# Robustness in the Life Sciences: Issues in Modeling and Explanation

by

**Morgan K. Thompson**

Undergraduate degree, University of Tennessee, Knoxville, 2010

Master degree, Georgia State University, 2013

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

**Morgan K. Thompson**

It was defended on

April 14, 2020

and approved by

Mazviita Chirimuuta, Associate Professor, Department of History and Philosophy of Science

David Danks, L.L. Thurstone Professor, Carnegie Mellon University, Department of Philosophy and Psychology

James Woodward, Professor, Department of History and Philosophy of Science

Kevin Zollman, Professor, Carnegie Mellon University, Department of Philosophy

Dissertation Director: Edouard Machery, Distinguished Professor, Department of History and Philosophy of Science

**Robustness in the Life Sciences: Issues in Modeling and Explanation**

Morgan K. Thompson, PhD

University of Pittsburgh, 2020

My dissertation introduces two new accounts of how robustness can be used to identify epistemically trustworthy claims. Through an analysis of research practices in the life sciences, I focus on two main senses of robustness: robust reasoning in knowledge generating inferences and explanatory strategies for phenomena that are themselves robust. First, I provide a new account of robustness analysis (called 'scope robustness analysis'), in which researchers use empirical knowledge to constrain their search for possible models of the system. Scope robustness analysis is useful for scientific discovery and pursuit whereas current accounts of robustness analysis are useful for confirmation. Second, I provide a new account of how researchers use different methods to produce the same result (a research strategy called 'triangulation'). My account makes two contributions: I criticize a prominent account of the diversity criterion for methods because it analyzes an inferential strategy (i.e., eliminative inference) distinct from the inferential strategy underlying triangulation (i.e., common cause inductive inferences). My account also better explains how triangulation can fail in practice by assessing points of epistemic risk, which I demonstrate by applying it to implicit attitude research. Finally, I contribute to a debate about another sense of robustness: phenomena that occur regardless of changes in their component parts and activities. I argue that some robust phenomena in network neuroscience are not best explained mechanistically by citing their constituent parts (e.g. individual neurons) and their activities, but rather by appealing to features of the connectivity among brain areas.

## Table of Contents

## List of Tables

## List of Figures

## Preface

Many people have helped me along my academic journey that led to this dissertation. Most directly, I have benefited from the intellectual contributions and mentorship of my advisor and committee members. I thank Edouard Machery for reading countless drafts and guiding the development of my ideas and arguments. No doubt my work is better for his critical feedback. Thank you to Mazviita Chirimuuta for crucial guidance early on in my dissertation project. Her work also was always a model for excellent philosophy of science in practice during my time in graduate school. I thank Jim Woodward for his careful readings of my work and suggestions that helped me clarify my own views.

Kevin Zollman and David Danks require special mention as both intellectual and professional mentors. Kevin's work and ideas in network epistemology has provided an excellent comparison case for much of the network neuroscience throughout this dissertation. I benefited greatly from many conversations exploring the similarities and differences of the use of network models in these different domains. Kevin has also been generous and supportive in his valuable feedback on my work. David has always been supportive of not only my work, but also my abilities as a philosopher. His feedback has helped me see the value of my work from a different perspective and to develop the positive accounts provided in this dissertation. David has also been a wonderful mentor in guiding me to reflect on the philosopher and person I want to be.

My work has also benefited from other faculty in Pittsburgh. Bob Batterman has been supportive of my ideas and his thoughts on multiple realizability and explanation were influential in my initial views on mechanistic explanation and ultimately, Chapter 4 of this dissertation. The influence of major ideas in his work can also be seen in Chapter 2. Jim Bogen also deserves

mentioning for his early support and excitement about my ideas in cognitive science. Jim's course on mechanistic explanation and his early suggestions of MVPA research led to some of my earliest work on mechanistic explanation, which influenced Chapter 4 of this dissertation. Ken Schaffner's course on philosophy of psychiatry also allowed me the opportunity to explore network neuroscience more fully at the beginning of this dissertation. I am grateful for Ken's vast intellectual knowledge and his willingness to share resources.

Less directly, my intellectual journal has also been supported by faculty from my undergraduate and Master's programs. At the University of Tennessee, Knoxville, Clerk Shaw's course on Ancient Philosophy helped me gain confidence in my philosophical abilities. Clerk has also remained supportive over the years. EJ Coffman supported my efforts in applying to graduate school, which I could not have done without him. Richard Aquila also taught me much about philosophical writing and research through his courses. At Georgia State, my Master's advisor Eddy Nahmias was incredibly helpful at further developing these philosophical writing and research skills. Collaborating with Eddy allowed me to see how to write an academic article and how to design empirical studies. Eddy also was incredibly supportive of my ideas; my work in the demographics of philosophy likely not exist without his eager support of my and Toni Adleberg's early ideas on the topic. I owe much of my intellectual development to Eddy's mentorship, which has continued even now.

The graduate students at Georgia State and Pitt have been incredibly giving of their time, feedback, and support. In the past year, I benefited so much from a working group with Annika Froese, Mahi Hardalupas, and Nedah Nemati. I would like to thank them all for their friendship, support, and advice. Annika, Mahi, and Nedah has been there for me throughout the difficult times. Thank you to Willy Penn for his support and friendship. No doubt my life would be poorer without

them. Trey Boone and Ashley Jardina have been excellent friends. Michael and I have often appreciated the weekends we've spent together. I am grateful for their support of my career as well. Joseph McCaffrey has also been supportive through all my struggles and a great friend. I would also like to thank Lauren Ross, Katie Tabb, Zina Ward, Siska De Baerdemaeker, Haixin Dang, Marina Baldissera Pacchetti, Aaron Novick, David Colaço, Daniel Malinsky, Mike Miller, Jen Whyte, and Marina DiMarco. During my time in Pittsburgh, I've also appreciated Kevin Zollman and Korryn Mozisek's friendship; both have always reminded me to celebrate the milestones throughout this long academic process. Thank you also to Liam Kofi Bright for his friendship and support of my intellectual projects. Conversations with Liam have helped me develop my thoughts on many topics, ranging from Carnapian explication to the demographics of philosophy to best practices in teaching and mentoring.

I also appreciate all of the friendships through my time at Georgia State. I would not have made it through the program and into my PhD program without their support. Blake Nespica has been a wonderful friend to both Michael and me. Toni Adleberg was an excellent friend during my time at Georgia State, to whom I am indebted for much of my early academic work. I am also grateful for the friendships that developed with Zack Hopper, Caitlin Hopper, Crawford Crews, and Chelsea Crews.

Finally, thank you to my family for being supportive of my career, even when I decided to pursue philosophy rather than psychology. Thank you to my parents, Wanda and Kerry, for supporting me emotionally and financially, each at different times. I appreciate that they have been open about their encouragement and their pride in my accomplishments. Most importantly, thank you to Michael Mahoney for being my emotional rock through many years of graduate school. He has been there for me through the times when I encountered frustrating writing blocks or lost

confidence in my abilities. I am grateful that Michael always lifts me up and grounds me through his calm demeanor and love. Even more, Michael has taught me to find the good in life and live in the moment. He is kind, patient, and thoughtful; Michael inspires me to be a better person each day.

## 1.0 Introduction

Scientific investigation comes with epistemic risk. Researchers are often in a position where they do not know how their methods work or they find their methods are prone to certain types of error. Yet, it is these fallible methods upon which we build our theories and construct scientific knowledge. To sort through this epistemic uncertainty, researchers need strategies and heuristics to sort noise and error from phenomena of interest. A variety of practices that fall under the term 'robustness' purport to play this role.

Robustness is broadly the invariance of a phenomenon, a result, or a prediction over variation. Robustness is said to help distinguish reliable claims from unreliable claims. It can both (1) guide how we search for robust results in our scientific investigations and (2) be a hallmark of phenomena that we'd like to explain.

'Robustness reasoning' is the idea that some claim is more strongly supported by multiple lines of argument or evidence than a single line of argument or evidence alone. Consider how this reasoning functions when evaluating the evidence provided by eye witnesses of a crime. Multiple independent eye witnesses provide much better evidence that the accused likely committed the crime than a single eye witness alone. However, as this example illustrates, it is important that these eye witnesses base their testimony on independent knowledge. If the eye witnesses talk and come to consensus about what they've seen before providing their testimony, then the evidence that the accused committed the crime is diminished. After all, it could be that one witness swayed the others with false information or that all of the witnesses colluded.

Robustness reasoning is particularly useful in science, where it can help resolve issues associated with epistemic uncertainty. Scientists use robust reasoning when they employ multiple

methods to investigate the same phenomenon or multiple models to make predictions about the same system. 'Robustness analysis' refers to the practice of using robustness reasoning over differences in modeling assumptions; 'methodological triangulation' refers to the practice of using robustness reasoning over differences in methods (including experimental protocols and scientific instruments).

Philosophical accounts of robustness in science owe much to William Wimsatt (1981). According to Wimsatt (1981, 62), robustness reasoning is characterized by the following activities:

> (1) To analyze a variety of independent derivation, identification, or measurement processes.
> (2) To look for and analyze things which are invariant over or identical in the conclusions or results of these processes.
> (3) To determine the scope of the processes across which they are invariant and the conditions on which their invariance depends.
> (4) To analyze and explain any relevant features of invariance.

Focused on a grand unifying project to demonstrate the importance of robustness to scientific practice, Wimsatt brings together cases ranging from using different sensory modalities to perceive some property, different experimental methods to generate a phenomenon, the discovery of invariance of a law or regularity over variation at a lower scale, and deriving the same result from multiple models. According to Wimsatt (1981, 63), all types of robustness have a common function in:

> distinguishing of the real from the illusory; the reliable from the unreliable; the objective from the subjective; the object of focus from artifacts of perspective; and, in general, that which is regarded as ontologically and epistemologically trustworthy and valuable from that which is unreliable, ungeneralizable, worthless, and fleeting.

Wimsatt himself was influenced by Donald Campbell (Campbell, 1969; Campbell & Fiske, 1959) and Richard Levins (1966). Campbell and Levins form the two central figures inspiring much of the work on robustness since. However, the influence of Campbell's work has been less recognized in the philosophical literature on robustness than the influence of Levins's work.

Campbell (1969) views robustness as multiple independent and imperfect measures that can be used to bootstrap ourselves into scientific understanding. Campbell introduces his ideas of robustness as multiple operationalism, stemming from operationalism in psychology and logical positivism more broadly (see Feest 2005). Operationalism holds that the meaning of a term can be defined by a measurement procedure. Multiple operationalism recognizes that terms can be operationalized in a variety of ways and so, denies that terms get their meaning from operational definitions. Instead, returning to Campbell's (1969, 33) specific multiple operationalism, he describes the problem as follows:

> [W]e have only other invalid measures against which to validate our tests; we have no 'criterion' to check them against… In this predicament, great inferential strength is added when each theoretical parameter is exemplified in 2 or more ways, each mode being as independent as possible of the other, as far as the theoretically irrelevant components are concerned.

I will return to Campbell's work particularly in Chapter 3.

Levins (1966) inspired most of the philosophical work on robustness reasoning in modeling (i.e., robustness analysis). When modeling complex systems in evolutionary biology, Levins (1966, 423) claimed that "our truth is the intersection of independent lies." That is, it is a problem that all of our models are strictly speaking false, in so far as they have simplifying abstractions and misleading idealizations, when we are often not in a position to tell if the results of our models are due to idealizing assumptions or to core features of the model. However, the problem can be addressed by using multiple models of the same system and looking for convergence in their results. If none of the idealizations in our models make a difference to this convergent result, then we can be more confident in their results.

My dissertation seeks to clarify some of the ideas in Wimsatt's work on robustness. First, I continue Wimsatt's emphasis on how robustness can be "illusory" (Wimsatt 1981, 64) in my

account of methodological triangulation (Chapter 3). Second, in my discussion of implicit attitudes in Chapter 3, I bring Campbell's ideas back to the fore in the philosophical literature on robustness, while the philosophical literature on robustness since Wimsatt has largely taken inspiration from Levins's work (e.g., Weisberg 2013). Finally, Wimsatt's work is again influential in my discussion of the explanation of robust phenomena in Chapter 4.

These senses of robustness reasoning are also related to robust phenomena. Wimsatt (1981, 79) discusses the robustness of some phenomena: "upper-level phenomena and laws [have] a certain insulation from (through their invariance over: robustness again!) lower-level changes and generates a kind of explanatory and dynamic (causal) autonomy of the upper-level phenomena and processes." In other words, some phenomena are stable over changes in their initial conditions, boundary conditions, or perhaps the organization of components and activities that produce them. One question that arises is how best to explain robust phenomena. Much of this debate has played out in terms of the scope of the mechanistic account of explanation. Some mechanists claim that all explanations are mechanistic—including explanations of robust phenomena. Some critics take their arguments to commit them to the idea that providing more detailed information about some phenomena always gives a better explanation.

## 1.1 Outline

This dissertation analyzes the concept of robustness, or the invariance of some feature despite variation in other features, in scientific practice. There are two main senses of robustness in science: robust reasoning and phenomena that are themselves robust. My dissertation covers

each sense in turn. My central aim in this dissertation is to expand and explicate the utility of robustness in science. In the case of robustness analysis, I argue that it is a useful strategy for contexts of discovery and pursuit, whereas the received view in philosophy holds that it is useful in contexts of justification. In the case of triangulation, I argue that it is a strategy more prone to inferential risks than philosophers of science have recognized. And in the explanation of some robust phenomena, I argue that some are best explained non-mechanistically.

I examine two scientific subfields in my analysis of how robustness is utilized during scientific investigation and how robust phenomena are explained. I consider both network neuroscience and social psychology. In both subfields, there is epistemic uncertainty about the methods and models employed. They are rife with robustness reasoning and in many cases, also aim for explanation of robust phenomena.

This dissertation proceeds in three chapters: Chapter 2 proposes a new type of robustness analysis that is useful in hypothesis selection rather than for the confirmation of some hypothesis or elaboration of a robust theorem. Chapter 3 provides a new account of triangulation focused on epistemic risk that aims to show some of the many ways triangulation can fail in practice. Finally, in Chapter 4, I argue that topological explanation is an alternative to mechanistic explanation and provide a case of topological explanation from network neuroscience. Below I discuss each chapter in more detail.

In Chapter 2, my work examines how the epistemic context of research projects affects the practice of searching for invariance between some result and set of models. When researchers have multiple models of a target system but are unsure which model best suits their purposes, they may examine whether all the models make some common prediction or have some common property. If convergence is found, an inference from the models to the predicted outcome or the attribution

of the property to the system is robust over variation among the models. Most work on robustness analysis has focused on whether convergence on results from multiple, diverse models can provide confirmation, and has thus examined robustness in the context of confirmatory science (where the goal is to support or undermine hypotheses). I argue that searching for a set of models that make the same prediction or have a common property can also be fruitful in contexts of scientific discovery and pursuit, particularly when empirical constraints are used to identify a set of possible models and rule out other models. Thus, I identify and characterize a new type of robustness analysis called 'scope robustness analysis'. I demonstrate that researchers use scope robustness analysis in assessing the relative contributions of generating principles on the organization of the *C. elegans* nervous system.

Robustness analysis is a process by which researchers search for a relation of invariance between a set of models of the same target system and some property or prediction. In traditional robustness analysis, it is determined whether some property or prediction is invariant over the set of models, which have varied assumptions. Traditional robustness analysis has been characterized by many philosophers on the basis of examples from climate science, biology, and economics (e.g., Lloyd 2015; Parker 2011; Weisberg and Reisman 2008; Kuorikoski, Lehtinen, and Marchionni 2010; Wimsatt 1981). Weisberg (2006; 2013; Weisberg and Reisman 2008; influenced by Levins 1966) offers the most influential account in the philosophical literature of the process by which researchers make predictions about a target system or ascribe a property to it from a set of models. The set of models must represent the same target system, but the models may make heterogeneous assumptions about it.

Much of the recent discussion of traditional robustness analysis has focused on whether (and how) agreement (or convergence) among the models confirms the ascription of some property

to the target (Houkes & Vaesen, 2012; Lloyd, 2015; Odenbaugh & Alexandrova, 2011; Orzack & Sober, 1993; Parker, 2011; Schupbach, 2015). For example, Orzack and Sober (1993) claim that convergence alone cannot confirm the ascription of a feature to the target unless all possible models of the target system are used in the analysis, which is almost never the case. Alternately, Lloyd (2015) argues that convergence among models on some common property or prediction can confirm the ascription of some feature to the target system, but only in conjunction with independent evidence for the shared model assumptions.

In Chapter 3, my primary contribution is a new account of triangulation that explains how the practice of triangulation can fail. To understand why this contribution is important for the literature on triangulation, I first explain and defend the diversity criterion in triangulation. Exactly how to understand what it means for methods to be sufficiently diverse for triangulation to lead to successful inferences has been the focus of the philosophical literature on triangulation. Most philosophers agree that triangulation (and robustness analyses, as well) require methods that make diverse assumptions. The simplest interpretation of this criterion is that each method needs to have completely independent assumptions. But it is unhelpful because it would mean triangulation will almost always fail to employ sufficiently diverse methods. Sometimes this is cashed out in terms of confirmational diversity (Fitelson, 2001; Lloyd, 2015). Otherwise, it is often cashed out in terms of the methods being subject to different sources of error (Wimsatt, 1981). A new account of the diversity criterion holds that only methods that could provide discriminating evidence between the hypothesis of interest and an alternative explanation are sufficiently diverse (Schupbach, 2015, 2018). In this chapter, I argue that Schupbach's argument for the new explanatory criterion of diversity changes the topic from triangulation to eliminative inference.

