[i]f models are accurate enough to describe observed phenomena and to predict what would have happened had circumstances been different, they are sufficient as explanations" (Chemero and Silberstein, 2008, p. 12). Similar arguments can ben proposed in regard to the unificatory potential of Bayesian approaches to the mind and the supposed explanatory power it provides (cf. Colombo & Hartmann, 2017; Griffiths et al., 2010).

While both prediction and theoretical unification are worthwhile desiderata for a scientific theory, they are, as was argued in more detail in section 3, conceptually distinct from and independent of explanation. Merely alluding to the predictive or unificatory potential of a given model is therefore not sufficient for establishing its explanatory power. By the power of what property, then, can mathematical models with no appeal to any mechanistic considerations achieve explanation? It remains unclear. Mechanism, on the other hand, provides a clear normative guideline for evaluating mathematical models of a phenomena.

In an essay building upon previous accounts establishing a detailed mechanistic framework of scientific explanation, Kaplan and Craver (2011) propose what they term the "model-to-mechanism-mapping requirement" (henceforth 3M). 3M states that a mathematical model is explanatory insofar as the individual components of the mathematical structure used to describing a phenomenon map onto components of the mechanism underlying it. This account, if applied to Bayesian modelling in cognitive science, fits with the approach described as "Bayesian enlightenment" By Jones and Love (2011), which was discussed earlier. This account, to reiterate, maintained that Bayesian modelers ought to the embrace the ontological commitments their models make, thus understanding them as sketches of mechanistic explanations rather than mere predictive instruments. Thus, mechanism can account for the explanatory force of mathematical models, whereas the supposed explanatory power of such models without a mechanistic background remains unfounded.

Furthermore, the ideal of emulating physics, as proposed by Russell (1912), in all types of scientific theorizing ought to be questioned. Physics deals with nature at a fundamental level which operates in accordance with – or so at least the received view in philosophical and scientific tradition – mathematically specifiable, fundamental laws. Trying to explain the fundamental entities of physics by appealing to a lower constitutive level from which its entities emerge is thus a futile endeavor, as physics deals with an aspect of nature that is in some aspect itself fundamental. What is left is thus to capture the properties and relations holding at this level using the language of mathematics. Cognitive science, however, is not physics. The

capacities of a cognitive system are brought forth by the complex interactions and organization of simpler constituents, and cognitive science aims to uncover how this emergence (in a non-theory-laden sense) occurs. This is the subject matter the mechanistic account is suited to deal with, and for which it can provide a notion of scientific explanation which is in line with common intuitions about scientific explanation (as showcased using various examples from the philosophical literature in section 3) as well as scientific practice. As argued before, given the vastly heterogenous subject matters and methods used in different scientific disciplines, there seems to be no a priori reason to assume that the same normative guidelines should apply to all of them.

Reductionism and Multiple Realization

A further objection one may raise is the seemingly reductionist nature of the proposed position. This need not constitute a reason for giving up the position per se, as reductionism is a defensible philosophical stance. However, it is a stance that has somewhat gone out of style due to certain philosophical problems it has, most prominently that of multiple realization, i.e., the thesis that mental states are multiply realized, and thus cannot be identified with (i.e., reduced to) any neuroscientific kind (Polger & Shapiro, 2016; Putnam, 1967). To what degree multiple realization is merely a logical possibility as opposed to a genuine fact about minds and their neural substrate partially depends on the conception of multiple realization one is using. For example, Aizawa and Gillett (2009) propose that minute differences at the neurochemical substrate of a mental state suffice for speaking of the multiple realization of said state. Bechtel and Mundale (1999) object that such examples are little more than artefacts of the way mental concepts and their physical substrates are described in the philosophical literature. They argue that alleged examples of multiple realization often involve very coarse-grained mental concepts, such as simply “learning”, as opposed to a more fine-grained alternative like “implicit associative priming”, a specific subtype of learning. These coarse-grained concepts are claimed to be multiply realized by differences in fine-grained details at the substrate level, such as small biochemical differences in an otherwise identical process, and a properly fine-grained conception of the mental kinds in question would, claim Bechtel and Mundale, lead to these alleged cases of multiple realization disappearing. Their argument casts serious doubt on the often uncritically accepted notion of multiple perception and makes an alternative, reductionist approach seem more palatable than it is often viewed to be in the philosophy of mind (cf. Bickle, 2003 for a more detailed defense of this stance). However, for our present purposes, this is not the argumentative route I will pursue. Although elsewhere I have endorsed Bechtel and Mundale’s (2009) view on this matter (Mattersberger, forthcoming), the current discussion does

not hinge upon the specific position taken regarding the multiple realization of mental kinds. Why is that?