I also introduce a new account of triangulation focused on types of epistemic risk and the locations they arise during the research process. I use this account of triangulation to analyze the failure of the triangulation argument for implicit attitudes. In particular, I argue that there are at least two types of epistemic risk that arise in implicit attitude research: (i) the risk that data cannot justifiably serve as evidence for the same hypothesis and (ii) the risk that there is insufficient evidence to infer the presence of the phenomenon due to the presence of other plausible hypotheses.

Finally, in Chapter 4 I analyze an explanation of the pattern of robustness and vulnerability in the human macroscale brain. My work draws on cases in network neuroscience to argue that the organization of the macroscale human brain, when considered from a network perspective, explains the pattern of robustness and vulnerability without appealing to cellular or molecular details. As a result, I argue that there are cases of non-mechanistic explanation in neuroscience.

Many mechanists believe that robust phenomena are best explained by mechanistic explanation, which describes how entities and activities in a particular organization produce the robust phenomenon. In neuroscience, Kaplan and Craver (2011) claim that all explanations are mechanistic. Other philosophers of science (Batterman & Rice, 2014; Huneman, 2010; Ross, 2015) argue that some explanations are not mechanistic by presenting various purported counter-examples of explanations that do not appeal to a system's entities and activities nor make use of heuristics like decomposition (Bechtel & Abrahamsen, 1991). Mechanists tend to respond to these critiques by arguing that the counterexamples are either (i) not explanatory or (ii) mechanistic explanations.

One alternative type of explanation that has been proposed in support of the counter-examples raised is topological explanation (Huneman, 2010, 2015). While mechanistic

explanations explain in virtue of the productive organization of components and activities of a system, topological explanations explain solely in virtue of the topological properties of the system. Topological properties are general properties of some system's non-spatial organization, often concerning the connectedness of elements in the system. I argue that topological explanations are explanatory, at least in some cases, because they meet two norms of explanation that are accepted by many mechanists: (1) answering w-questions and (2) the asymmetry of explanation. I argue that some cases of topological explanation are not mechanistic because they include so few mechanistic details.

## 2.0 Scope Robustness Analysis: Modeling Possibilities

I argue that there is a role for robustness analysis in discovery and pursuit that has not yet been explored in the philosophical literature and is distinct from traditional accounts of robustness analysis. My account—scope robustness analysis—and the received view are distinct in the following ways: appropriate in different epistemic situations, have different success and failure conditions, and have different functions. Scope robustness analysis better accounts for robustness analysis used to explore growth principles in network neuroscience and biology where well accepted models of the target system already exist (e.g., *C. elegans* wiring diagram).

### 2.1 Introduction

Robustness analysis is a process by which researchers search for a relation of invariance between a set of models of the same target system and some property or prediction. In traditional robustness analysis, it succeeds when some property or prediction is invariant over the set of models, which have varied assumptions. If not all models of the target system have the same properties or make the same predictions, then the set of assumptions that differ among them should be re-examined. If all of the models do converge on some property or prediction despite their different modeling assumptions and the set of models are all plausible models of the target system, then perhaps there is defeasible reason to ascribe the property to the target system.

I argue that some cases of robustness analysis ('scope robustness analysis') are not well-described by traditional accounts of robustness analysis. In fact, we need a new type of robustness

analysis to account for this practice and its utility. I describe the two types in section 2.2. Then I describe existing views on modeling possibilities, such as how-possibly explanations and perspectival modeling in section 2.3. Scope robustness analysis is the practice of investigating the scope of possible models of a system for which some particular known property or observed outcome is invariant. The research questions and epistemic contexts of network neuroscience make the subfield particularly apt for scope robustness analysis. I illustrate scope robustness analysis with a case from network neuroscience in section 2.4. The two types of robustness analysis are most appropriate when researchers are in different epistemic situations regarding the target system, have different success and failure conditions, and follow different rules for defining the set of models explored. In section 2.5 apply these criteria to argue that the case from section 2.4 is best seen as an example of scope robustness analysis. Then I consider objections to my claim that there are two types of robustness analysis in section 2.6.

## 2.2 Two Types of Robustness Analysis

Here I describe the predominant view and cases of robustness analysis, which I call 'traditional robustness analysis.' Then I introduce my new account of 'scope robustness analysis.' On my view, it is fruitful to view general robustness analysis as: the search for an invariance relation between a set of models and some property or prediction.

### 2.2.1 Traditional Robustness Analysis

Call it 'traditional robustness analysis' when researchers are investigating whether some property of interest X is invariant over a set of models that make different assumptions. Woodward (2006) calls this general version of robustness analysis 'inferential robustness'. Traditional robustness analysis has been characterized by many philosophers using cases from climate science, biology, and economics (e.g., Lloyd 2015; Parker 2011; Kuorikoski, Lehtinen, and Marchionni 2010).

In more sophisticated forms of traditional robustness analysis, it involves the search for robust relationships between parts of the model (i.e., particular assumptions) and the results of interest while altering other assumptions of the models. Traditional robustness analysis can demonstrate the extent of reliability of our inferences from assumptions in our models to new predictions, or in the more sophisticated version, causal claims.

Weisberg (2006; 2013; Weisberg and Reisman 2008; influenced by Levins 1966) offers the most influential account of the sophisticated view. According to Weisberg, researchers can identify and assess the stability of successful inferences from some features of the models in the set to some properties or predictions of the models (call these inferences 'robust theorems'). The set of models must represent the same target system but make heterogeneous assumptions about the system.

Weisberg (2006) provides four-step procedure for robustness analysis:

(1) Determine whether a robust property or prediction (RP) follows from each [model in the set] $M_1$, $M_2$, …, $M_n$;
(2) Determine whether $M_1$, $M_2$, …, $M_n$ share a common core set of assumptions (CC);
(3) Formulate a 'robust theorem' connecting CC and RP using the following general form: 'Ceteris paribus, if CC obtains, then RP will obtain';
(4) [Optional] Assess the scope and strength of robust theorem by analyzing stability of relationship between common structure and robust property.

The fourth and final step of in inferential robustness is optional, but it involves determining the limits of the application of the robust theorem. Here Weisberg acknowledges that "[o]bviously every model cannot be investigated," so instead researchers can selectively assess the limits of the robust theorem (2013, 159). This exploration will determine the assumptions under which the robustness of the theorem fails. If a particularly wide investigation is undertaken, then researchers might replace the ceteris paribus clause in the robust theorem with particular conditions under which the inference holds.

In biological modeling, Weisberg and Reisman (2008) note that the Lotka-Volterra principle holds across a variety of predatory-prey interaction models with different assumptions.[1] The Lotka-Volterra principle states (Weisberg, 2006, 737): "Ceteris paribus, if the abundance of predators is controlled mostly by the growth rate of the prey and the abundance of the prey controlled mostly by the death rate of predators, then a general pesticide will increase the abundance of the prey and decrease the abundance of predators." The robust theorem for Lotka-Volterra models demonstrates a causal relationship between growth rates, death rates, and population levels of predators and prey. It is general in the sense that it can apply to many different actual and possible systems.

[1] Here and throughout I assume that varying modeling assumptions results in distinct models. This position makes sense of the use of multiple models in these classic robustness analysis cases. Model families are the general types or categories of models with small changes in parameters or minor assumption changes. However, my argument does not hinge on this claim. If the reader finds it unpalatable to proliferate the number of models, then they can translate many of my claims about robustness analysis into their own terminology according to their favorite theory of model individuation. For example, if new models are created only when varying the value of biologically significant variables, then scope robustness analysis might be referred to as 'sensitivity analysis' on some views (e.g., Raerinne, 2013, 289-290). Still there are important differences between the way that sensitivity analysis is typically characterized (as checking the stability of a model's results under changes in initial and boundary conditions) compared to features of scope robustness analysis (as examining the ability of different principles (or the extent of their contribution) to produce the result).

Ultimately, Weisberg's account of robustness analysis is a special case of traditional robustness analysis in which all the models have an additional commonality (the CC) beyond the robust property or prediction and thus, a robust theorem can be defined. This aspect of Weisberg's account also inspired Knuuttila & Loettgers's (2011) account of causal isolation robustness analysis, wherein researchers aim to identify the core causal mechanism. Robustness analysis is used to isolate the core causal mechanism by searching for what causal features are common to all models that produce the result, but unlike Weisberg's robustness analysis, Knuuttila and Loettgers claim that causal isolation robustness analysis does not aim to confirm the common result. Instead, the focus is to sufficiently isolate the causal mechanism producing the common result, which Knuuttila and Loettgers recognize will often require the use of empirical evidence. Raerinne (2013) calls this sufficient parameter robustness analysis.

Much of the recent discussion of traditional robustness analysis has focused on whether it provides empirical confirmation, non-empirical confirmation, or increases the reliability of our inferences (Houkes & Vaesen, 2012; Lloyd, 2015; Orzack & Sober, 1993; Parker, 2011; Schupbach, 2015).[2] Most interlocutors agree robustness analysis alone does not provide empirical confirmation. However, some philosophers elaborate views on how robustness analysis could provide confirmation indirectly and/or in conjunction with empirical support for modeling assumptions without relying on a notion of non-empirical confirmation (Lloyd, 2015; Weisberg, 2013). There has been comparatively little focus on how robustness analysis might be a tool for discovery and pursuit of hypotheses. One exception is Odenbaugh & Alexandrova (2011), who assume but do not argue for the claim that robustness analysis is widely used in biology and

[2] I set aside Schupbach's view of robustness analysis here, but I address his account in Chapter 3. His account primarily rests on an interpretation of a case of triangulation, not robustness analysis.

economics to construct schematic causal hypotheses that are later tested. Still even Odenbaugh and Alexandrova focus on how robustness analysis can be used to formulate hypotheses about causal mechanisms—akin to Weisberg's "common core" assumptions or Knuuttila and Loettgers's "core causal mechanism" represented by a set of modeling assumptions.

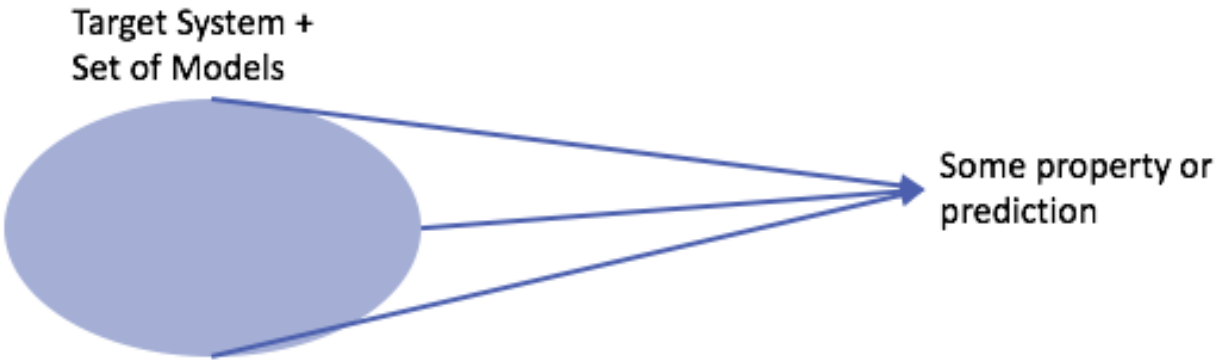


**Figure 1. Schematic of the traditional account of robustness analysis.**

### 2.2.2 Scope Robustness Analysis

There are cases that look similar to traditional robustness analysis in that the robustness relation between some set of models and a property or prediction made by those models is examined by researchers. However, unlike traditional robustness analysis, there is already sufficient evidence to ascribe the property to the target system. Often the study is aimed at determining how the target system comes to have the properties that it does rather than identifying those properties in the first place. For example, Solé, Pastor-Satorras, Smith, & Kepler (2002) aim to model different growth principles for the yeast *S. cerevisiae* protein-protein interaction network model. The network represents proteins as nodes and the potential for interactions between any two proteins as edges. As the protein-protein interaction network model for *S. cerevisiae* and other organisms (e.g., viruses, prokaryotes) have already been modeled, the researchers know the

properties of these models: scale-free degree distribution, small-worldness, and robustness against random but not targeted node deletion. Rather than build further network models from empirical data on protein interactions, Solé et al. (2002, 3) aim to model simplistically "proteome evolution aimed at capturing the main properties exhibited by protein networks."

To do so, they generate network models according to different growth principles. In particular, they model gene duplication and mutation by: (i) copying a randomly chosen node in the graph (i.e., duplication), (ii) removing the edges from this new node with a certain probability (i.e., mutation), and (iii) creating new edges between the new node and other nodes with some probability. The effects of modeling gene duplication and mutation using a data model of the protein-protein network might lead to a resulting non-actual protein-protein network model. The idea is to generate models using possible gene duplications and mutations that could have altered a protein in such a way that it either now interacts with or fails to continue to interact with some other protein in the network. However, the values of the probability in steps (ii) and (iii) are constrained by empirical data but varied to generate different synthetic models. Then Solé and colleagues compared the synthetic models to the data model of the *S. cerevisiae* protein-protein interaction network. Specifically, they look for whether the synthetic models have the same network properties as the data model and the boundaries at which the synthetic models no longer have similar properties to the data model. This case appears distinct from traditional robustness analysis because it is primarily focused on modeling possibilities and identifying the boundaries of possible protein-protein network configurations under different generating principles.

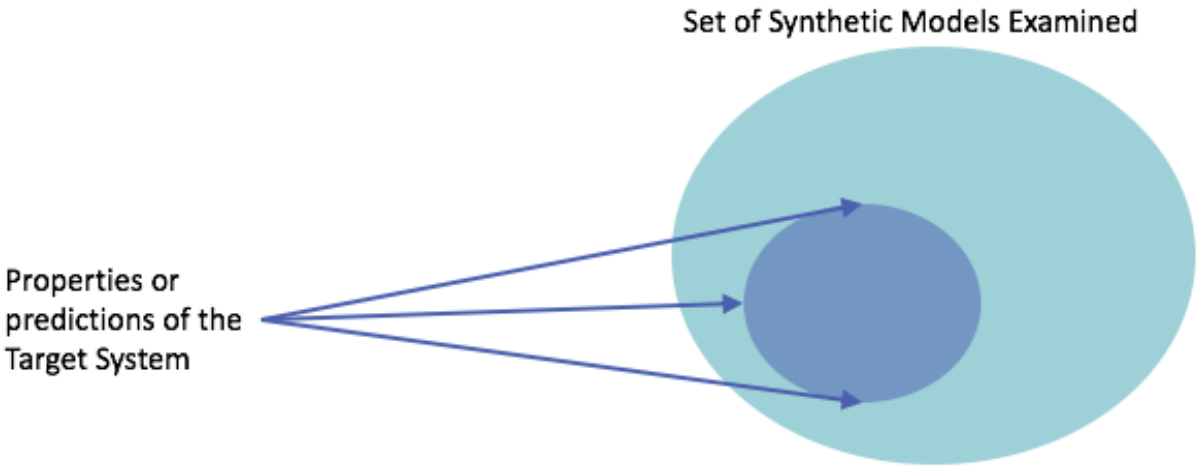


**Figure 2. Schematic of scope robustness analysis.**

Solé and colleagues are learning information about the (im)possible principles for the growth of the *S. cerevisiae* protein-protein interaction network model. They do so following the general logic of robustness analysis—by examining invariance of properties over differences in assumptions and idealizations in a set of models. The current literature on robustness analysis does not take a stance on whether this type of case is an instance of robustness analysis and whether existing views of robustness analysis can be adequately extended to account for these cases.

Call it 'scope robustness analysis' when researchers search for the scope of the set of models Y for which some particular property or prediction is invariant. While both processes involve evaluating the robustness relation between a property of interest and a set of models, these types of robustness analysis are useful in different epistemic contexts and serve different purposes. These hypothesized models of the system can be constructed according to different organizational or developmental principles and these principles are the ultimate target of inquiry. While both scope robustness analysis and traditional robustness analysis involve evaluating the robustness relation between a property of interest and a set of models, they examine different relata of that relationship. Traditional robustness analysis gives researchers more confidence in common

predictions or properties ascribed to the target. Scope robustness analysis, on the other hand, identifies possible models of the target, such as the developmental processes that result in some specific neural organization. These possible models can answer questions about other ways a system could be organized to produce the same behavior or more indirectly, to illustrate which developmental principles can generate possible models of the system consistent with known properties of the system.

**Table 1: A Comparison of Key Features of Traditional and Scope Robustness Analysis.**

| *Features* | *Traditional Robustness Analysis* | *Scope Robustness Analysis* |
|---|---|---|
| Epistemic Context Relative to Target System | Under-described; uncertain what modeling assumptions are warranted | Well-accepted descriptive model required; know which modeling assumptions are warranted |
| Type of Research Question | How best to describe the target system? | How did the target system come to have the properties that it does? |
| Function | Identifying differences in modeling assumptions that do not make a difference for certain properties or predictions (need not assess every assumption) | Identifying principles (or trade-offs among competing principles) that can produce models with the observed properties |
| Conditions of Failure | Do not identify any invariant property | Do not identify properties of descriptive model as invariant |
| Success can allow… | Increasing the reliability of inferences from some core modeling assumptions to the common result; indirect confirmation of the result | The generation of possible models useful for empirical investigation; Ruling out principles (e.g., growth principles) that are unable to generate models with the properties of the descriptive model |

In what follows, I will argue that traditional robustness analysis and scope robustness analysis answer different research questions, have different functions, and conditions of failure.

To do so, I will illustrate what I call scope robustness analysis and provide a detailed account of the differences from traditional robustness analysis.

## 2.3 Models Demonstrating Possibility

Because scope robustness analysis generates models that represent the possible states of a system, I will now turn to existing philosophical accounts models of possibilities: how-possibly explanations and perspectival modeling. There are some similarities to the generated models in scope robustness analysis and the models in these views.

In the context of explanations, some models contribute to how-possibly, how-plausibly, or how-actually explanations (Craver, 2007; Machamer, Darden, & Craver, 2000). Intuitively, how-possibly explanations are provided when model(s) demonstrate that some outcome could have happened. How-plausibly explanations are ill-defined and sit between how-possibly and how-actually explanations in terms of how much confirmation these models have and the extent to which they support inferences from the models. How-actually explanations by indicating how some outcome actually happened. The philosophers who introduced these notions take them to be a continuum of explanations with less to more empirical confirmation.

Other views hold that how-possibly and how-actually explanations are not on a continuum of empirical confirmation. Instead, how-possibly explanations have a different logical form than how-actually explanations (Forber, 2010). According to Forber, global how-possibly explanations work by examining general processes in idealized models to identify relationships between processes and particular outcomes. Local how-possibly explanations work much the same as global how-possibly explanations, but specifically aim to explain an outcome in a particular

system. I take no stance on whether scope robustness analysis provides explanations and instead, I focus on how it can be useful in discovery and pursuit. Still in Forber's account of local how-possibly explanations, the general (evolutionary) processes play a role similar to the role principles play in my account of scope robustness analysis. They set the bounds of what possible states we could have observed in the target system. In scope robustness analysis, researchers are discovering which principles could play a role in the development of the actual target system. In how-possibly explanation, researchers use these principles (and the evidence supporting them) to derive modal consequences for the target system.

We find a useful account of how to model possibilities in Michela Massimi's (2018) work on perspectival modeling, though Massimi's work is situated in a debate about whether and how multiple, inconsistent models of the same target system are consistent with realism. Perspectival modeling is the practice of generating models with *sui generis* representational content for exploratory purposes. The *sui generis* representational content of perspectival models distinguishes them both from non-representational models (discussed in Batterman and Rice 2014) and models that represent the target system via a mapping relation (e.g., isomorphism, homomorphism). Rather, perspectival models represent possibilities about the target system (not states of affairs that are known to be either actual or fictional). It is for this reason that Massimi emphasizes perspectival models' modal role in scientific exploration and discovery: "it is about exploring and ruling out the space of possibilities in domains that are still very much open-ended for scientific discovery" (2018, 339). Massimi's perspectival models are a helpful starting place to think about how scope robustness analysis models possibilities.

In scope robustness analysis, the generated models reflect possible states of the system, conditional on the different principles that generate them. Importantly for scope robustness

analysis, these models are only evaluated for their similarities with data models of the system, not for their ability to represent the system directly. Unlike Massimi's perspectival modeling, however, researchers may also include models known to be generated using false principles (or false weightings of true principles). These models are known to be false and so, would not satisfy Massimi's criterion that perspectival models represent without researchers knowing the representations to be true or false.

## 2.4 Scope Robustness Analysis in Network Neuroscience

Two main research questions that drive the network neuroscience research program are: Which topological properties best characterize the organization of the neural system? and how does the neural system come to have that specific organization? I will focus on network neuroscientists' research projects aimed at answering the second question. I address the first question in Chapter 4.

Network scientists are interested in how network models come to have the properties that they do as opposed to other properties. Some network neuroscientists use generative modeling—synthetic models are created using simple principles for building the model—to explore possible answers to this question. These synthetic network models that are not intended to represent any particular target system, but rather to explore the relationship between various growth principles and resulting topological properties in the network models (e.g., Erdős and Rényi 1960; Watts and Strogatz 1998; Barabási and Albert 1999). These synthetic models are built according to different principles or differently weighted principles (e.g., generalizations that describe dynamic and functional relations within a set of systems; Green 2013) and then compared to the descriptive

models of the target system to determine which of the synthetic models have the same properties. If some growth principles cannot generate models with those properties, then researchers can rule out those growth principles that could not account for the properties of the target system. However, those definitive cases may be rare. Often multiple principles will be involved in generating and constraining the outcomes and sometimes principles will even produce competing constraints. In these cases, researchers may seek to estimate the relative contributions of different growth principles.

Scope robustness analysis is best used in cases where the researchers are knowledgeable about the target system and have already ascribed properties to the target system through empirical examination or via consensus among researchers in the research program. When a well-accepted descriptive model (e.g., Ankeny 2006) exists (as is the case for many systems in animal models), properties of that model can be used to restrict the set of synthetic models (and principles to those that could possibly generate the principles of interest).

### 2.4.1 Scope Robustness Analysis and the *C. elegans* Wiring Diagram

One example of scope robustness analysis comes from contemporary uses of research on the neural system of the nematode *Ceanorhabditis elegans* wiring diagram. *C. elegans* was chosen by Sydney Brenner for a research project to use genetic and molecular biological information to learn about nervous systems. Brenner identified hopes for the project's ability to discover "some underlying principles that […] are applicable at many different levels of biological complexity" (Adler, 1976). Insight into the *C. elegans* nervous system combined with complete genetic information were predicted to shed light on the growth and abstract developmental principles

involved in neural systems in general. Researchers identified a near complete wiring diagram for the *C. elegans* hermaphrodite neural system. By examining specimens using sequential electron micrographs to identify neurons and synaptic connections, Brenner and colleagues generated a schematic model of the *C. elegans* nervous system, abstracting from any differences in synaptic connections found between the specimens (Ankeny, 2001; White, Southgate, Thomson, & Brenner, 1986).

Nearly forty years later, network neuroscientists use this well-accepted wiring diagram to examine simple principles that could generate the set of network properties observed in the wiring diagram.[3] Chen et al. (2013) perform a robustness analysis in order to examine which trade-off between competing principles (i.e., wiring cost and efficiency of information transfer principles) can generate models with many of the topological properties of the *C. elegans* wiring diagram. Based on their knowledge of the topological properties of the *C. elegans* wiring diagram (i.e., the descriptive model), Chen and colleagues search for the following properties among the synthetic networks: a set of interconnected hubs, modularity, and robustness of the network to targeted deletion of the node with the highest degree. The information about the descriptive model is used to constrain the synthetic models.

To determine which models to examine in the robustness analysis, Chen et al. generate models that trade-off two competing hypotheses concerning general growth principles for neural systems. These hypotheses are that the brain evolved to be an efficient information processing system and also to minimize the developmental and metabolic costs associated with longer axonal connections between spatially distant neural bodies. Information traveling different spatial

[3] This strategy has been performed in the network modeling of other target systems as well such as *Saccharomyces cerevisiae* (Solé et al., 2002; Vázquez, 2010) and *Drosophila melanogaster* (Middendorf, Ziv, & Wiggins, 2005) protein–protein interaction networks.

distances in the system are subject to different costs (call these 'wiring costs'), including the conduction of action potentials along axons of different diameters and whether the neuron is myelinated (Cherniak, Changizi, and Kang 1999; Chklovskii and Stepanyants 2003). Longer axons also have further spatial distance for the action potential to conduct through, often resulting in delayed times in information transfer. Further, longer axonal connections are subject to other costs such as being developmentally and metabolically costly (Bullmore & Sporns, 2012). Thus, long distance axonal connections may be minimized.

However, initial research suggested wiring costs were not completely minimized. Network neuroscientists have examined the extent to which wiring cost minimization can be increased by non-actual spatial arrangements of the *C. elegans'* ganglia (which involve collections of neurons) and random rewiring from the actual network model (Cherniak 1995; Klyachko & Stevens 2003; Cherniak et al. 2004; Chen, Hall, and Chklovskii 2006; Pérez-Escudero and de Polavieja 2007). If the system were solely minimizing the costs of connections between parts of the brain, then only connections between local parts would exist. However, there are long distance connections in neural systems that would not be predicted by a solely wiring cost minimizing model (Bullmore & Sporns, 2012). Some evidence suggests that component placement is near, but not fully optimal for the minimization of wiring costs in the *C. elegans* wiring diagram (Kaiser & Hilgetag, 2006).

Network neuroscientists also view the brain as a system that maximizes efficiency of information transfer. Neural signals should travel quickly between brain areas including spatially distant ones. Chen et al. (2013) operationalize the principle as driving the minimization of average path length—the average of the minimum number of edges necessarily traversed in order to reach one node from another node for each set of nodes in the network. Chen and colleagues reason that if both competing principles play a role in the organization of the *C. elegans* neural system, the

trade-off between the two principles could result in many of the observed network properties of the *C. elegans* wiring diagram.

Chen and colleagues reconstructed particular connections while holding fixed the spatial layout of the nodes and the number of edges in the network. Using the following equation, the researchers set $\alpha$ to various levels to fix the trade-off between minimizing wiring cost ($L_p$) or minimizing path length ($L_g$): $L = (1 - \alpha)L_g + \alpha L_p$. For this equation, when $\alpha = 0$ the average path length in the network is minimized without regard to spatial constraints and when $\alpha = 1$ the wiring cost is minimized without regard to efficiency of possible information transfer. Using a simulated annealing optimization algorithm, they searched for organizations that minimized the function L describing the trade-off between wiring costs and efficiency of information transfer at different values of $\alpha$. The simulated annealing algorithm was performed on random networks with the same spatial layout for nodes and number of edges.

In evaluating the $\alpha = 1$ network model, Chen and colleagues note that the topological organization is significantly different from the known *C. elegans* wiring diagram. The wiring cost ($L_p$) was only 1.25% lower than the *C. elegans* wiring diagram, however, only 27% of edges are recovered in the $\alpha = 1$ model. There are no hubs in the wiring cost minimization network and the efficiency of that network is much lower. On the other hand, in the $\alpha = 0$ network the FLPL neuron, which is the most spatially central node, was the single hub so that all path lengths either connected to that neuron or through it. In this extreme model, Chen et al. note that the network would be vulnerable to targeted deletion of the hub node, which is inconsistent with the properties of the *C. elegans* wiring diagram. Chen and colleagues conclude that setting L at $0.8 < \alpha < 1$ generates network models with similar hubs and modules to the *C. elegans* wiring diagram and so, this trade-off between wiring cost minimization and efficiency of information transfer maximization is

sufficient for the known topological properties. While it is possible that other principles generate similar topological properties, the empirical support for wiring cost minimization from neurophysiology studies makes it likely that the two principles are involved in generating the known network properties.

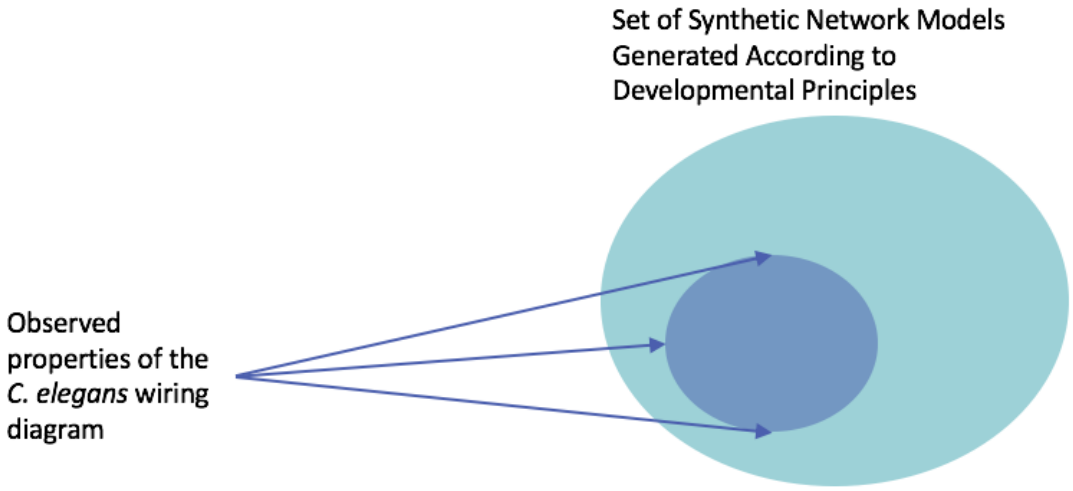


**Figure 3. Schematic of scope robustness analysis particular to Chen and colleages' research.**

One major advantage of exploring the characteristics of synthetic models at different levels of $\alpha$ is that values of $\alpha$ that produce networks with properties that are inconsistent with the properties of the descriptive model can be ruled out as possible descriptions of the generation of the *C. elegans* wiring diagram. For example, the synthetic models driven predominantly by the efficiency maximization principle had only one hub with a very high degree in the network; the wiring diagram model is characterized as having multiple hubs with high degree. It can therefore be ruled out that low levels of $\alpha$ and the trade-off balanced favoring maximization of efficiency of transfer is not a growth principle for the *C. elegans* neural system.

### 2.4.2 Contrast with Parameter Estimation

It might seem that Chen and colleagues are simply engaging in parameter estimation because they generated their models using a single parameter, which represents a trade-off between two growth principles. The suggestion implies that what Chen and colleagues are doing is estimating the range of actual values that $\alpha$ takes for the *C. elegans* neural system. On this interpretation, Chen and colleagues know how to model the growth principles of the network, but they simply do not know the quantitative value(s) the parameter can take. Their robustness analysis then, in a minimal sense, searches the possible parameter values to identify the range of empirical values for the development of the target system.

I think this interpretation fails to take into account the fact that Chen and colleagues do not take the two growth principles to be the only principles or constraints at work in the development of the *C. elegans* neural system. Interpreting the trade-off between the two principles allows them to more easily determine which growth principle drives the majority of the connections in the *C. elegans* wiring diagram. They can then use this information to guide and limit future directions for their research. Further, within the maximal range of reproduction of the properties of the known wiring diagram, it can be determined how many connections and properties cannot be accounted for. For example, Chen et al. suspect that another important constraint will be the tendency for neurons that process sensory information to have cell bodies that are spatially close to the pathways for incoming sensory information. Chen et al.'s current model cannot incorporate this growth principle as it would require including information about each neuron's functional role in the network. So, I do not think that Chen and colleagues view themselves as estimating the true range of values of $\alpha$ as if $\alpha$ were a real value of the *C. elegans* neural system.

## 2.5 Comparing Types of Robustness Analysis

Now that I have laid out the case for scope robustness analysis, I will argue that it is indeed distinct from traditional robustness analysis. At a coarse level of description, both types of robustness analysis involve similar activities: build a set of models varied in some of their assumptions and then examine the properties of those models looking for invariance. Chen et al.'s (2013) research process fits this coarse-level sketch of robustness analysis. They built a set of synthetic models varied in terms of the growth principles used to generate them. Then they examined the set of models and their properties looking for invariance of certain properties of interest. However, the case departs from traditional robustness analysis in a few distinct ways that I will enumerate in the rest of this section.

### 2.5.1 Goals of the Robustness Analysis and Epistemic Context

Viewing Chen and colleagues' research as the search for robust properties, as in traditional robustness analysis, would obscure the goals of these researchers. The studies by Chen and colleagues would be pointless if they were engaged in traditional robustness analysis because the wiring diagram model of the *C. elegans* nervous system is known on the basis of empirical investigations. When researchers agree on the best descriptive models of the system, there is no reason to perform a traditional robustness analysis to confirm the ascription of properties to the system. Their goal is to exploit this empirical information to learn more about other aspects of the system. In domains with vast amounts of data but little guiding theory, such as neuroscience, these strategies for exploiting empirical knowledge to direct future theorizing are particularly important.

Chen and colleagues also use scope robustness analysis to determine which growth principles cannot adequately represent the target, i.e., developmental processes for the system. By ruling out possible growth principles, researchers narrow the search space for future theorizing and experimentation. Because Chen et al. focused on the trade-off between two growth principles, they rule out ranges of $\alpha$ that are inconsistent with the descriptive model. If they had examined different growth principles that did not trade-off, they may have been able to rule out some principles completely as contributors to the development of the *C. elegans* neural system.

The extent to which scope robustness analysis can rule out some principle depends on whether all relevant principles were included in the analysis. The issue of what principles to include, and thereby what models will be included, in scope robustness analysis is one that researchers need to justify. Otherwise, researchers will not be justified in inferring from some principle generating no how-possibly models to that principle not playing a role in the organization or development of the system. It is possible that had some other principle been included in the analysis, the first principle could produce how-possibly models in conjunction with the other principle. This issue is similar to a problem in traditional robustness analysis: the choice of what models to include in the analysis can impact the inference that some hypothesis is correct (Orzack and Sober 1993). Orzack and Sober argue that researchers can guarantee that a hypothesis is correct in robustness analysis only if the researchers know that one model in the set is true. Otherwise it is possible that all models included in the robustness analysis are false and any convergence on some hypothesis by a set of false models gives no reason to increase credence in that hypothesis. So, researchers need to justify their choice of models in both types of robustness analysis.[4] One

[4] However, I do not agree with Orzack and Sober's argument as stated. All models are strictly speaking false because they idealize and abstract away from details of their target. So, researchers will never be in a position to utilize robustness analysis if it requires that researchers know that one of their models is true. Their argument can be softened

way to justify the choice of models in traditional robustness analysis is to appeal to evidence and theoretical reasoning in support of these models (Lloyd 2015; Winsberg 2018). The same solution could be extended to justifying the choice of principles to include in a scope robustness analysis.