I shall argue that the account I have endorsed is either reductionist in a weak sense, for which the multiple realization thesis is not problematic, or not reductionist (in the strong sense) at all. In other terms, an account of reductionism capable of incorporating the view I have presented would need to give up so many of the claims traditionally associated with reductionism that multiple realization no longer poses a problem for it. What is this traditional view of reductionism, and how does it differ from the mechanistic account I have endorsed?

Reduction in the traditional sense occurs when terms and the logical relation in which they stand to each other of higher-level theories (e.g. a theory in psychology) are identified with terms of lower level theories (such as physics or biology), so that the axioms and laws of the higher level theory can be logically deduced from the axioms and laws of the lower level theory using so called bridge laws (Nagel, 1961; van Riel & Van Gulick, 2024). Bridge laws specify the relationships holding between the terms and statements of the higher and the lower-level theory (Nagel, 1961; van Riel & Van Gulick, 2024). This type of reduction is proposed as model of scientific unification by Oppenheim and Putnam (1958).

The core assumption of this type of reductionism is that the natural world can be classified into hierarchical strata with a clear order, from highest to lowest level (van Riel & Van Gulick, 2024). Different scientific disciplines, following this view, investigate the natural world at different levels. Reduction is achieved by establishing a logical relationship between these theories at different levels so that the higher-level theory is deducible from the lower-level theory (Oppenheim and Putnam, 1958; van Riel & Van Gulick, 2024). Thus, reduction in the traditional sense is an inter-theory relationship, with both of the involved theories describing (an aspect of) a different stratum of the natural world (van Riel & Van Gulick, 2024).

This type of traditional reductionism shall be termed “strong reductionism” for the purposes of the present discussion. Is strong reductionism a fitting characterization for the view I have endorsed? I argue that it is not. Following Craver (2005), I propose that the mechanistic program offers an alternative view on interdisciplinary integration and unification which significantly differs from strong reductionism How so?

The key differences, so I shall argue, are first, that reductionism, other than mechanism, is concerned with “eliminating” higher level theories by showing how they are merely logical consequences of lower level theories, second, that reductionism assumes that scientific fields

neatly correspond to distinct strata of the natural world, which mechanism does not, and third, that reductionism is concerned with formal derivability of higher-level theories from lower level theories, which is not the case for mechanism.

Regarding the first of these differences, whereas reductionism is focused on "looking down", i.e., reducing a higher-level theory to a lower level one, the mechanistic approach to cognitive and neuroscience must, as in the aptly titled paper by William Bechtel, "[look] down, around, and up", (Bechtel, 2009, p. 544). They must focus on integrating mechanism constraints from different fields and different levels, with no clear hierarchical direction from the higher to the lower, as reductionism proposes. This is directly connected to the second difference mentioned above. In the different disciplines studying the mind and brain, there is often no clear hierarchical relationship between different fields and theories. Researchers may be concerned with problems and topics as heterogenous as the structure of mental representation, temporal aspects of information processing, the correlates of activity in certain brain regions, the connectivity patterns between different neuronal populations, the interactions and inhibitions between different processes, the amplitudes of EEG activity at different frequencies averaged over the whole brain, and many more. There is no clear stratified relationship between these various research fields and models proposed in them, and thinking about them as a clear hierarchy from top to bottom with intermediary bridge laws is not helpful, nor is it an accurate representation of scientific practice. The mechanistic view proposes that these various fields make different constraints on mechanism identification, some of them temporal, some of them spatial, some of them in terms of organization or interaction, some of them in terms of clearly delineating the explanandum phenomenon. The goal is not to reduce these different fields to a common lower level, but rather to integrate them by taking into consideration the different mechanistic constraints they pose. Craver (2005) gives an example of this approach in the research on long term spatial memory in mice, which integrates constraints and findings in different disciplines and at different levels for theorizing about a unitary underlying mechanism:

> [R]esearchers deleted the gene for the NMDA receptor [in mice] and tested for effects at a number of different levels. They found that mice without the NMDA receptor lost LTP induction, had distorted spatial maps, and had difficulty solving mazes. Experiments of this sort draw on molecular biology (for the deletion), electrophysiology (to assay for LTP), higher-level electrophysiology and computation (for detecting and modeling the spatial map), and psychology (to monitor LM performance), piecing their findings into a single multilevel theory. (Craver, 2005, p. 392)

This also relates to the last difference between mechanistic integration and reduction: just as mechanism does not hinge on the assumption of a clear hierarchy of strata, it does not hinge on the assumption that some theories are formally derivable from lower-level theories. The mechanistic view takes the contributions of different field studying the same phenomenon as constraints on the mechanism underlying the phenomenon. These different fields can inform our understanding of the phenomenon without the need for intertheoretical reduction or derivability.

For the present discussion, this means that the theoretical vocabulary describing higher level, abstract characterizations of Bayesian computation supposedly carried out in the mind need not be defined in terms of the vocabulary of a lower-level theory (e.g. that of neuroscience), nor is there a need to postulate bridge laws between the axioms of different theories. Rather, the goal is to identify more clearly the mechanisms at play in producing the phenomenon under consideration. This may seem like a very minor distinction but is has far reaching philosophical consequences, which brings us back to the question of multiple realization. To reiterate a classical, perhaps *the* classical, argument against reductionism is the impossibility of identifying higher level theories with lower-level theories due to the formers' multiple realizability. A given abstract cognitive architecture cannot be identified with a lower-level theory, e.g. of neural activity, because such an abstract architecture is multiply realizable. The mechanistic view renders this critique irrelevant. Although an abstract cognitive architecture may be multiply realizable, one could still identify and describe the mechanism involved in realizing it in a particular type of cognizers, such as humans. One could explore how deviations in the mechanism relate to differences in the architecture, or how variance in the mechanism can nonetheless give rise to very similar cognitive profiles. One need not be worried about inter-theory derivability or formally defining one's high-level vocabulary in terms of lower-level vocabulary, because such considerations of formal reduction and derivability are no concern under the mechanistic framework. Thus, when, for example, specifying the mechanism underlying human learning, one need not be concerned with the objection that the concept "learning" cannot be identified with the description of that mechanism, as such an identification was never the goal to begin with. The goal was to describe (a prototype of) the causal mechanism giving rise to learning in the human brain. This can then lead to interesting follow-up research, such as comparing the mechanisms of human learning with that of learning in different species or artificial forms of intelligence, in order to identify commonalities or differences. This could in turn lead to interesting empirical results, such as that a very similar mechanisms underlies learning across these different domains, or that the underlying

mechanisms are vastly different. The latter result, however, would not then lead the researchers arriving at it to the conclusion that they never understood learning in the first place because their inter-theoretical reduction failed.

Critics may argue, despite the foregoing arguments, that the proposed view is still a form of reductionism. Although not committed to the clearly stratified nature of scientific disciplines or the goal of defining higher level concepts in terms of lower-level concepts, it still rejects the independence of the special sciences in theorizing, calling for inter-field integration of the cognitive and neural sciences instead.

I gladly concede that the proposed view is reductionist in this weak sense, but do not deem it problematic. The main argument in favor of the independence of the special sciences is that of multiple realizability and therefore the irreducibility (in the strong sense) of theories in the special sciences (Fodor, 1974). As I have argued, multiple realization is not a concern for the mechanistic view endorsed here, hence making the type of weak mechanistic reductionism (if one is inclined to call it that) I have endorsed unproblematic.

To summarize, strong reductionism differs in various core assumptions from the mechanistic view defended in the present thesis. If the latter can nevertheless be said to be reductionist, then it must be a weak form of reductionism which is not affected by the problem of multiple realization, for it does not wish to identify higher level concepts with lower-level ones, nor is it concerned with formal derivability of higher-level theories. Craver (2005, p. 393) aptly summarizes the key differences between mechanism and strong reductionism as follows:

> Findings at different levels constrain the links between levels accommodatively, spatially, temporally, and experimentally. Fields are integrated across levels when different fields identify constraints on mechanisms at different levels. This gradual process of accumulating constraints on mechanisms at multiple levels in no way resembles efforts to translate one theory to another or to create a homomorphic image of one in terms of another. Instead, different fields elaborate the multilevel mechanistic scaffold with a patchwork of constraints on its organization, thereby revealing different hints as to how the mechanism can and cannot be organized.

Conclusion

In the introductory remarks, it was stated that the present thesis aims to pursue three distinct goals. The first of these goals was to introduce and give an overview of the field of Bayesian cognitive science. To this end, it was first elucidated how Bayesian inference operates in general, independent of any concrete domains of scientific application. Barring the mathematical details, this type of inference integrates a new piece of information into the assessment of a hypothesis' probability, arriving at the posterior probability by combining the prior probability of the hypothesis with the likelihood of the new evidence given the hypothesis in question. In doing so, it strikes a compromise between prior assumptions and newly available data. It was then elucidated how Bayesian inference typically assesses various different hypotheses at once, with the totality of hypotheses under consideration being termed the "hypothesis space", which can either be discrete or continuous. Last, we saw how approximating algorithms, such as different Monte Carlo methods, are often used to circumvent the possible computational intractability of assessing complex hypothesis spaces.