The epistemic context with respect to the properties of the system are different for traditional and scope robustness analysis. The system in the *C. elegans* case is the *C. elegans* nervous system, which was already well-understood by the descriptive model (i.e., the wiring diagram). Still some questions about the system remain: How did it develop? How else could it have been organized? Thus, the researchers investigate their target (i.e., developmental processes) that produce the observed properties of the system (i.e., the *C. elegans* nervous system). Contrast this with traditional robustness analysis, which is useful when relatively little is known about how to describe the target system.

### 2.5.2 Criteria for Inclusion in the Set of Models

Another difference between the two types of robustness analysis involves their criteria for including models in the analysis. In traditional robustness analysis, researchers include only plausible models of the target system. They aim to determine whether any properties are invariant over differences in existing assumptions. Including implausible models in the set of models input to traditional robustness analysis could potentially obscure the invariance of some property among the models and thus fail to identify a true property of the target system.

to require only that one model in a robustness analysis provides the correct truth value regarding the hypothesis (Parker 2011).

Just as in traditional robustness analysis, researchers begin with a set of models, but they identify them according to different rules. Before scope robustness analysis, researchers do not yet know which principles can generate how-possibly models with the properties of interest, so they ought to include a number of plausible principles. Regardless of what criteria are chosen for determining which principles to include in an analysis, those for scope robustness analysis will be more inclusive than traditional robustness analysis. In the case of Chen et al., they included models constructed at regular intervals across the entire range of possible values of $\alpha$. They even included a model generated by completely minimizing wiring costs despite previous research demonstrating that complete minimization could not account for the spatial placement of nodes in the *C. elegans* wiring diagram.

My account of scope robustness analysis also better explains why researchers might include models they *already* know to be false, as Chen and colleagues do. Including some false models helps identify the scope of how-possibly models. Imagine there is a space of different hypothesized models for some target where models are located next to models generated with similar principles (or trade-off values for multiple principles). Researchers using scope robustness analysis to search the space of hypothesized models are looking for a whole range of how-possibly models, which may reflect similar growth principles used to generate them. Including false models, and thus searching more of the model space, allows researchers to identify the boundaries of how-possibly models. Consider the case of Chen and colleagues. They knew from previous research that the *C. elegans* wiring diagram doesn't have fully minimized wiring costs, but they could not rule out that wiring costs were *nearly* fully minimized. These hypothesized models need to be included in the analysis because they might be how-possibly models. By searching systematically

across the trade-off $\alpha$ for the two principles, Chen and colleagues are better able to identify where in the trade-off hypothesized models are consistent with the descriptive model.

### 2.5.3 Conditions of Failure

Failure for scope robustness analysis involves a failure to generate any models that have the properties of the descriptive model. In those cases, it is likely that the growth principles used to generate the network models do not represent their target: the developmental processes or organization of the system. Researchers should re-evaluate the growth principles and determine whether there are other possible growth principles that could be included in the robustness analysis. It could also be the case that the descriptive model does not accurately describe the target and so researchers should also ensure there is sufficient evidence for the descriptive model.

Traditional robustness analysis, on the other hand, fails when researchers cannot find some property or prediction on which all the models converge. In that case, there is no compelling reason provided by the robustness analysis to ascribe any property to the target system or to make any prediction about it. Suppose that in some robustness analysis of climate models mostly predicted that a particular glacier would melt within the next twenty years, but one climate model predicted that the glacier would not melt. There is no convergence among this set of climate models on this hypothesis. When this occurs, researchers re-evaluate the set of plausible models. It is possible that some of the models make a mistaken assumption about the target system that produces the lack of convergence.

### 2.5.4 Functions

Traditional robustness analysis typically is viewed to have the function of providing (dis)confirmation for the attribution of particular properties or confidence in some particular prediction (Lloyd, 2010; Weisberg, 2013). While there is debate about whether traditional robustness analysis alone can play a confirmatory role (Odenbaugh & Alexandrova, 2011; Orzack & Sober, 1993; Parker, 2011), its function is either confirmation of the common result or an increase in the reliability of our inferences from modeling assumptions to that result. Even so, many actual cases of traditional robustness analysis fall short of the ideal of confirming hypotheses sufficient to warrant acceptance or belief in the hypothesis (Parker 2011).

Scope robustness analysis, on the other hand, aims to identify the set of the models that are consistent with empirical knowledge about the system. This process can result in two related outcomes: (A) the growth principles used to generate how-possibly models for some system and (B) ruling out other growth principles as possible descriptions of the system's development. Scope robustness analysis identifies the boundaries of how-possibly models among the set of examined models (A). Scope robustness analysis is one strategy for identifying how-possibly models for some system and using that information to shed light on possible organizing principles for the system.

In the process of identifying how-possibly models, researchers also identify plausible models that are inconsistent with the descriptive model and rule them out (B). More importantly, when certain growth principles produce only inconsistent models, they can rule out the growth principles used to generate them as worthy of their limited resources (financial, time, etc.). As such, scope robustness analysis also plays a role in hypothesis evaluation by directing researchers to the growth principles that are most likely to be fruitful in future research.

## 2.6 Objections

In this section, I discuss a range of objections to my argument that: (1) scope robustness analysis and traditional robustness analysis are distinct types and (2) that scope robustness analysis is truly a type of robustness analysis in general.

### 2.6.1 Why Not Regular Old Confirmation?

One might wonder why scope robustness analysis is any different than garden variety confirmation in which ruling out potential hypotheses thereby increases the confirmation of all remaining hypotheses. By ruling out some possible models of the system and the growth principles that generated them, scope robustness analysis does (slightly) confirm all the remaining possible models. Consider the similarities with eliminative inferences for a set of hypotheses in which ruling out potential hypotheses thereby slightly increases confirmation in all remaining hypotheses. According to this objection, if we accept that traditional robustness analysis has the function of providing confirmation and scope robustness can also confirm certain models, then perhaps the two types of robustness analysis are not distinct after all.

I agree that scope robustness analysis can provide confirmation by ruling out principles that cannot generate models consistent with our current evidence, but I claim that this is not the function of scope robustness analysis. Confirmation is rather a side-effect of ruling out possible competing hypotheses. When scope robustness analysis does confirm a hypothesis, it does so by eliminating competing hypotheses as principles that generate models inconsistent with our current evidence. In a typical case, it also will provide relatively small confirmational boosts compared to

the size of confirmational boosts that proponents of traditional robustness analysis have claimed (Lloyd, 2015; Weisberg, 2013).

### 2.6.2 Why Not the Optional Step in Weisberg's Robustness Analysis?

Despite the similarities, scope robustness analysis is not simply the final step of Weisberg's account of robustness analysis. First, the order of operations in Weisberg's account incorrectly describes Chen and colleagues' work. As I have argued in this article, the properties of interest of the system were already well-established empirically (i.e., the *C. elegans* wiring diagram) and so a scope robustness analysis could be performed without first having completed steps 1-3 of Weisberg's account of robustness analysis. In fact, steps 1-3 would be superfluous given the empirical knowledge about the system. Second, Chen and colleagues learned something beyond the removing of the *ceteris paribus* conditions; they learn which growth principles can produce systems with the known properties of the system and which growth principles cannot.

Now one might wonder whether a traditional robustness analysis can be performed after a scope robustness analysis. This order is possible but not necessary. For example, scope robustness analysis could be used to determine some set of possible models $M_1 \ldots M_n$ with property Q (known to be ascribed to the system based on empirical knowledge) and then a traditional robustness analysis could be conducted using the identified set of models $M_1 \ldots M_n$ to determine what predictions U are common to all those models. In this way, traditional robustness analysis can be used to complement scope robustness analysis.

In fact, this abstract case can be filled out according to the case study here. Scope robustness analysis is used to generate possible models of the developmental processes for the *C. elegans* nervous system, a subset of which are identified by those models of the developmental

processes that produce network models with the properties known from empirical work on the *C. elegans* nervous system. If researchers were to find that all of these models (with the properties similar to the *C. elegans* nervous system) also had some other properties R, then the researchers could reason from a traditional robustness analysis that the property R ought to be attributed to the *C. elegans* nervous system. Note though that this property R would not contribute to the question researchers began with: how does the *C. elegans* nervous system come to have the properties that it does? Instead, it answers research questions about what properties the *C. elegans* nervous system has.

In order to draw conclusions from the joint scope and traditional robustness analysis, however, the researchers would need to provide justification that the set of models generated according to particular growth principles were the only relevant network models to include, which would be difficult given the plausibility of other growth principles. As I mentioned previously, Chen and colleagues themselves take there to be at least one more principle guiding the development of neural systems: the tendency for neurons that process sensory information to be located spatially close to the inputting sensorimotor neurons.

### 2.6.3 Why is This Robustness Analysis?

The basic unifying theme of robustness, according to William Wimsatt (1981), are multiple determinations of something that is invariant arrived at via independent means. Although there are a number of differences in the epistemic context and practice of each type of robustness analysis, they are similar in terms of the relationship sought: invariance between a set of models and some property or predicted outcome of those models. Researchers use what they already know about the system to fix one side of the invariance relationship. Robustness analysis can then be used to fill

in the other side of the relationship. As a result of successful robustness analysis, researchers can draw justified conclusions about either the scope of possible models of the system or which hypotheses about the target system are confirmed. Which type of inference is justified differs based on empirical knowledge, but the invariance relationship is the same across both types of robustness analysis.

My account of scope robustness analysis may not appear to be a case of robustness due to some ambiguities in the use of the term 'robustness.' There are at least two general senses of robustness relevant to the current discussion of robustness analysis. In particular, we can ask which sense is relevant when answering the question: what in a particular robustness analysis is robust? The first sense of robustness concerns its magnitude. It is a judgement about a large swath of variation over which some property, inference, or phenomenon is stable. So, to say that some result of a robustness analysis is robust in the magnitude sense is to say that the result holds under a relatively large (perhaps surprising) range of varied conditions. The second sense of robustness concerns its stability. To say that some result is robust in this sense is to say that some stable invariance relationship holds over variation in other assumptions that produce the result. It need not imply that the stable relationship between some models and the result holds under a large or even moderate range of assumptions. In fact, this sense of robustness allows for there to be significant divergence of results under some particular set of assumptions (this sense of robustness is also used in the context of traditional robustness analysis, see Kuorikoski et al., 2010). Still this sort of robustness can be epistemically useful and significant, depending on the specifics of the case. My claim is that the objections that scope robustness analysis is not really a form of robustness analysis or does not describe robustness trades on the use of the first sense of robustness (magnitude of stability) rather than any stability relationship.

## 2.7 Conclusion

Robustness analysis has traditionally been understood a research strategy that helps scientists when there is uncertainty about whether particular idealizing modeling assumptions misleadingly drive common results among a set of models. The main goal is to provide confirmation of a common result, or at the very least, to increase confidence in inferences from modeling assumptions to the common result.

I have argued that robustness analysis plays a role in selecting among several hypotheses that has been obscured by the focus of philosophical accounts on confirmation and the reliability of our inferences relying on modeling assumptions. For this reason, I argue that there are two types of robustness analysis: scope robustness analysis and traditional robustness analysis. In both types, robustness analysis is used to identify an invariance relationship between a set of models and some property or predicted outcome common to a set of models, but the justified inferences that can be drawn from such an invariance relationship depend on the context of the research. However, the types of robustness analysis are distinct enough to warrant different accounts.

In the new account of scope robustness analysis presented here, I've argued that researchers are interested in questions about the development or organizational constraints on a system. Researchers use scope robustness analysis to identify possible models among a set of hypothesized models. The models that are possible gives researchers information about the developmental principles that are live hypotheses for the development of the nervous system. As the cases in this paper demonstrate, scope robustness analysis is a powerful tool for investigating which hypotheses to further study in early research programs. It becomes increasingly important, given the emphasis of many "big data" studies, to have philosophical analysis on fruitful research strategies in domains with large numbers of hypotheses but little guiding theory. Scope robustness analysis offers one

such tool for leveraging what we know about a system to guide future research in a fruitful direction.

## 3.0 How Triangulation Fails: Epistemic Risks in the Triangulation Argument for Implicit Attitudes

One important strategy for dealing with error in our methods is triangulation, or the use multiple methods to investigate the same hypothesis. There are two major success criteria in all current accounts of triangulation: (i) the methods employed need to be sufficiently diverse and (ii) the methods need to provide evidence about the same phenomenon. I address each in turn. First, I argue against one recent account of what it means for the methods in triangulation to be sufficiently diverse. Second, I argue that an account of triangulation focused on epistemic risk is better able to describe how triangulation fails and to normatively guide future triangulation research. I provide an account of triangulation that focuses on types of epistemic risk and the conditions under which it arises.

### 3.1 Introduction

With the development of new techniques, there is always a risk that their results may be due to systematic error rather than the phenomenon of interest. Consider the case of mesosomes—organelle-like structures in bacteria viewed through electron microscopes in the 1950s (Culp, 1994; Hudson, 1999). At first, researchers thought mesosomes might be a new internal structure in cells with a ribbon-like structure similar to ribosomes. However, after a series of experiments, they determined that mesosomes were observed only when the cells were prepared using certain chemical fixation techniques and not observed when those chemical fixation techniques were not

employed. So, scientific consensus ultimately settled on the claim that mesosomes were the result of systematic error introduced by the fixation techniques and not a newly discovered internal structure of cells. How did researchers in the 1980s discover that mesosomes were not in fact organelles but rather systematic errors from the process of specimen preparation? And more generally, how do scientists justify their claims about the existence of some phenomenon, given the risk that the data produced by any method could be subject to systematic errors?

One strategy is to use different experimental methods to investigate the same question—a practice called methodological triangulation. In cases where different methods converge on a common result, then researchers ought to increase their credence that the result is due to some phenomenon of interest rather than noise and error in the data. Most philosophical accounts of triangulation focus on why it is a successful and wide-spread practice. In social science, researchers might use both quantitative surveys and qualitative interviews; in neuroscience, researchers might use neuroimaging methods and electrophysiological methods; and in psychology, researchers might use self-report and informant report methods.

In this chapter, I examine the two major commitments of existing philosophical accounts of methodological triangulation. Current accounts of methodological triangulation agree on two success criteria: (i) the methods employed need to be sufficiently diverse and (ii) the methods need to provide evidence about the same phenomenon. I have two major goals in this chapter. First, I will argue that the first success criterion (i) can be understood in terms of the methods having different propensities to fail. Second, I will argue that a descriptively and normatively adequate account of methodological triangulation needs to account for the ways triangulation is susceptible to failure in its practice, especially concerning success criterion (ii), rather than focusing primarily on how and why it succeeds in ideal cases.

In the first half of this chapter, I consider existing views of the diversity criterion (i) and argue for the failure independence view. In section 3.2, I argue that the function of methodological triangulation as controlling for *unknown* systematic error and not likely or suspected systematic error. In section 3.3, I begin by providing some clarification of the primary features of methodological triangulation, including an account of both the diversity of methods and what is being triangulated upon. In section 3.4, I argue that a recent account of the diversity of methods in triangulation changes the subject.

My second claim is that in order to have an account of triangulation that's useful for scientific practice, we need a sense of the types of epistemic risks and their typical locations in the practice of triangulation. Only by identifying potential pitfalls of reasoning can we provide a descriptively and normatively satisfying and *useful* account of triangulation. In particular, by looking to insights and developments in the literature on the role of values in science, I develop an account of triangulation that can explain why in practice it can fail. This account is necessary to better understand how to improve triangulation arguments in practice, including what specific types of inferences need justified and what types of errors in reasoning may arise.

## 3.2 Methodological Triangulation

Methodological triangulation involves the use of multiple methods to examine the same research question. Current accounts of methodological triangulation agree on two success criteria: (i) the methods employed need to be sufficiently diverse and (ii) the methods need to provide sufficient evidence to accept the hypothesis (e.g., Heesen, Bright, and Zucker 2016; Munafò and Smith 2018; Soler 2012). Much of these statements must be spelled out to analyze particular cases.

What unites current accounts of methodological triangulation? Why is methodological triangulation a successful practice? When is methodological triangulation useful? I will attempt to spell these out in the next three subsections.

### 3.2.1 The Received View of Methodological Triangulation

Most current accounts of triangulation (and robustness reasoning, more generally) are cashed out in terms of its success. One view of triangulation sets out to: "identify at an abstract level the logic behind successful robustness arguments [and…] to determine what is required for a specific form of robustness analysis to be successful" (Kuorikoski and Marchionni 2016, 230). On another view, triangulation is defined as: "the use in empirical practice of multiple means of investigation to validate an experimental outcome" (Schickore and Coko 2013, 296). The philosophical accounts of triangulation primarily seek to explain why it works in the cases where it does and what conditions need to be met for triangulation to be successful in new domains.

One exception to the focus on success is the extensive discussion about what it means for triangulation to employ sufficiently diverse methods (discussed more in section 3.3). Still even here philosophical discussions tend to develop accounts of what diversity should be required for triangulation on the basis of successful cases. For example, Jonah Schupbach (2015, 275) claims: "The most important and challenging question an account of [robustness] can answer is what sense of evidential diversity is involved in [robustness analyses]." To the extent that failure of triangulation is discussed, it is primarily taken to involve the failure to have sufficiently diverse methods.