Having established this fundamental understanding of Bayesian inference, it was then explained how this type of inference is employed as a model of cognition in contemporary cognitive science, the general idea being that cognizers need to make inferences about uncertain external stimuli and achieve this by employing a process of probabilistic inference. Cognition, in this framework, is thought of as an integration of the information provided by uncertain stimuli into a prior world model, which correspond to the likelihood and the prior in Bayesian terms.

The research conducted in this framework was, for our purposes, divided into two general categories, with the caveat that these categories are not to be considered as entirely mutually exclusive and jointly exhaustive, but rather as two broad and generalized directions of research within the field. These categories were termed the 'normative' and the 'descriptive' Bayesian approach. The former of these, normative Bayesianism, operates by first specifying, often in mathematical detail, the inferential problem a cognitive system is confronted with in a certain context. It then provides a computational solution to this problem, which is devised to adhere to normative standards of optimality. These normative standards are provided by the mathematical framework of Bayesian statistics. Only in a subsequent step is this normative solution then compared to the observed performance of the cognitive system in question. The most prominent example of this broad category of research is Anderson's (1990, 1991, 2013, 2014) method of rational analysis, which follows a distinct schema of model development in order to pursue the theoretical goal just laid out.

The normative approach to Bayesian cognitive science was contrasted with what was termed the descriptive approach. The latter does not start with a description of the computational problem faced by a cognizer and devising an optimal solution to it, but rather attempts to show directly how the cognitive system in question can be understood as Bayesian in some regard. Thus, its starting point is not devising an optimal solution to an inferential problem, but rather constructing a theory of how a cognitive system's architecture could implement (approximately) Bayesian processes. The most influential representative of the descriptive category that was discussed, and possibly the most influential Bayesian framework of cognition in general, is the predictive processing approach. The core tenet of this framework is that cognition fundamentally operates via a hierarchically structured network of predictive levels, each of which try to predict the state the next lower node in the network is in. The deviance between the prediction and the actual state (the so-called 'prediction error') is reported back up the hierarchy and used to revise the model employed in generating the initial predictions. This process of generating predictions based on an internal world model and updating said model based on the observed empirical data (the prediction error) is Bayesian at its core, and explicitly devised as such in the predictive processing framework.

The second aim of the present paper established in the introductory remarks was to propose a philosophical framework of explanation in cognitive science. To this end, first we considered the covering-law model and the U-model of explanation. It was attempted to show that both accounts, taken by themselves, fail to establish adequate criteria for scientific explanation as they are unable to account for certain intuitively relevant features of explanation, for example the asymmetric nature of explanation. The latter denotes the intuitively appealing idea that the explanans part of an explanation explains the explanandum, but the inverse does not hold.

The mechanistic framework of scientific explanation was then introduced as a viable alternative to the previously discussed accounts. The core idea of the latter is that explanation aims to lay bare what Salmon (1984) termed the "causal structure of the world". In other words, it aims to uncover the causal interaction of physical constituents giving rise to a phenomenon. This general approach, so the defended claim, is suitable for characterizing explanation in cognitive science, though unlikely to be applicable to the entirety of scientific pursuits due to their methodological heterogeneity. The particular mechanistic framework then introduced and further discussed was that of Craver (2006, 2007), which was originally proposed by its author specifically for the field of neuroscience, but is equally applicable, it was argued, to the related discipline of cognitive science. Craver distinguishes etiological mechanistic explanation, which specifies the causal history leading to a phenomenon, and constitutive mechanistic explanation,

which specifies how the causal interaction and organization of constituent components give rise to a phenomenon, with his focus being on the latter kind, as it lends itself more handily to neuroscience (as well as, as was argued, cognitive science). Such constitutive explanations can exist at various levels of abstraction and completeness, sitting on a spectrum ranging from ideally complete mechanism descriptions, which specify every causally relevant aspect of a mechanism, and merely phenomenological models, which do not specify any mechanistic details but merely a mechanism's input-output mapping. These models are to be assessed, according to Craver, by the extent to which they are able to specify the phenomenon in question, its constituent parts, the parts' activities and interactions, and their organization, with the caveat that models should identify real, causally active entities as opposed to giving mere how-possibly stories about how a phenomenon could be brought about by fictional entities.

It was then argued that to the extent to which they are explanatory, alternative models of explanation in cognitive science - rational explanation, functional explanation, and the tripartite model of explanation - adhere to the standards set by this framework. These alternative frameworks can thus be considered special cases of mechanistic explanation, typically mechanism sketches, in which concrete physical entities are replaced by functionally defined abstract components. However, the latter still fulfill the role of causal explanatory entities, as they, like all mechanisms, narrow down the range of plausible mechanisms. The commitment to mechanism, however, entails that scientific explanation should not end with these prima facie models of target phenomena, but instead aim to fill in the blanks left by invoking merely abstractly defined causal entities, a commitment which does not necessarily follow from the alternative frameworks taken by themselves.

Equipped with this philosophical toolkit, we then proceeded with the third aim of the present thesis, namely, to evaluate the present state of Bayesian cognitive science from a philosophy of science perspective. To reiterate, the goal was not to evaluate the scientific merit of Bayesianism as a paradigm in cognitive science, i.e., how well it can account for empirical phenomena, how predictive it is of new phenomena, how well it fits into our general scientific worldview, etc., but merely the more modest task of applying the standards given by mechanistic philosophy of science to the types of models and theories (again, there is no philosophical commitment attached to these terms in the way I use them) proposed in this field. In other terms, our question was, given that Bayesian models fulfill other scientific assessment criteria, how well are they able to explain their target phenomena due to their explanatory structure?

In this context, we differentiated between the two categories of Bayesian cognitive science identified earlier, descriptive and normative Bayesianism. Whereas descriptive Bayesianism, most prominently the predictive processing Framework, is typically explicitly guided by mechanistic concerns, trying to identify the organizational structure underlying cognition, normative Bayesianism is often content with demonstrating that human behavior can be reconstructed using Bayesian inference and is hence justified in respect to this normative framework. This normative assessment, it is argued, has explanatory character in and by itself. While these observations are generalized and certainly not applicable to each case of Bayesian theorizing in cognitive science, the very distinction between normative and descriptive approaches being tentative, it does highlight some general trends and trajectories within the field.

The predictive processing framework (if empirically adequate over various contexts, which, to reiterate, is a claim which to assess goes beyond the scope of our current investigation) narrows the space of possible mechanisms by postulating an abstract structure of the realizing mechanism of cognition. It specifies its temporal organization and the nature of the interaction between components. Although it lacks many specific details required for speaking of fully fletched mechanistic explanation (in the general framework, little is said about the concrete physical entities fulfilling the only functionally characterized component roles and spatial characteristics are, if at all, only touched upon very superficially), the predictive processing framework aims to give genuine mechanistic insights. Its current state is best described as devising a mechanism sketch (or mechanism sketches), the details of which ought to be filled in by subsequent research, a task which is taken seriously and pursued by researchers in the field.

A somewhat contradictory trend can be identified in the normative framework, wherein normative justification and computational modelling without consideration of mechanistic implementation is often taken to be explanatory in itself. However, whereas the mechanistic framework provides a detailed account of what it is to be explanatory (in many special sciences), namely, by specifying how a phenomenon is situated in the larger causal structure of the world, the merely normative approach lacks any detailed proposals about the nature of explanation or justifications of a purely normative evaluation's supposedly explanatory prowess. In other words, it remains unclear how normative justification can provide explanatory value.

It was argued, however, that such approaches still possess explanatory power, although to a very limited degree, in that they provide merely phenomenological models of a target phenomenon, thus specifying the phenomenon itself in detail and hence contributing to narrowing down the space of possible constitutive mechanisms. A detailed specification of the phenomenon under investigation, to reiterate, is one of the core criteria by which to assess a mechanistic explanation following Craver (2006, 2007).

Last, it was shown how some researchers make attempts to fill in the mechanistic details of purely computational models by taking serious a mechanistic interpretation of the algorithms used to implement Bayesian models, which are usually brushed aside as non-representational of any real life aspect of cognition. This approach, I argued, independent of whether it turns out to be empirically accurate and predictive on a larger scale, as well as approaches considering possible neural implementation of Bayesian cognitive architectures, are moving the field in the right direction, by going beyond a merely phenomenological model fit to considering how the computational frameworks in question could be implemented in the natural world.