### 3.2.2 Explaining the Success of Methodological Triangulation

The reason that methodological triangulation is considered superior to single method research is that any given method may be subject to systematic errors (Wimsatt 1981; Kuorikoski and Marchionni 2016). (I will discuss systematic error more in section 3.2.3.) The more diverse a set of methods used to investigate the same question, the less likely that a single systematic error is causing all the methods to produce the same result (Cartwright, 1991; Cartwright, 1983; Salmon, 1984).

Consider a case where researchers use different methods (that are sufficiently diverse, according to your preferred view of the diversity criterion) and they get the same result from each method. What would explain the convergence of results? It could be due to error or the phenomenon of interest. Let's consider each possibility in turn. If the convergence were caused by error, then we need to posit different errors for each of the methods. Given our assumption that the methods employed are sufficiently diverse, they will be unlikely to be subject to the same types of error. So, we'd explain the convergence of results in terms of different systematic errors—one for each method used. Further, we need to posit that each different systematic error nonetheless interacts with the method's underlying processes to produce the same result. On the other hand, if the convergence were caused by the phenomenon of interest, then we need only posit a single cause to explain multiple different results. Considering both possibilities, Cartwright argues, it is more likely that there is a single cause driving the convergent results. This claim gives us reasons to prefer the hypothesis that the phenomenon of interest, rather than different systematic errors, cause the convergent results.

### 3.2.3 The Function of Methodological Triangulation

In this section, I argue that the primary function of a successful methodological triangulation is guarding against unknown systematic error when inferring to some claim. To support this conclusion, I analyze the notion of systematic error and the difficulty in controlling for it. Then I distinguish cases where systematic error is unknown from cases where researchers have plausible suspected systematic errors in mind. I argue that triangulation is a particularly important research strategy for dealing with unknown systematic error.

A systematic error (i.e., artifact, confound) occurs when some aspect of a method (e.g., protocol, instrument) causes a consistent pattern in the data and this pattern is mistaken for the phenomenon of interest. Systematic error can be distinguished from random error. Largely, for multiple experiments following similar protocols on the same population, random error (i.e., noise) can cancel out when aggregating over many experiments, depending on the distribution of the random error. That is, if we hypothesize that we will find some positive effect, in the long run after conducting many experiments, we can reasonably assume that the random error biases our aggregate results equally in both directions (positively and negatively). If the experiment has a source of systematic error, however, that error will continue to influence the results for each successive experiment in the same way. Performing the experiment multiple times will not remove the source of this systematic error and thus, will not avoid its biasing influence on the results.

Let me now motivate the importance of removing sources of systematic error from studies. Consider a recent set of neuroscience studies that sought to use information from functional connectivity to explain developmental or psychiatric differences in human populations. Some of these studies aimed to distinguish people with autism and people from normal populations on the basis of their brain's functional connectivity (Jones et al. 2010). And these studies did successfully

sort autistic and non-autistic participants. However, their success was not based on the functional connectivity of the individual participants. It turns out that the classifiers were using information produced by the motion correction processes to sort participants into the two populations. People with autism in the sample tended to move their heads more in the MRI than people from normal populations (Jones et al. 2010). We can see the same issue in some developmental studies that indicated a decrease in short-range functional connections in adult populations compared to younger populations (for review see Power, Schlaggar, and Petersen 2014). But again, the phenomenon of interest (cognitive development, in this case) was not the relevant feature used for correctly distinguishing child and adult populations. The finding of more short-range function connections in younger compared to older populations is also what one would expect to see if there is a difference in head motion between child and adult populations (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012). So, it can often be difficult to sort out systematic error as the cause of our methods.

Now I would like to make one further distinction between different types of systematic error. Researchers can be in different epistemic situations with respect to the systematic errors influencing their methods. Researchers can know what systematic errors impact their results, they may suspect possible systematic errors are impacting their results, or there may be unknown systematic errors impacting their results. Known systematic errors can be accounted for with different strategies, such as removing confounding variables from the experimental set-up. The process of identifying potential systematic errors requires both experimental skills, such as tacit knowledge of how to work with a particular instrument (Feest, 2016), and often a sophisticated understanding of the theoretical implications of an experiment. Detailing the ways that researchers identify and test for potential systematic errors has not been a focus in recent philosophy of science,

but as we will see in section 3.3.2, Schupbach (2018) provides some much needed analysis of this practice. However, researchers can never be sure they have eliminated all sources of systematic error in their methods. It is always possible that some systematic error that is yet to be articulated is driving their results. These systematic errors I call unknown systematic error.

My main claim is that a unique function of triangulation is to control for unknown systematic error. Controlling for known or suspected systematic errors need not lead to the use of triangulation as there are a number of other strategies for addressing the contribution of known systematic error; controlling for *unknown* systematic error in our methods can only be dealt with by using methodological triangulation. My argument is based on the common cause argument, which is the most widely accepted argument for the success of triangulation. Triangulation allows researchers to infer that it is unlikely that unknown systematic error produces the same result for each method. This unlikely scenario would require that each method is subject to the same systematic error unbeknownst to the researchers *and* that each systematic error produced the same result. It does so even in cases where researchers do not know much about plausible sources of error in their methods. Even though the systematic error in some methods are not known, they can still be controlled for to some extent by appealing to the degree of independence of assumptions among the set of methods. The greater the independence, the greater confidence that unknown systematic errors are controlled for. Thus, even in the impoverished epistemic situations researchers often find themselves in, gradual increases in the confidence of the detection of phenomena can be achieved by methodological triangulation.

## 3.3 Diversity in Methodological Triangulation

In this section, I consider four views of the diversity criterion for successful methodological triangulation: the assumption independence view, the failure independence view, the confirmational independence view, and the explanatory independence view. I argue for the failure independence view of diversity for triangulation. In particular, I consider and reject Schupbach's (Schupbach, 2015, 2018) arguments for the explanatory independence view because he changes the topic from triangulation to eliminative inference.

### 3.3.1 Views of the Diversity Criterion

There are numerous ways methods can be diverse. I will illustrate a few different types of diversity: diversity of the sample, diversity of the instrumental apparatus, and diversity of the theoretical assumptions. Thinking informally about diversity illuminate connections between triangulation and replication. A direct replication of some experimental protocol is completed using a different sample from the same population under study (Machery draft). In these cases, researchers are interested in whether they can find the same result in each study, despite potential random bias in the data. In cases of human studies, direct replications may simply use the same experimental protocol but vary the participants in the study within the same population. In this extreme case, the similarity in protocols means that there is no diversity in the probability of failure of the method. If the protocol fails to appropriately identify the phenomenon, then it will fail in both experiments.

In other cases, researchers produce minimal variations in the experimental protocol. For example, the instrumental apparatus might remain the same, but other aspects of the protocols are

different, such as using different chemical solutions to prepare cells for viewing under an electron microscope. As a result, ways that the cells are likely to be damaged in the process of preparation are varied across the different protocols. For these moderate cases of diversity, there will be partial independence of the probability of failure among the methods. A failure of the instrument will lead to error in all variations of the experimental protocol, but a failure in one preparation solution protocol may not indicate error in the others. Finally, two methods may use different experimental paradigms or instruments, such that each method has almost non-overlapping sets of assumptions (Hacking 1983, 201).

Now that we've seen a range of intuitive ways methods can be diverse in experimentation, we can ask what it means to have 'sufficient diversity' on accounts of successful triangulation. For example, it seems there is not sufficient diversity for triangulation among experiments that are conducted on different days of the week but are otherwise identical. More formally, there have been four views of the diversity criterion put forth in philosophical discussions of triangulation: the assumption independence view, the failure independence view, the confirmational independence view, and the explanatory independence view. I will discuss each in turn.

First, the most simplistic view is that the methods employed in triangulation need to have fully or partially independent sets of assumptions. Assumptions might be made to make mathematical equations more tractable (e.g., assuming a continuous variable is discrete) or make causal claims about how the method works (e.g., that the magnet in an MRI is sufficiently strong to align the axes of hydrogen protons and that their rotations decay at different rates depending on the type of tissue). The strongest reading of this diversity criterion requires all assumptions to be independent. On the strong version, the practice of triangulation in science would almost never meet this ideal. A more reasonable version of the assumption independence view is that diversity

among the methods should involve partial independence of their assumptions. On the weaker version, diversity is a matter of degree. It can be determined more precisely: two methods are diverse to the extent that they produce results that are unconditionally probabilistically independent.

There are two problems with the weaker assumption independence view. First, most methods—especially ones that can measure the same phenomenon—share many assumptions, such as the assumption that the methods measure features of the world and so on. In some cases, these assumptions may be shared for the purposes of mathematical tractability (Weisberg, 2007) or to make multi-scale models work together ("kludges" on Winsberg's 2010 view). In these cases, the fact that the assumptions are shared is unlikely to be the cause of the convergent results. Shared assumptions only undermine our inferences from triangulation when they could be responsible for the convergence of the results in the absence of the phenomenon of interest. Second, the assumption independence view conflicts with one claim in the robustness literature stemming from Levins (1966) and developed by Weisberg (2006, 2013; Weisberg & Reisman, 2008): that in robustness analysis, one searches for models that share a common biological assumption. It may be that we should be revisionary about the views of robustness analysis following from Levins, but I will demonstrate below that more sophisticated views of the diversity criterion have other benefits and avoid some of the problems.[5]

Another diversity criterion is the failure independence view, which holds that methods are diverse to the extent that each is prone to different systematic errors (Wimsatt 1981). I subscribe to the failure independence view for three reasons: it fits with the general account of triangulation

[5] In Chapter 2, I argued for a more general view of robustness analysis that does not assume that a common biological assumption for all models must be found.

as a strategy helpful in contexts of epistemic uncertainty, it explains why the assumption diversity criterion is *prima facie* plausible, and it is consonant with accounts of the success of triangulation. First, it retains the core of robustness, which Wimsatt held to be such an important strategy in science. Searching for convergence can give us a strategy for distinguishing trustworthy and untrustworthy results, which can be especially useful in contexts of uncertainty. The relevant uncertainty for triangulation is uncertainty about the causal processes underlying our methods and their interaction with the phenomenon of interest. According to this view, robustness in general has this function because robust practices look for invariance over models and methods that fail in different ways.

Second, it makes sense of the intuitive plausibility of the assumption independence view. When there is greater partial independence of assumptions of some set of methods, then those methods are also likely to have diversity in the ways the methods can fail. For example, the use of fMRI in cognitive science often relies on the assumption that increases in the BOLD signal (i.e., increased blood flow) is indicative of increased cognitive activity in some localized area. If this assumption were wrong under certain conditions, the use of fMRI and its associated inferences in much of cognitive science would be undermined. It would not undermine evidence produced by EEGs though. EEGs measure brain activity electrically and do not assume that changes in blood flow is a good proxy for changes in cognitive activity. However, the failure independence view avoids the issues with the simplistic assumption independence view. Particularly, the failure independence view need not claim that every difference in assumptions between methods increases the power of triangulation; for differences in assumptions that make no difference to the way the methods might err, the difference in assumptions do not increase the power of triangulation, according to the failure independence view of diversity.

Finally, the failure independence view is also consonant with philosophical arguments for the success of triangulation (Cartwright, 1983, 83-86; Salmon, 1984). In the classic case of triangulation—the estimation of Avogadro's number, Perrin does not need to know for each of thirteen methods, which method is prone to particular systematic errors. Instead, the likelihood that the methods are prone to different systematic errors is sufficient to explain the likelihood that the common cause is the phenomenon and not error.

Still philosophers have introduced other diversity criteria. Lisa Lloyd (2010; 2015) introduces and defends the confirmational independence view, which states that two methods are diverse to the extent that the increase in confirmation for some hypothesis based on one method is independent of whether or not we have previously confirmed that hypothesis with a different method (Lloyd 2009). Confirmational independence can be more formally understood in terms of conditional probabilistic independence relative to a hypothesis. Lloyd draws explicitly on Fitelson's (2001) account of mutual confirmational independence by the screening-off condition. Formally, two methods that produce the same result (R) have confirmational independence with respect to some hypothesis H when: $Pr(R_1\&R_2|H) = Pr(R_1|H) \times Pr(R_2|H)$ and $Pr(R_1\&R_2|\sim H) = Pr(R_1|\sim H) \times Pr(R_2|\sim H)$.

Schupbach (2018, 284) points out a crucial problem with the confirmational independence view. Many cases of triangulation, especially when all methods are not employed simultaneously, will not satisfy the formal definition of confirmational independence. That is, according to the confirmation function of a particular hypothesis, the $c(H, R_1|R_2) < c(H, R_1)$. Schupbach points out that often different experiments in the triangulation, when they successfully confirm the hypothesis, result in small increases of confirmation of the hypothesis for future experiments. That is, when we confirm some new hypothesis, the initial confirmational boost can be large. But the

confirmational boost for the hypothesis with subsequent experiments will be lower. So, following Schupbach, I think the confirmational independence view (as it is formulated) is often incorrect about what ought to be considered sufficient diversity among methods.

Schupbach (2015, 2018) introduced the explanatory independence view of the diversity criterion. The basic idea is that methods in triangulation are sufficiently diverse when they can rule out alternative explanations of the observed results. More formally, Schupbach's (2018, 288) proposal is that methods are sufficiently diverse:

> with respect to potential explanation (target hypothesis) H and its competitors to the extent that their detections ($R_1$, $R_2$, ... , $R_n$) can be put into a sequence for which any member is explanatorily discriminating between H and some competing explanation(s) not yet ruled out by the prior members of that sequence.

Suppose that some method produces a particular result. Researchers can most appropriately use triangulation by designing similar experiments that can test alternative hypotheses that would explain the presence of the result. So, researchers ought to design their experiments to rule out likely sources of systematic error.

### 3.3.2 Schupbach's Explanatory Diversity Criterion

In setting out his argument for the explanatory diversity criterion, Schupbach (2018) poses an objection to the failure independence criterion: it is too restrictive because it cannot account for classic examples of methodological triangulation such as Perrin's work on Brownian motion. His analysis of this case is also his major positive argument for the explanatory diversity criterion. (His other major case is the Lotka-Volterra predator-prey model, which he claims follows similar reasoning as the Brownian motion experimental triangulation case. I will set aside the case of the Lotka-Volterra model here.) In the remainder of this section, I argue that Schupbach's objection

to the failure independence criterion and argument for the explanatory independence criterion rests on the mistaken assumption that Perrin's experimental work on Brownian motion is a case of methodological triangulation. Instead, I argue that he is conflating a classic case of triangulation (i.e., Perrin's estimation of Avogadro's number) with Perrin's explanation of Brownian motion. The choice of case matters because, on my view, the two have different underlying inferential strategies. Perrin's estimation of Avogadro's number is an inductive inference following the probabilistic common cause argument (e.g., Cartwright, 1991), whereas Perrin's experimental work on Brownian motion is best characterized as an abductive eliminative inference.

Let me now explicate Schupbach's interpretation of Perrin's Brownian motion experiments. Perrin performed a series of experiments to successively rule out alternatives to Brownian motion that could explain the motion of particles in a medium. The variations included, for example, testing for the characteristic motion with both organic and inorganic particles. However, many of the assumptions and protocols for these successive experiments remained the same: the medium was water, only fine particles were used, and the same container was used. Yet, Schupbach holds that these experimental variations are sufficient to establish the existence of Brownian motion and that this conclusion is established through methodological triangulation. I will reject the latter claim.

Schupbach argues that these conclusions constitute a problem for the failure independence criterion (and other existing criteria of diversity). According to the criterion of failure independence, the Brownian motion case is not a very successful case of methodological triangulation; many of the experiments had similar protocols to the majority of other experiments. If a systematic error was found to be due to, say, the shape and material of the container, it would

be present in the majority of experiments. So, Schupbach suggests, a different diversity criterion must be at work in methodological triangulation.

I agree that the Brownian motion case would pose a problem for the failure independence criterion, if it were a case of methodological triangulation. I also agree with Schupbach that: there is some sense of diversity among the set of experimental protocols employed by Perrin; successive experiments ruled out possible alternative explanations of the data; and the set of experiments are not very diverse according to the failure independence criterion. As Schupbach notes, the case for Brownian motion is best made as a successive set of experiments aimed at ruling out particular likely systematic errors. Where Schupbach and I disagree is whether Perrin's experiments are a case of methodological triangulation. In particular, our disagreement lies in both what function methodological triangulation plays and what type of inference underlies its practice.

As I've already argued (section 3.2.3), one unique function of methodological triangulation is to control for unknown systematic error. On my view, Perrin's design of experiments does not aim to control unknown systematic error. Instead, Perrin designs his experiments to rule out suspected sources of systematic error. Each experiment is carefully designed in succession to rule out plausible alternative explanations of the observed results (including systematic error and in some cases, predictions by alternative theories). Unknown systematic error and suspected systematic error pose different problems for experimentation. When dealing with known or suspected error, one needs strategies such as abductive eliminative inferences, calibration of methods, or adjusting protocols or data to control for known confounding variables.