In summary, the present thesis aimed to showcase how mechanism can provide a proper framework of what it means to explain a cognitive phenomenon, and then to assess to which extent this type of explanation is common in the field of cognitive science. The perspective thus endorsed, although the scope of our investigation was limited to cognitive science, is congruent with plausible assessments of explanatory power in other fields of the special sciences. Consider the explanatory power of Darwinian natural selection. Although fully specified in algorithmic terms, Darwin's theory gained traction as a genuine, detailed explanation only when the mechanistic details of its inner workings were made explicit through the study of modern genetics (cf. Dennett, 1995). This addition to the purely algorithmic framework Darwin provided did not change anything about the core proposals of the latter. It did, however, fill in the missing mechanistic pieces, thus transforming the idea from an – although powerful and plausible – mechanism sketch to a genuine, fully fletched mechanism model, and by doing so laid the groundwork for the entire field of modern biology (cf. Dennett, 1995).

This example aims to demonstrate that mechanistic implementation matters. It is not merely a philosopher's abstract ideal far removed from actual scientific practice. By specifying the details of a phenomenon, one can escape the mere hypothetical character of how-possibly models and gain genuine insights into how a phenomenon fits into the world at large, how it is causally situated, and how it fits together with the domains of other fields, thus allowing for robust, genuine explanation.

For scientific practice, this entails, contra Fodor (1974, 1997), that sciences like psychology and cognitive science (insofar as the latter aims to gain an understanding of natural cognition as opposed to devising artificial cognizers) ought not to operate independently and autonomously, but rather seek close integration with the biological and neurosciences. Such a close integration is able to flesh out the tentative insights of these more abstract fields, hence solidifying their insights by, to borrow Salmon's (1984) terminology one last time, demonstrate how cognition is situated in the "causal structure of the world".

References

Aizawa, K., & Gillett, C. (2009). The (multiple) realization of psychological and other properties in the sciences. *Mind & Language, 24*(2), 181-208.

Anderson, J. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum Associates.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Anderson, J. R. (2013). *The Adaptive Character of Thought*. Psychology Press.

Anderson, J. R. (2014). A rational analysis of human memory. In H. L. Roediger III & F. Craik (Eds.), *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving* (1st ed., pp. 195-210). Psychology Press.

Barnes, E. (1992). Explanatory unification and the problem of asymmetry. *Philosophy of Science, 59*(4), 558-571. https://doi.org/10.1086/289695

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology, 22*(5), 543-564.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science, 66*(2), 175-207.

Bellhouse, D. R. (2004). The Reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth. *Statistical Science, 19*(1), 3-43. https://doi.org/10.1214/088342304000000189

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Springer Science & Business Media.

Bromberger, S. (1965). An approach to explanation. In R. S. Butler (Ed.), *Analytical Philosophy: Second series* (pp. 72-105). Basil Blackwell.

Chater, N. (2014). Cognitive science as an interface between rational and mechanistic explanation. *Topics in Cognitive Science, 6*, 331-337. https://doi.org/10.1111/tops.12087

Chater, N., & Oaksford, M. (Eds.). (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford University Press.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. MIT Press.

Chemero, A., & Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science, 75*, 1-27.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204.

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science, 68*(2), 451-484. https://doi.org/10.1093/bjps/axw007https://doi.org/10.1093/bjps/axv036

Craver, C. F. (2005). Beyond reduction: Mechanisms, multifield integration and the unity of neuroscience. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 373-395.

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*(3), 355-376.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*.

Craver, C., & Tabery, J. (2023). Mechanisms in science. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Fall 2023 Edition)*. https://plato.stanford.edu/archives/fall2023/entries/science-mechanisms/

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy, 72*(20), 741-765. http://links.jstor.org/sici?sici=0022-362X%2819751120%2972%3A20%3C741%3AFA%3E2.0.CO%3B2-Q

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian Cognitive Science* (pp. 59-75).

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00102

DeGroot, M. H. (1975). *Probability and Statistics*. Addison-Wesley.

Dennett, D. C. (1993). *Consciousness Explained*. Penguin UK.

Dennett, D. C. (1995). Darwin Dangerous Idea. Simon & Schuster.

Faulkenberry, T. J., Ly, A., & Wagenmakers, E.-J. (2020). Bayesian inference in numerical cognition: A tutorial using JASP. *Journal of Numerical Cognition, 6*(2), 231-259. https://doi.org/10.5964/jnc.v6i2.288

Fink, S. B., & Zednik, C. (2017). Meeting in the dark room: Bayesian rational analysis and hierarchical predictive coding. In T. Metzinger (Ed.), *Philosophy and Predictive Processing* (pp. 226-238). MIND Group. https://doi.org/10.25358/openscience-637

Fodor, J. A. (1974). Special sciences (Or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97-115. https://doi.org/10.1007/BF00485230

Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Philosophical Perspectives, 11*, 149-163.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy, 71*(1), 5-19.