Rather than viewing Perrin's experiments Brownian motion as employing inductive inferences of the sort in the common cause argument (section 3.2.2), the case is better viewed as Perrin's attempt to rule out alternative explanations of the observed phenomenon through an

abductive eliminative inference. Eliminative inferences support an explanation of some result by systematically ruling out alternative explanations until only one explanation remains undefeated. In this case, Perrin's eliminative inference is an inductive inference for the existence of Brownian motion as due to the collision of particles composing the fluid and those suspended in the fluid justified by ruling out a series of plausible alternative explanations for his observations. First, Perrin rules out alternative explanations that result from plausible confounds to his method. Perrin then devises key tests of different theoretical explanations for the observed phenomenon. As more plausible alternative explanations are ruled out through targeted experimental manipulations, more support is generated for the claim that Brownian motion is caused by the collision of particles floating in the medium and composing the medium. Perrin is simply following good experimental practice in varying aspects of one's experimental protocol to determine whether the results still obtain (Franklin & Howson, 1984). Eliminative inference, but not common cause inductive inference, requires the identification of potential other causes of the observed results. As a result, eliminative inference works best in cases where researchers have fleshed out competing theories or well understood protocols. It will apply in more limited cases than common cause inductive inference, but eliminative inference, as Schupbach notices, requires much less diversity among methods.

Here it is important to note that the rhetorical force of Schupbach's framing trades on an ambiguity about which of Perrin's conclusions is the result of triangulation. Most philosophical discussion of Perrin's triangulation arguments focus on either the conclusion that atoms exist or, more specifically, the triangulation on a particular estimation of Avogadro's number. Schupbach instead changes the discussion to only a subset of these experiments: the ones on Brownian motion. In my view and from the philosophical literature, the more clear instance of methodological

triangulation is Perrin's estimation of Avogadro's number from distinct experiments (e.g., Brownian motion, x-ray diffraction, Blackbody radiation) (Cartwright, 1983; Mayo, 1996; Salmon, 1984).

In the estimation of Avogadro's number, Perrin does not need to know any likely sources of systematic error for each method. Instead, he employs an inductive inference using the common cause argument: the diversity among the methods makes it highly unlikely that there is one unknown systematic error producing the same result for each experiment. Nothing about the logic of this classic case of triangulation requires the potential systematic error to be described or even named. The importance of having tools and strategies for reducing the likelihood of unknown systematic error influencing our results cannot be underestimated.

One interesting consequence of my argument is that there are two types inferences underlying cases of robustness in experimentation—common cause inductive inference and abductive eliminative inference—but each has a different function and different procedures. First, my view is supported by Cartwright's (1983, 85) arguments that the inferential strategy underlying common cause argument is distinct from inference to the best explanation. Both historically and philosophically, the common cause argument provides the explanation for how and why methodological triangulation is an important practice. Second, eliminative inference plays a different sort of role in scientific research. Eliminative inference is most useful when there are clear competing theories or suspected experimental confounds. That is, eliminative inferences will be a less effective strategy for cases of epistemic uncertainty regarding alternative theoretical predictions or causal understanding of how some method produces data. No doubt a large part of scientific practice and experimenter reasoning falls under this description. Still it is important to

keep different research strategies separate. They are useful in different contexts and have different norms (as the work of Wimsatt and Schupbach elaborates).

To summarize the dialectic: Perrin notes that on the failure independence criterion of diversity, Perrin's arguments for an explanation of Brownian motion would not count as methodological triangulation. I agree that were the Brownian motion case an instance of triangulation, this would constitute a problem for the failure independence view. But I argue that the Brownian motion case is not a case of methodological triangulation because triangulation employs common cause inductive inferences and not abductive eliminative inferences. Perrin's experiments aim at identifying and ruling out plausible sources of systematic error and as such, are best understood as a practice based on underlying eliminative inferences. I've argued above that one unique function of methodological triangulation is to control for *unknown* systematic error. So, Schupbach's analysis of the case does not pose a problem for the failure independence view and instead, his work contributes to our understanding of the norms of eliminative inference.

### 3.4 How Triangulation Fails

In this section, I move to defend my section claim in this chapter: that a descriptively and normatively adequate account of methodological triangulation needs to account for the ways triangulation is susceptible to failure in its practice, especially concerning success criterion (ii), rather than focusing primarily on how and why triangulation succeeds in ideal cases.

A theory or account of a practice should highlight potential failures in order to be useful. Consider some ethical theory that gives an account of right and wrong actions. In order to use this ethical theory to guide my actions, I need to know not just what makes an action right or wrong,

but also some features of my moral psychology. What are the ways that I am likely to err? Should I be worried about having a weak will and lacking follow through for actions that I deem right? Knowledge of the ways in which I might err allows me to better use the ethical theory to guide my actions. Analogously, I argue that an account of triangulation that is useful in practice ought to explain not just why triangulation is successful in ideal cases, but also how it can fail in practice. To do so, I will appeal to the idea of epistemic risk from the literature on the types and roles of values in science, medicine, and technology. By identifying types of failure, this lays the groundwork for future normative work developing strategies to avoid or mitigate these risks in triangulation research.

Before providing my account of triangulation, I will first demonstrate that the backbone of existing accounts of triangulation are insufficient to explain why the practice of triangulation can fail. Current accounts of triangulation are cashed out in terms of its success.[6] One view of triangulation sets out to: "identify at an abstract level the logic behind successful robustness arguments [and…] to determine what is required for a specific form of robustness analysis to be successful" (Kuorikoski and Marchionni 2016, 230). On another view, triangulation is defined as: "the use in empirical practice of multiple means of investigation to validate an experimental outcome" (Schickore and Coko 2013, 296).

How would this received view of triangulation account for cases of failure in practice? Recall that I elucidated two success criteria for triangulation: (i) employ sufficiently diverse

[6] One exception is Stegenga (2009) who considers various problems with the use of triangulation as a strategy to deal with the problem of epistemic uncertainty in science. However, many of his critiques are not internal to the practice of triangulation. Stegenga's main concern is that philosophical accounts of triangulation provide no guidance when evidence both confirms and disconfirms the same hypothesis. But most centrally to this chapter, Stegenga does not examine the epistemic risks in triangulation arguments when they *appear* to be successful. These potential errors are all the more suspect because they masquerade as successes.

methods and (ii) provide evidence about the same phenomenon. There is substantial discussion of the failure to have sufficiently diverse methods (i), which is what Wimsatt (1981) called "illusory robustness." Still these accounts of diversity are based on successful cases of triangulation (e.g., Schupbach 2018).

We can also consider the other success criterion in triangulation: that each method produces data about the same phenomenon (ii). While most philosophers working on triangulation recognize that this is a success criterion, relatively little has been said about how researchers can *know* they have met this criterion.[7] Even less has been said about how researchers can fail to meet this success criterion.

### 3.4.1 Epistemic Risk

In order to flesh out an account of triangulation that explains how it can fail in practice, I appeal to the concept of epistemic risk, which is "any risk of epistemic error that arises anywhere during knowledge practices" (Biddle and Kukla 2017, 218). There are many types of epistemic risk that occur at different parts of the research process. The most discussed kind of epistemic risk is inductive risk (Douglas, 2016), which is particularly predominant in discussion about the role of values in science, medicine, and technology. Although the name implies it is any risk in inductive inferences, it is a technical term that refers only to specific inferences: inferences from some body of evidence to acceptance or rejection of a hypothesis.

[7] One exception is Kuorikoski and Marchionni (2016), who argue that triangulation primarily consists in justifying data-to-phenomena inferences. Relying on Bogen and Woodward (1988), Kuorikoski and Marchionni argue that researchers can use empirical reasoning to justify these inferences, such as intervening on the phenomenon to determine whether there are corresponding differences in the data. While I think their view is on the right track, it is (1) susceptible to the criticism of not explaining why triangulation sometimes fails and (2) does not provide a sufficiently developed account of the practice of triangulation. I aim to rectify these two issues here.

Following Biddle & Kukla (2017), I hold that focusing exclusively on inductive risk makes our philosophical accounts of epistemic risk deficient. There are other risks in knowledge production that do not neatly fit into the category of inductive risk. Other types of epistemic risk include the risk in deciding whether to characterize some datum as evidence for a hypothesis, such as whether some particular slide contains tumors and whether the tumors were malignant (Biddle's 2016 interpretation of Douglas 2000, 569; see also work on phenomena are constituted, Colaço 2018). Another example is risk in the inference from animal models to the target system of interest (usually in humans) as in research on exposure to bisphenol A in a particular rat model (Biddle's 2016 interpretation of Wilholt 2009). Another risk is the way a problem is implicitly or explicitly framed, including setting bounds on what counts as the problem, such as the framing of the ethical problems associated with genetically engineered crops (Biddle, 2018). Finally, there is risk in the diagnostic criteria for diseases, such as infertility (Biddle, 2016; Kukla, 2019). I discuss and expand upon the distinction between epistemic risk (the general category) and inductive risk (one type of epistemic risk) below in section 3.6.1.

Current accounts of triangulation focused on success can only account for two types of epistemic risk: the failure to have sufficiently diverse methods (or Wimsatt's "illusory robustness") and, on my view, inductive risk. Drawing on the tradition of epistemic risk laid out in Biddle & Kukla's recent work, I argue that an account of triangulation that explains failure will need to make use of epistemic risk more broadly. There are types of epistemic risk present in triangulation that do not neatly fall under either the risk of illusory robustness or inductive risk. Further, the form that inductive risk takes in triangulation is a particular kind that warrants its own exploration. That is, the inductive risk is the risk of being wrong that each method provides evidence about the *same*

phenomenon, whereas inductive risk in the science and values literature typically is merely the risk of being wrong that there is (or is not) some predicted effect.

### 3.4.2 Schema for Triangulation in Practice

In order to develop an account of triangulation that highlights points of failure, I turn away from abstract success conditions and to the details of knowledge production via triangulation. I highlight important steps in the practice of triangulation from the causal production of data to its transition to playing an evidential role to the increased or decreased credence in some hypothesis. In this section I provide a schema for the practice of triangulation.

Let me first distinguish between data and phenomena (Bogen & Woodward, 1988). Data are publicly observable reports that result from experimental or observational processes. They are not repeatable because they are the actual reports produced through experimentation or observation. Phenomena on the other hand are stable patterns in the world. Phenomena are often not directly observable and are characterized and explained by theory.

In the practice of triangulation, researchers identify multiple methods that are likely to produce data relevant to the same phenomenon. Each method may include some sources of error, such as random error from sampling or systematic error due to the instruments and procedures of the method. Unfortunately, researchers are often unaware of all sources of error in their methods. And these errors causally impact what data is produced. Yet, it is this data produced by imperfect methods that is the input for our inferential reasoning.

Here let me make a further distinction between data and evidence. Rather than thinking of evidence as a separate kind of entity, we can think of it as a role that data play in confirming or disconfirming some hypothesis. In some cases of triangulation, this step may not be trivial: when

data is produced in radically different experimental and theoretical contexts, many assumptions may be required to get from these different datasets to evidence that bears on (some particular) hypothesis. This evidential role problem is what Stegenga (2009) calls this the problem of incongruity.

Consider also that the data may be used as evidence in relation to multiple hypotheses. That is, despite of the fact that it may have been collected with some particular purpose in mind, it can serve as evidence for or against other hypotheses. In the case of triangulation, we're interested only in data that can be used as evidence for the same hypothesis. I'll focus on hypotheses about the existence of a phenomenon, though triangulation can also be used to estimate parameters and constants (e.g., Avogadro's number). At this point in the practice of triangulation, it needs to be demonstrated that all of the diverse datasets can serve as evidence for or against the *same* hypothesis. Then once the evidential role of the datasets with respect to the same hypothesis has been established, researchers can make an inference to accept or reject the hypothesis. Even if all of the datasets provide supporting evidence for the hypothesis, a judgement still needs to be made about whether sufficient evidence has been collected to accept the hypothesis.

Theory can help reduce the uncertainty for some cases of triangulation. If researchers are triangulating on a claim about the existence of a phenomenon, then they should use some theoretical characterization of that phenomenon that describes its features. Researchers need a sufficiently developed characterization of a phenomenon in order to distinguish between inferences to the phenomenon of interest from inferences to other phenomena.

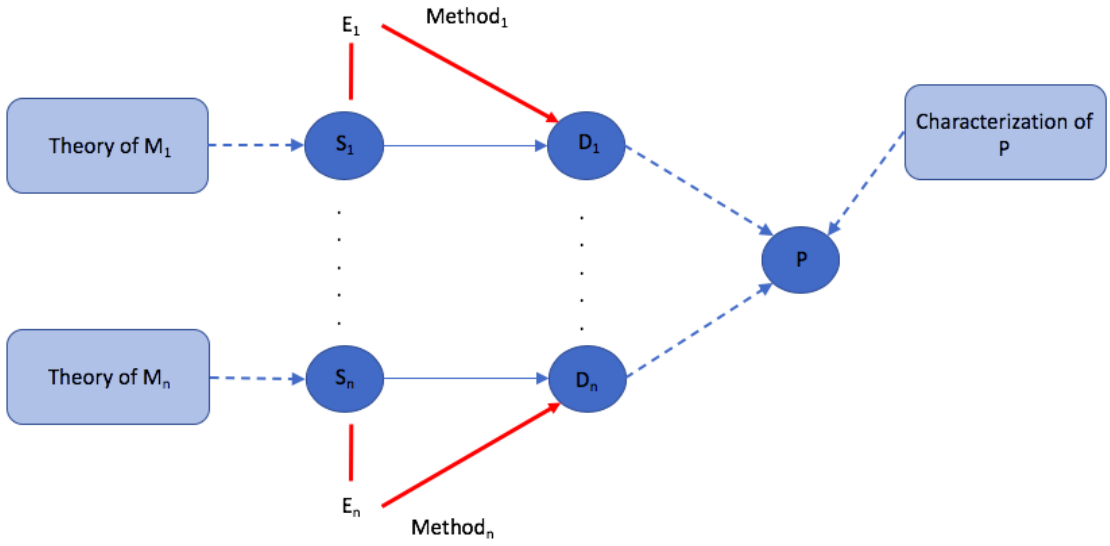


**Figure 4. Schema of triangulation.**

So, to summarize my view: Triangulation begins with the use of multiple methods to generate diverse datasets. Here researchers need to deal with the problem of error as a potential and partial cause of the data. One partial solution is to appeal to theories about how the methods work—how is that they causally produce the data—in order to identify and remove possible sources of error. Still other unknown sources of error remain and so, researchers can rely on the diversity of ways their methods can fail and the common cause argument when drawing conclusions. Researchers then need determine whether all of these diverse datasets can serve as evidence for the same hypothesis. This step will often require an appeal to assumptions about what is being measured. For hypotheses about phenomena, they need to provide a theoretical characterization of the phenomenon's features and how to distinguish it from other potential phenomena.

## 3.5 Triangulation in Implicit Social Cognition

Now that I've described the process of triangulation, I will demonstrate how it locates different types of epistemic risk. To do so, I will analyze the triangulation argument for implicit attitudes in social psychology.

By the mid-1990s, the majority of participants in psychology studies no longer self-reported holding explicitly racial attitudes (e.g., Dovidio and Gaertner 2000). In fact, many participants began to view racist acts as socially unacceptable and avoided committing racist actions themselves (Sue 2010). Yet, widespread racially discriminatory practices and racial disparities in economic, social, and health spheres persisted. Social psychologists posited that an explanation for these apparently contradictory features was that individuals still held racially biased attitudes, but that they were not reporting them when asked directly about their attitudes. So, researchers developed new techniques to control for the social desirability of appearing egalitarian (e.g., the "bogus pipeline" Jones and Sigall 1971). Indirect measures get around participants' ability and motivation to present themselves in a particular way to the researchers and instead measure their less controlled responses. As a result, researchers posited 'implicit attitudes' as a mental state or process. Implicit attitudes are automatically activated evaluative judgments about which participants are typically unaware or unable to control.

### 3.5.1 The IAT and the Evaluative Priming Task

The study of implicit attitudes bloomed. There are now nearly two dozen methods for measuring implicit attitudes. The two initial and most well-developed of these methods are the

Implicit Association Test (IAT) (e.g., Greenwald, McGee, and Schwartz 1998) and the evaluative priming task (EPT) (e.g., Fazio et al. 1986). I discuss each in turn.

During a racial IAT, participants view stimuli from four categories: two racial groups and two evaluative groups. On any trial, each racial group is paired with a different evaluative category and these pairing are displayed on either side of the display screen. On typical racial IATs, two of the categories are stimuli related to two racial groups (e.g., faces of White and Black individuals) and two of the categories are evaluative stimuli (e.g., positive and negative words). Participants are asked to quickly categorize stimuli by pressing one of two keys on the right and left sides of the display, each corresponding to the disjunctive categories listed. Researchers can compare participants' reaction times on trials in which Black-positive and White-negative are paired to those in which Black-negative and White-positive are paired. A faster response time to the latter compared to the former is thought to indicate racial attitudes that more closely link Black people with negative concepts and White people with positive concepts (e.g., Mitchell, Nosek, and Banaji 2003). There is some evidence that IAT scores are also influenced by the salience of stimuli, the perceptual similarity of stimuli, and a participant's cognitive skills (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009).

Evaluative priming tasks instead use stimuli from the categories of interest to prime participants before participants perform a categorization task on unrelated evaluative target stimuli. If researchers are interested in racial attitudes, they might use images of Black or White people to prime participants. Then during the categorization task, participants are asked to categorize positive- and negative-valence words (target stimulus). Researchers reason that reaction times on the categorization task will be influenced by the evaluative valence of the prime stimulus. If a participant holds negative attitudes towards White people, then after viewing a White stimulus

prime, they will categorize negative target words more quickly than positive target words. In general, the test-retest reliability of evaluative priming scores for the same participant is very low (Bosson, Swann, & Pennebaker, 2000), even compared to test-retest reliability of IAT scores (Lane, Banaji, Nosek, & Greenwald, 2007).

### 3.5.2 The Triangulation Argument for Implicit Attitudes

Even early in their research, social psychologists took indirect measures like the IAT and EPT to triangulate on the same phenomenon—implicit attitudes. Theories of implicit attitudes also often assume that it is a unified psychological kind (a claim discussed by Holroyd, Scaife, and Stafford 2017), though some recent philosophical theories do not (Machery, 2016). Social psychologists take indirect measures like the IAT and EPT to triangulate on the same phenomenon—implicit attitudes. Here I will offer some evidence for this claim.

Discussing the views of the field at the time in a review article on the nature of implicit attitudes, Gawronski, Hofmann, and Wilber (2006, 486; citations removed) state:

> A widespread assumption underlying the application of indirect measures is that they provide access to unconscious mental associations that are difficult to assess with standard self-report measures. Specifically, it is often argued that self-reported (explicit) evaluations reflect conscious attitudes, whereas indirectly assessed (implicit) evaluations reflect unconscious attitudes.

While Gawronski and colleagues go on to critique this widespread assumption (at least, its attribution of 'unconscious' to implicit attitudes), this quote demonstrates the ubiquitous assumption among implicit attitude researchers that first-generation indirect methods measured implicit attitudes.

More recently social psychologists have developed a neutral characterization of implicit attitudes that does not commit to any particular view of 'implicit'. This is to broadly accommodate

issues that participants are able to predict the evaluative direction of their implicit attitudes (Hahn, Judd, Hirsh, & Blair, 2014). As Greenwald and Lai write in a review article this year, "The currently dominant understanding of "implicit" among social cognition researchers is "indirectly measured." The labels "indirectly measured attitude" and "implicit attitude" are used interchangeably in this review" (Greenwald and Lai 2020). Still the assumption remains: whatever indirect measures are measuring, it is the same phenomenon.

## 3.6 Two Examples of Epistemic Risks in Triangulation

In this section, I use my account of triangulation to highlight two examples of epistemic risks and where they arise in implicit attitude research. My account better explains what goes wrong in these cases than accounts of triangulation focused on success. That is, my account provides a better descriptive account of scientific practice, where triangulation does not always succeed. Here I identify two types of epistemic risk: (1) epistemic risk when data is taken to be evidence for some hypothesis and (2) inductive risk in determining a sufficient level of evidence for the acceptance or rejection of a hypothesis.

### 3.6.1 Moving from Data to Evidence

One major epistemic risk in triangulation is that we may mistakenly think that the different datasets can serve as evidence for the same hypothesis. We are particularly at risk of this error when we do not justify the claim that our methods measure aspects theoretically related to the same hypothesis. Data do not automatically bear on hypotheses. A datum can be an image from

electron microscopy, a mark selecting an answer on a survey, or recorded video of a researcher interacting with participants. So, data needs to be interpreted in relation to the hypotheses for which they may serve as evidence. In doing this, researchers must infer on the basis of data and some assumptions to the confirmation or disconfirmation of a hypothesis.

I argue that this epistemic risk is relevant to the triangulation argument for implicit attitudes. The data produced and current assumptions in social psychology do not support the claim that the data produced by the IAT and EPT serve as evidence for the same hypothesis. In fact, according to some implicit attitude researchers, they serve as evidence for slightly different hypotheses.

In IAT studies, the categories of interest are made explicit to the participant as the categories must be identified and paired to perform the categorization task. Thus, IAT scores are thought to measure attitudes toward the general social category. Thus, they can serve as evidence for hypotheses about associations between evaluative categories and social categories.

In an evaluative priming task, on the other hand, the instructions do not explicitly determine the relevant categorical membership of the priming stimulus. It is generally accepted that due to this feature, evaluative priming tasks measure attitudes toward the stimuli rather than the category (Mitchell et al., 2003; Olson & Fazio, 2003). Consider that the priming stimulus is often an image of a person's face. Researchers may wish to contrast Black and White faces as priming stimuli in an evaluative priming task; however, for each stimulus, the individual it represents will also belong to other social categories (e.g., attractiveness, gender). Because the categorization task is only along the evaluative dimension, it is not made salient which of these categories a participant is responding to. Consider the case of a participant who when primed with a particular image of a Black face, categorizes positive stimuli more slowly than when primed with an image of a White

face. The response discrepancy could be caused by a negative evaluations of the person-represented-in-the-image's perceived race, attractiveness, perceived gender, or any combination of these and other features. Good task design will control for these differences as much as possible, but due to the design of the task, it is impossible to identify what features influence the participant's reaction times in the categorization task in any given case. Further, indirect measures generally have low test-retest validity (Bosson et al., 2000).

In order to address this epistemic risk, researchers need to provide justification for the claim that the IAT and EPT produce data that can serve as evidence for the hypothesis that participants have a negative association with the social category of interest. For the IAT, this justification already exists. For the EPT, it is less obvious. So, using my account of triangulation, I have highlighted a particular weak point in the triangulation argument for implicit attitudes and emphasized a place for the development and elaboration of norms for successful triangulation.

In section 3.6.3, I argue that this case cannot be reduced to the failure of sufficient method diversity according to different diversity criteria. Here I will argue that it cannot be reduced to inductive risk. Biddle and Kukla note the tendency to interpret inductive risk broadly to accommodate any risk during experimental knowledge production at the locus of our inferences from evidence to the acceptance or rejection of a hypothesis. This broad application of the notion of inductive risk fails to disambiguate different cases. Inductive risk in the classic sense impacts judgements about whether there are sufficient levels of evidence to accept or reject the hypothesis of interest. However, earlier on in the research process, before the production of data that can serve as evidence, researchers make many types of inferences.

Take the example of choosing an appropriate animal model for some series of experiments. Suppose researchers are interested in whether exposure to bisphenol A causes cancer in humans

(Wilholt 2009). Researchers need to choose an appropriate animal model to test this hypothesis. They will need to rely on inferences about the appropriateness of different animal models to do so, such as the inference that the relevant causal mechanism for accumulating and processing bisphenol A in humans is analogous to the causal mechanism in a rat animal model. This inference is where the risk that researchers are wrong arises. They can be wrong if the causal mechanism works very differently in the rat model than in humans and therefore, is not as sensitive to the presence of bisphenol A. This error also undermines our inferences from the evidence produced to, in this case, the rejection of the hypothesis that bisphenol A is harmful to humans. But it undermines the inductive inference from evidence to hypothesis rejection because the previous inference about the causal mechanism in rats and humans is incorrect. So, it is best to locate the epistemic risk where it first arises rather than at the locus of inductive risk, despite the fact that errors earlier in the production of evidence will impact the inferences we can make on the basis of that evidence.

### 3.6.2 Inductive Risk in Triangulation

Once we know data can serve as evidence for the same hypothesis, we can ask: How do researchers know there is sufficient evidence to accept the hypothesis? On my view, the epistemic risk of error here is best characterized as inductive risk. However, in the context of triangulation inductive risk takes a particular form. Specifically, researchers ought to be concerned about the risk of accepting the hypothesis when it is false. In cases where our hypothesis is about the existence of some phenomenon (as it often is in triangulation), there is a particular risk that unbeknownst to researchers, the data produced support the hypothesis that there are distinct

phenomena. In other words, there is an inductive risk in accepting the hypothesis that some phenomenon of interest exists on the basis of triangulation when researchers have not sufficiently ruled out the possible hypothesis that multiple phenomena are differentially driving the results.

Psychologists evaluate the validity of their tests using psychometrics. Relevant to my arguments, convergent validity is the extent to which two methods that are predicted to measure the same phenomenon are in fact measuring the same phenomenon. Low convergent validity suggests that two methods measure different phenomena. Psychologists often assess convergent validity by examining correlation coefficients.[8] If two methods measure the same phenomenon, they are expected to have high correlations in their scores. However, given that the two methods are distinct in some ways, there should not be a perfect correlation in their scores. There is no well accepted threshold for what counts as sufficiently high convergent validity. But social psychologists hold that the IAT and EPT ought to have high convergent validity (e.g., Banaji 2001).

Unfortunately, researchers have found low correlations between the IAT and other implicit measures and thus, low convergent validity (Fazio & Olson, 2003). The correlation in scores for the IAT and EPT range between r=.24 and r=.13. These are very low positive correlations. So, a participant's score on the IAT provides very little information about their EPT score, and vice versa.

One possible cause of the low correlations between IAT and EPT scores is the low reliability of EPT (De Houwer et al., 2009). Perhaps the scores do not correlate well due to noisiness in the data produced by unreliable methods rather than the methods measuring different

[8] Other methods such as the multi-trait multi-method matrix (Campbell & Fiske, 1959) have been used less frequently and less completely in the context of implicit attitudes.

phenomena. A recent comparison of seven indirect measures of attitudes Bar-Anan and Nosek (2014), the EPT had weak correlations with other indirect measures (including the IAT, r=.24).

However, there are two reasons to remain neutral with respect to these explanations. First, as Bar-Anan and Nosek (2014, 677, original emphasis) suggest, low convergent validity and low reliability may *both* contribute to the low correlations of scores on indirect measures of attitudes:

> the most likely explanation for this pattern, coupled with the similar rank ordering for internal consistency, is that [Affective Misattribution Priming] and EPT are both relatively distinct, and *also* less effective in reliably assessing the target evaluation than are the other measures. […] it could still be the case that both measures assess unique components of evaluation that are not assessed by other indirect measures (including each other).

Still one promising finding is that unlike the Affective Misattribution Priming task, Bar-Anan and Nosek (2014) do not find a strong correlation between the EPT and direct measures of racial attitudes (i.e., self-report on surveys), which would have indicated the potential influence of deliberate evaluation in the indirect measurement. There are further reasons to be concerned that the contribution of low reliability and low convergent validity cannot be distinguished for indirect measures. Schimmack argues that low reliability is indicative of validity problems. Schimmack (2019, 5) states: "even if low convergent validity is caused by low reliability, it poses a problem for the validity of the IAT as a measure of individual differences." He argues that the reliability of a measure sets the upper bound for its validity. Having low reliability thus constrains the highest possible validity for a measure. So, while some of the low correlations between the measures may be due to the low reliability of the EPT, it is likely that both low reliability and low convergent validity cause the low correlation among indirect measures of implicit attitudes.

### 3.6.3 Why Can't These be Understood as a Failure of Diversity?

One potential objection to my claim that current accounts of triangulation cannot sufficiently capture what is going on in the implicit attitude cases appeals to accounts of sufficiently diverse methods. On this objection, the IAT and EPT are not sufficiently diverse methods, unbeknownst to researchers, and thus the failure to triangulate in this case is explicable on existing success-focused accounts of triangulation. The basic idea is that whatever diversity criterion we accept (see Schupbach 2018), the IAT and EPT are too similar to count as distinct methods for the purposes of triangulation. I respond to this objection in two ways: first, by clarifying that these methods are historically descendant from different theories in psychology and second, by arguing that on our best understanding of the mechanisms underlying the measurement tasks, the IAT and EPT measure slightly different psychological processes. In addition to my arguments that they produce data relevant to different hypotheses (section 3.6.1), this gives us some reason to think the methods are sufficiently diverse on any appropriate diversity criterion.

The two methods I discuss were developed out of different historical traditions in psychology (Payne & Gawronski, 2015). Drawing on Shiffrin and Schneider's (Shiffrin & Schneider, 1977) work on selective attention, Fazio and colleagues (Fazio, Jackson, Dunton, & Williams, 1995) developed the evaluative priming task to distinguish automatic and controlled processing. Controlled processing requires attention and can be altered voluntarily, whereas automatic processing takes place on memories stored in long-term memory, is automatically activated given the appropriate inputs, and is difficult to suppress.

Greenwald and Banaji's (1995) work on implicit attitudes came out of cognitive psychological research on implicit memory, which describes the way that earlier experiences can influence current performance on learned tasks without conscious awareness of the past

experiences. Most famously, the patient H.M., who had a medial temporal lobectomy and thus lacked bilateral hippocampi and other structures, was unable to create new episodic memories. However, H.M. demonstrated the formation of new implicit memories through the time-savings in relearning motor skill tasks (Corkin, 2002). As Greenwald et al. (1998) constructed it, the IAT is a measurement of implicit memory. So, both measures were designed based on different theories. In short, the evaluative priming task was designed to measure a construct that is typically uncontrolled or automatic while the IAT is designed to measure a construct that is typically unconscious or about which the individual is unaware.

They are also causally distinct, though there is a caveat to this claim that it is based on current evidence and is defeasible. It is generally unclear what mechanisms underlie the racial IAT and evaluative priming tasks. Many mechanisms for the IAT have been proposed, but little evidence bears on their plausibility. Often what evidence exists is consistent with multiple mechanisms. However, for the sake of this chapter, I will outline one proposed mechanism for each method to motivate the claim that, given the state of current research, the racial IAT and the racial evaluative priming task likely measure distinct phenomena; affect plays a more significant role in the underlying mechanism of the latter than the former.

While there is no one well-accepted mechanism underlying the IAT (De Houwer, 2001; Greenwald et al., 1998), I will focus on the familiarity account. Rothermund and Wentura (2004) argue that the IAT task is simplified by participants when they only focus on one of the categories and thus turn the task into a decision task. Whichever category is more salient (or familiar) will have the stimulus that captures attention. Then that category is the "figure" category (rather than the "ground" category) and is used to make judgements about which key to press. Kinoshita and Peek-O'Leary (2005) endorse Rothermund and Wentura's view and further argue that in general

the more familiar category will be the "figure" and the less familiar categories the "ground." So, in the case of the racial IAT, typically White participants will demonstrate a pro-White bias on IAT because their familiarity with White people and the positive valence of the category White people. According to the salience/familiarity mechanism, the IAT includes affect to a lesser extent than other indirect measures.

Contrast the role of affect in the IAT with its role in evaluative priming tasks. The mechanism for evaluative priming tasks is response priming. The prime stimulus automatically readies a response to the target stimulus. The response can be congruent or incongruent to that required during that categorization task towards the target stimulus. When the two responses are congruent, the prime stimulus has already readied the appropriate response for the target stimulus, and thus, responses are faster and there are fewer errors. In the case of incongruent response preparation, that initial response must be aborted and the alternative response prepared, which increases response time and errors (Wentura and Degner 2010; De Houwer et al. 2009).

Much of Fazio's research into evaluative priming has coincided with research about emotional states. After all, the association between target and prime is supposed to transfer affective responses from the prime stimulus to the target stimulus in the evaluative judgement task. For example, Fazio and Hilden (2001) are interested in the way these attitudes feed into emotional states like guilt or conflict-avoidance. In one experiment, they showed participants a commercial that depicts a photograph of a Black man and revealed text describing a scenario about the apprehension of a criminal. The text was revealed in a way to mislead the viewer to interpret the Black man as the criminal rather than the police officer. Feelings of guilt from watching the commercial were also associated with more positive automatic attitudes towards Black people. Given the prominence of motivation and the

link to negative emotional states, it is clear that affect is a prominent component or result of evaluative priming.

According to the different diversity criteria, the IAT and EPT count as sufficiently diverse. Suppose we accept the failure diversity criterion. There are methodological differences between the IAT and EPT, namely, the number and types of dimensions along which participants must categorize. This methodological difference makes the IAT more susceptible to the influence of researcher demand effects compared to the EPT. So, the IAT scores could fail to track the phenomenon when researcher demand effects are present, whereas EPT scores would be unaffected. Suppose instead that we held the assumption diversity criterion. Here there are a few differences in assumptions between the IAT and EPT. In particular, there are assumptions about what is being measured from the theoretical traditions preceding the development of each method and some features of the causal mechanism underlying what is measured. First, the IAT assumes that the attitude being measured is unconscious to the participant due to its development from work on implicit memory. The EPT does not assume participants are unaware of the direction or content of their attitudes. Instead, it assumes that participants cannot control the impact of their primed responses during the categorization task. Similarly, the IAT does not assume that a participants' sensitivity to certain emotions (such as guilt) would impact their scores, whereas the EPT as a measure of affective responses from the prime stimulus would be impacted if participants were particularly prone to emotions like guilt.

## 3.7 Conclusion

In this chapter, I argue that Schupbach's explanatory independence view of the diversity criterion in triangulation is unsuccessful because his argument changes to topic. To do so, I distinguished between two epistemic positions researchers can have with respect to the systematic error in their methods: known or suspected systematic error and unknown systematic error. Then I argue that triangulation is uniquely able to guard against unknown systematic error, whereas eliminative inference is one strategy for identifying and controlling the influence of suspected systematic error. Schupbach's explanatory account describes how to design future experiments in the context of an eliminative inference argument, but it is distinct from the common cause inductive inferences underlying methodological triangulation.

I have also provided an account of triangulation that highlights locations and types of epistemic risk. In particular, I diagnosed two epistemic risks in implicit attitude research: (1) the risk that data do not serve as evidence for the same hypothesis, and (2) the particular inductive risk that there is insufficient evidence provided to conclude that there is a single phenomenon (given the plausibility of alternative hypotheses positing multiple phenomena). Neither is sufficiently described by illusory robustness and (1) is not a case of inductive risk either. Finally, I demonstrated that current accounts of triangulation focused on successful cases cannot provide explanations of why triangulation sometimes fails in practice and thus, do not develop sufficient norms to guide future triangulation research.