Gallagher, S., Hutto, D., & Hipólito, I. (2022). Predictive processing and some disillusions about illusions. *Review of Philosophy and Psychology, 13*(4), 999-1017.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41*(6), 711-724. https://doi.org/10.1016/S0042-6989(00)00277-7

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2020). *Bayesian Data Analysis* (3rd ed.). http://www.stat.columbia.edu/~gelman/book/

Gervais, R. (2015). Mechanistic and non-mechanistic varieties of dynamical models in cognitive science: Explanatory power, understanding, and the 'mere description' worry. *Synthese, 192*(1), 43-66. https://doi.org/10.1007/s11229-014-0548-5

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*(S3), 342-353. https://doi.org/10.1086/341857

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357-364.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*(4), 263-268. https://doi.org/10.1177/0963721412447619

Gull, S. F. (1988). Bayesian inductive inference and maximum entropy. In G. J. Erickson & C. R. Smith (Eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering: Foundations* (pp. 53-74). Springer Netherlands. https://doi.org/10.1007/978-94-009-3049-0_4

Hacking, I. (2006). *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press.

Hájek, A. (2023). Interpretations of probability. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*. https://plato.stanford.edu/archives/win2023/entries/probability-interpret/

Halonen, I., & Hintikka, J. (1999). Unification: It's magnificent but is it explanation? *Synthese, 120*(1), 27-47.

Hempel, C. G. (1968). Maximal specificity and lawlikeness in probabilistic explanation. *Philosophy of Science, 35*(2), 116-133.

Hempel, C. G. (2012). *Aspekte Wissenschaftlicher Erklärung* (Aspects of Scientific Explanation). De Gruyter. (Original work published 1976)

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135-175. https://doi.org/10.1086/286983

Hempel, C. G. (1962). Deductive-nomological vs. statistical explanation. In H. Feigl & G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science* (Vol. 3, pp. 98-169). University of Minnesota Press.

Hohwy, J. (2013). *The Predictive Mind*. OUP Oxford.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language, 35*(2), 209-223.

Hubel, D. H., & Wiesel, T. N. (1979). Brain mechanisms of vision. *Scientific American, 241*(3), 150-163. http://www.jstor.org/stable/24965293

Hyman, S. E., Malenka, R. C., & Nestler, E. J. (2006). Neural mechanisms of addiction: The role of reward-related learning and memory. *Annual Review of Neuroscience, 29*(1), 565-598.

JRank Science Encyclopedia. (n.d.). The Newtonian synthesis. In *Physics*. Retrieved June 5, 2024, from https://science.jrank.org/pages/10737/Physics-NEWTONIAN-SYNTHESIS.html

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*(4), 169-188.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron, 100*(2), 424-435. https://doi.org/10.1016/j.neuron.2018.10.003

Kharratzadeh, M., & Shultz, T. (2016). Neural implementation of probabilistic models of cognition. *Cognitive Systems Research, 40,* 99-113.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific Explanation* (pp. 410-505). University of Minnesota Press.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712-719.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Foundations of the Theory of Probability). Julius Springer.

Köver, H., & Bao, S. (2010). Cortical plasticity as a mechanism for storing Bayesian priors in sensory perception. *PLOS One, 5*(5), e10497.

Kruschke, J. (2014). *Doing Bayesian Data analysis: A Tutorial with R, JAGS, and Stan*. https://nyu-cdsc.github.io/learningr/assets/kruschke_bayesian_in_R.pdf

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). University of Chicago Press. (Original work published 1962)

Kyburg, H. E. (1965). Salmon's paper. *Philosophy of Science, 32*(2), 147-151. https://doi.org/10.1086/288034

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Blackwell.

Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1-25.

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

Martin, O. (2016). *Bayesian Analysis With Python*. Packt Publishing Ltd.

Matamala-Gomez, M., Malighetti, C., Cipresso, P., Pedroli, E., Realdon, O., Mantovani, F., & Riva, G. (2020). Changing body representation through full body ownership illusions might foster motor rehabilitation outcome in patients with stroke. *Frontiers in Psychology, 11*, 1962. https://doi.org/10.3389/fpsyg.2020.01962

Mattersberger, M. (forthcoming). *On Granularity Matching and Scale Mapping: A Novel Challenge for Multiple Realization* [Paper presentation]. 45th International Wittgenstein Symposium, Kirchberg am Wechsel, Austria

Meehl, P. E. (1992). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. In R. B. Miller (Ed.), *The Restoration of Dialogue: Readings in the Philosophy of Clinical Psychology* (pp. 523-555). American Psychological Association. https://doi.org/10.1037/10112-043

Menzies, P., & Beebee, H. (2024). Counterfactual theories of causation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Summer 2024 Edition)*. https://plato.stanford.edu/archives/sum2024/entries/causation-counterfactual/

Miłkowski, M., & Litwin, P. (2022). Testable or bust: Theoretical lessons for predictive processing. *Synthese, 200*(462). https://doi.org/10.1007/s11229-022-03891-9

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Harcourt, Brace & World.

Oaksford, M., & Chater, N. (2008). Probability logic and the Modus Ponens-Modus Tollens asymmetry in conditional inference. In *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 97-120).

Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota Studies in the Philosophy of Science* (Vol. 2, pp. 3-36). University of Minnesota Press.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*, 283-311.

Polger, T. W., & Shapiro, L. A. (2016). *The Multiple Realization Book*. Oxford University Press.

Port, R. F., & van Gelder, T. (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37-48). University of Pittsburgh Press.

Putnam, H. (1979). *Mind, Language and Reality* (Vol. 2). Cambridge University Press.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science, 45*(2), 206-226.

Railton, P. (1981). Probability, explanation, and information. *Synthese, 48*, 233-256. https://doi.org/10.1007/BF01063889

Ramakonar, H., Franz, E. A., & Lind, C. R. (2011). The rubber hand illusion and its application to clinical neuroscience. *Journal of Clinical Neuroscience, 18*(12), 1596-1601.

Reichenbach, H. (1956). *The Direction of Time*. University of Chicago Press.

Reijula, S. (2017). How could a rational analysis model explain? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017*. Cognitive Science Society.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386-408.

Ross, L., & Woodward, J. (2023). Causal approaches to scientific explanation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. https://plato.stanford.edu/archives/spr2023/entries/causal-explanation-science/

Russell, B. (1912). On the Notion of Cause. *Proceedings of the Aristotelian Society*, *13*, 1–26. http://www.jstor.org/stable/4543833

Salmon, W.C. (1977). A third dogma of empiricism. In: Butts, R.E., Hintikka, J. (eds.) *Basic Problems in Methodology and Linguistics. The University of Western Ontario Series in Philosophy of Science* (Vol. 11). Springer. https://doi.org/10.1007/978-94-017-0837-1_10

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Salmon, W. C. (1989). *Four Decades of Scientific Explanation*. University of Pittsburgh Press. https://doi.org/10.2307/j.ctt5vkdm7

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593-1599.

Sher, S., & McKenzie, C. R. (2006). Information leakage from logically equivalent frames. *Cognition, 101*(3), 467-494.

Sher, S., & McKenzie, C. R. (2008). Framing effects and rationality. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, (p. 79-96). Oxford University Press.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review, 17*(4), 443-464.

Sprenger, J., & Hartmann, S. (2019). *Bayesian Philosophy of Science*. Oxford University Press.

Sternberg, R. J., & Sternberg, K. (2016). *Cognitive psychology* (7th ed.). Wadsworth.

Stigler, S. M. (1983). Who discovered Bayes's theorem? *The American Statistician, 37*(4), 290-296.

Teigen, K. H. (2002). One hundred years of laws in psychology. *The American Journal of Psychology, 115*(1), 103-118. https://doi.org/10.2307/1423676

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

van Riel, R., & Van Gulick, R. (2024). Scientific reduction. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Spring 2024 Edition)*. https://plato.stanford.edu/archives/spr2024/entries/scientific-reduction/

Von Eckardt, B., & Poland, J. S. (2004). Mechanism and explanation in cognitive neuroscience. *Philosophy of Science, 71*(5), 972-984.

Ward, J. (2019). *The Student's Guide to Cognitive Neuroscience*. Routledge.

Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Sciences, 2*(9), 338-347.

Woodward, J., & Ross, L. (2021). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*. https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/

Yalim, T., & Mohamed, S. (2023). Meltdown in autism: Challenges and support needed for parents of children with autism. *International Journal of Academic Research in Progressive Education and Development, 12*(1), 850-876.

Zou, J., Huang, L., Wang, L., Xu, Y., Li, C., Peng, Q., Zeng, H., Zhang, S., & Li, L. (2022). Neuronal cell-type and circuit basis for visual predictive processing. *bioRxiv*. https://doi.org/10.1101/2021.12.31.474667