## 4.0 Topological Explanation in Neuroscience: From Network Models to the Brain

Some mechanists (Kaplan, 2011; Kaplan & Craver, 2011) hold that all explanation in neuroscience is mechanistic. Other philosophers of science have argued that topological explanations, in which the behavior of a system is the mathematical consequence of the topological properties of that system, are an alternative explanatory strategy. I argue topological explanations are both non-mechanistic and employed in neuroscience. In particular, network neuroscientists can explain the robustness and vulnerability of the macroscale human brain to perturbation by appeal to the topological properties of small-worldness and modularity. This explanandum is an example of functional robustness, which holds that some functions can be retained in a system over variation in its underlying parts and activities. As such, I argue it is a prime candidate for a non-mechanistic explanation because details about the underlying parts and their activities are not relevant. I consider objections that the case is either: (i) not explanatory or (ii) mechanistic. I conclude that mechanists such as Kaplan and Craver should deny the claim that all explanation in neuroscience is mechanistic. There are two benefits for mechanists taking this approach: (1) they can retain their commitment to difference-making and asymmetry as norms of explanation in general and (2) they can retain the distinctive norms of mechanistic explanation in particular.

### 4.1 Introduction

Some mechanists (Kaplan, 2011; Kaplan & Craver, 2011) have recently argued for the wide scope of mechanistic explanation in neuroscience—the claim that all explanation in

neuroscience is mechanistic. They claim to allow space for non-mechanistic explanations, but in practice these mechanists attempt to accommodate all alleged counterexamples raised against the wide scope claim. To defend the wide scope claim from these counterexamples, mechanists adopt one of two strategies: (i) deny that the cases are explanatory by appealing to the difference-making and asymmetry norms of explanation or (ii) deny that the cases are non-mechanistic by relaxing the distinctive norms of mechanistic explanation.

In this chapter, I present evidence that neuroscientists sometimes use explanatory strategies that are not mechanistic. Some of these strategies can be characterized as topological explanations (a specific type of mathematical explanation), in which the behavior of a system is mathematical consequence of the topological properties of that system. I argue that the current use of network models in neuroscience sometimes provides topological explanations. In order to account for this fact, I argue that mechanists should reject the wide scope claim and allow room for topological explanation as an alternative form of explanation in neuroscience.[9]

One particular test of the wide scope claim is whether some robust phenomena in neuroscience are better explained by topological explanation than mechanistic explanation. It is intuitive to think that mechanistic explanation will not cope well with some types of robust phenomena, where the spatio-temporal locations of parts and activities changes under different initial conditions.

In Section 4.2, I discuss the mechanists' view that all explanation in neuroscience is mechanistic as well as the two typical mechanistic responses to alleged counterexamples. In Section 4.3, I introduce Huneman's (2010, 2015) account of topological explanation and recent

[9] Some mechanists like Bechtel (2015c) already reject the wide scope claim, though in other work he tends to accommodate potential counter-examples to the wide scope claim under mechanistic explanation by expanding the norms of mechanistic explanation (Bechtel, 2013, 2015a).

clarifications by Kostić (2012, 2020). Then in Section 4.4 I argue that the account of topological explanation clearly describes some cases of explanation in network neuroscience. Finally, in Section 4.5 I present possible objections that mechanists might make to my interpretation of the case and argue that their responses are deficient for the case presented. Further, these responses would either require denying norms of explanation that they accept or diminishing the distinctive norms of the mechanistic account of explanation. Neither outcome is desirable for the mechanistic account of explanation, so I conclude that mechanists like Kaplan and Craver should give up the wide scope claim instead.

## 4.2 Mechanistic Explanation

The canonical definition of a mechanism comes from Machamer, Darden, and Craver: "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al., 2000, 3). In mechanistic explanation, describing how the underlying entities and activities (in some particular organization) bring about the phenomenon provides an explanation. Initially, mechanists were cautious about the scope of mechanistic explanation. Machamer, Darden, and Craver (2000) express hope that the picture of mechanistic explanation can be expanded to scientific fields other than cellular and molecular neuroscience, but do not predict that it will account for all explanatory practices. However, more recently some mechanists have expanded the scope of mechanistic explanation to areas beyond cellular and molecular neuroscience (Bechtel & Abrahamsen, 2010; Craver, 2016; Kaplan & Craver, 2011).

Kaplan and Craver suggest that mechanistic explanation may accommodate all explanation in systems and cognitive neuroscience: "we articulate and defend a mechanistic approach to explanation for dynamical and mathematical models in systems neuroscience and cognitive neuroscience. Such models…carry explanatory force to the extent, and only to the extent, that they reveal (however dimly) aspects of the causal structure of a mechanism" (2011, 602). Kaplan and Craver claim that their stance is not "imperialistic" because some instances of explanation may be non-mechanistic. However, they list only physics and folk psychology as domains in which mechanistic explanation may be inappropriate. Both in print and in practice, Kaplan and Craver endorse the wide scope claim—that all explanation in neuroscience is mechanistic.[10]

Some philosophers have pushed back against the widened scope of mechanistic explanation (e.g., Silberstein and Chemero 2013; Ross 2015; Chirimuuta 2014). The mechanists' responses to these alleged counterexamples follow two lines of response. The mechanists either argue that (i) the purported counterexample is explanatory but also mechanistic and so it does not actually counter the wide scope claim or (ii) the purported counterexample is not actually explanatory and so lies outside the scope of theories of explanation.

Concerning network models, Zednik (2014, 18) defends a form of response (i). The idea is that network models can provide a non-mechanistic explanation stems from a misunderstanding about what is required for an explanation to count as mechanistic. Network models highlight the functional or structural organization of a system and according to Zednik, the emphasis on organization is sufficient for network models to provide a (however incomplete) mechanistic

[10] See also Piccinini and Craver (2011, 292) and Zednik (2014, 16-7).

explanation.[11] In that sense network models might contribute to mechanistic explanations despite deviating from the canonical definition by failing to represent components and their activities.[12]

Craver & Kaplan (2020) rely on the idea of a mechanism sketch to explain when some detail about a mechanism is unknown, but a model refers to at least some internal detail about a mechanism. The problem is the assumption that mechanism sketches are ultimately to be fleshed out in a mechanistic explanation. While it may be descriptively accurate to say that mechanistic details are identified through experimental practice or that models are refined to include further mechanistic details, that does not necessarily imply that the original model can only contribute to mechanistic explanations. The use of the "mechanism sketch" model could be used in other types of explanations or it may provide a satisfactory explanation on its own. In these cases, calling the model a 'mechanism sketch' gives the impression that its only use is to be later turned into or accommodated by a complete mechanistic explanation. This assumption is imperialistic.

Kaplan and Craver (2011, 602) offer a version of response (ii). Although Kaplan and Craver suggest that mathematical models in cognitive and systems neuroscience may be integrated into the project of providing mechanistic explanations, they claim the models themselves do not provide explanations. Kaplan and Craver (2011, 623) acknowledge that descriptive models can improve our understanding of the phenomenon to be explained and act as an important precursor to experimentation, but on their view, this does not amount to explanation proper.

[11] In some places, Kaplan (2011, 347) endorses something close to this response: "Far from requiring a perfect correspondence or isomorphic mapping between model and mechanism, 3M requires only that some (at least one) of the variables in the model correspond to at least some (at least one) identifiable component parts and causal dependencies among components in the mechanism responsible for producing the target phenomenon."

[12] Levy and Bechtel (2013) make a similar point about network motifs. Abstraction from the structural properties of the components is compatible with mechanistic explanation insofar as causal relationships in the system are represented.

### 4.2.1 When Do Details Matter?

One major barrier to providing an explanation for a robust phenomenon would be if the mechanistic account of explanation did not allow for abstraction or the elimination of irrelevant detail in an explanation. The majority of authors in the wide scope debate agree that an explanation of a robust phenomenon ought to include all and only the relevant details (e.g., Chirimuuta, 2014; Craver & Kaplan, 2020; Kaplan & Craver, 2011; Ross, 2015). The main debate between defenders of the wide scope claim and critics is about the sense of 'relevance' in this claim.

According to Kaplan (2011, 347), on the mechanistic account of explanation "a model carries explanatory force to the extent it reveals aspects of the causal structure of a mechanism, and lacks explanatory force to the extent it fails to describe this structure." Kaplan articulates this claim more precisely in the model-to-mechanism-mapping constraint on mechanistic explanation:

> (3M) A model of a target phenomenon explains that phenomenon to the extent that (a) the variables in the model correspond to identifiable components, activities, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these (perhaps mathematical) variables in the model correspond to causal relations among the components of the target mechanism.

More recently, Craver and Kaplan (2020, 297) revisit the 3M principle to be a relative notion of explanatory completeness with respect to one phenomenon compared to another and define 3M*:

> A constitutive mechanistic model has explanatory force for phenomenon P versus P' if and only if (a) at least some of its variables refer to internal details relevant to P versus P', and (b) the dependencies posited among the variables refer causal dependencies among those variables (and between them and the inputs and outputs definitive of the phenomenon) relevant to P versus P'.

In condition (a), 'internal' refers to details that are internal to a phenomenon rather than external details, which rules out details related to eliciting conditions and so on. It serves to distinguish norms for mechanistic explanation (which is both causal and constitutive) from the norms of

aetiological causal explanations. The biggest difference between 3M and 3M* (besides making it a comparative criterion) is that condition (a) has been significantly weakened. Many took the condition (a) from 3M to specify that isomorphism is required, though Kaplan (2011, 347) clarifies 3M is not intended to imply isomorphism. Condition (a) from 3M* clarifies that not every variable needs to correspond to constitutive elements of the mechanism and thus allows some variables to be black-boxes. 3M and 3M* provide a kind of hierarchy of the extent to which models contribute to an explanation: the more *relevant* details provided by the model, the more explanatory power. So, what details are relevant to a mechanistic explanation?

Craver and Kaplan claim that 3M* implies that only some details are necessary for any constitutive explanation, namely:

> Some Details Are Necessary (SDN): A putative constitutive explanation for P versus P' has explanatory force for P versus P' only if it constitutively describes some of the entities, activities and organizational features relevant to P versus P'.

They propose SDN as primarily a clarification of a norm for any (constitutive) explanation (as opposed to aetiological explanations). It cannot be a norm of causal explanation in general because, as they mention, some aetiological causal explanations do not involve constitutive relations. However, SDN is very permissive—intentionally so—to allow "that even models describing very abstract details…cf. (Batterman and Rice 2014)… can count as explanatory" (Craver & Kaplan, 2020, 298). Thus, even the most recent work by Craver and Kaplan demonstrates they think the majority explanations are either mechanistic (constitutive) or aetiological. We will see later in section 4.5.6, Craver and Kaplan's recent account of 3M* and SDN provide reason to agree that there are genuine counterexamples to the wide scope claim.

## 4.3 Topological Explanation

One promising alternative explanatory strategy is topological explanation (Huneman, 2010, 2015; Jones, 2014; Kostic 2020). While mechanistic explanations explain in virtue of the productive organization of components and activities of a system, topological explanations explain in virtue of the topological properties of the system. Topological properties are general properties of some system's non-spatial organization, concerning the connectedness of elements in the system. Huneman (2010) introduces the idea of topological explanation to account for explanations that demonstrate the mathematical entailment of some pattern, behavior, or property of the system on the basis of its mathematical structure. Huneman cites a wide range of scientific domains that provide topological explanations, including ecology, molecular biology, evolutionary biology, and the social sciences. He speculates that there may be topological explanations in neuroscience as well but does not discuss any cases. I will clarify Huneman's characterization of topological explanation and argue that some explanations in neuroscience are topological explanations, and thus are counterexamples to the claim that all explanation in neuroscience is mechanistic.

Let me now describe the network approach to explanation. A classic problem in network science is the "problem of the seven bridges of Königsberg" described by Leonhard Euler (1741). In this problem, Euler describes the layout of the seven bridges of Königsberg. The task is to explain why one cannot cross all seven bridges only once (an *Eulerian path*). Euler explains that one's particular starting point and choice of path is irrelevant to solving the problem. Instead, to answer why one cannot cross all seven bridges in Königsberg without crossing at least one bridge twice, Euler provided a mathematical explanation of the impossibility of an Eulerian path based on a network representation of the bridges. To do so, he represents Königsberg as an abstract network of connected nodes and edges. Nodes represent the landmasses and edges represent the

bridges. Their connections provide a structure of possible ways one could travel to the different landmasses. Each landmass has a different number of bridges connecting it to other landmasses, which can be quantified as the degree for each node (i.e., the number of connected edges for a node).

Euler reasoned that if the degree of each node (number of edges adjacent (or connected) to the node) were even, then an *Eulerian path* would be possible. However, in the network representing the Königsberg bridges, each node has an odd degree. Given the number of nodes with odd degrees, it is a mathematical consequence that an Eulerian path is impossible. It is the topological properties of the network, namely its organization as represented in the network model, that entail that an Eulerian path does not exist.

### 4.3.1 An Account of Topological Explanation

The problem of Königsberg's bridges has been seen as an exemplar of what is now called topological explanation.[13] An explanation is topological when the explanandum, some behavior or property of the system, is shown to be the mathematical consequence of some of the system's topological properties. Consider Batterman's (2002) type-two explanatory question: why do distinct systems all exhibit the same pattern of behavior? The answer provides a topological explanation, roughly, when distinct systems exhibit the same behavior because they all have the same topological properties. These answers are also contrastive: the trait, property or propensity

[13] Pincock (2007) argues that the problem of Königsberg's bridges is solved by giving a structural explanation. His account of structural explanation can be seen as a predecessor of topological explanation. Abstract explanations provide explanations by mapping elements of a physical system into a mathematical space. While this realization is a crucial part of a topological explanation, it is not sufficiently characterized.

to behave in a certain way is because the systems have the same topological properties rather than some other topological properties (more topological properties are discussed in section 4.4.1).

Following Huneman (2010), I characterize the steps of topological explanation as:

(1) The topological properties X (as opposed to topological properties Y) entail the trait, property, or propensity for behavior Z.
(2) The real system realizes a network model with topological properties X (as opposed to topological properties Y).
(3) Therefore, the real system exhibits the trait, property, or propensity for behavior Z solely in virtue of its topological properties X (rather than Y).

Topological explanations need to show that in virtue of the relationship between the network model and the real system, some behavior or property of the real system can be explained.[14] For Huneman, abstract (or mathematical) realization ("A-realization") plays this role. A topological property of a network model can be attributed to the real system when the real system A-realizes the network model. Huneman relies on Gillett's characterization of A-realization: X A-realizes Y when X maps onto or is isomorphic to Y.[15,16] Realization occurs when there is

[14] Kostić (2020) argues that there are two types of topological explanations: vertical and horizontal. This distinction relies on a difference in topological properties: they can be properties of the network ('global') or properties of particular nodes ('local'). However, it is also important to keep in mind that some local properties may rely on properties about the organization of the network as a whole, such as the centrality of a node in a network. Vertical topological explanations are when "A describes a global topology of the network, B describes some general physical property, and had A had [sic] not obtained, then B would not have obtained either" (Kostić 2020, 2). A horizontal topological explanation is when "A describes a set of local topological properties, B describes a set of local physical properties, and had the values of A been different, then the values of B would have been different" (Kostić 2020, 2).

[15] A-realization is importantly different from what Gillett calls M-realization, where the relationship between X and Y involves metaphysical constitution.

[16] According to Kostić (2018), there is a particular kind of realization in topological explanation that is distinct from existing constitutive accounts of realization. Kostić (2018, 87) calls it topological realization when: "The realization relation stands between a topology T and a system S, such that the system S realizes topology T when the *elements* of S are interconnected in ways that display the pattern of connectivity characteristic of T." Kostić clarifies that elements here refers to either spatially-located parts of a target system or aspects of an abstract representation of data. One major difference between topological realization and constitutive realization is what can stand as a realizer of the target system. For topological realization, the realizer is at the global level of the topology of a system. For constitutive realization, the realizer is at the level of parts and activities. According to Kostić's view, the topological realization itself is not explanatory. Rather for topological explanation the "explanatory relation stands between the topology and its mathematical consequences" (Kostić 2018, 88).

isomorphism between elements of the real system and the topological network model (Gillett, 2010).

## 4.4 Network Models in Neuroscience

I aim to show that neuroscientists do sometimes give topological explanations using network models and thereby expand the scope of Huneman's account of topological explanation to neuroscience. However, I do not endorse the claims made by some neuroscientists (e.g., Seung 2012) that the connectome—a network of all neural connections in the brain—is a privileged level of analysis for the brain, nor the idea that topological explanation is the only kind of explanation in neuroscience.

Network neuroscientists are interested in the topological properties that can be attributed to the macroscale brain as a result of creating network models. In macroscale neuroscience, researchers consider primarily functional or structural networks, neither of which represent causal interactions between brain areas.[17] In functional networks, the researchers use neuroimaging to determine pair-wise correlations above a certain threshold to create the edges in the network. Nodes are often based on single voxels or prior regions of interest (or sets of voxels).[18] However, with some instruments, such as electrical encephalogram, nodes may be defined as the area of brain activity measured by the position of the electrodes in the extracranial cap. In structural

[17] While there are some attempts to create effective connectivity network models (where edges are causally directed) such as the Granger causality test (e.g., Knight 2007; Kim et al. 2011), these methods are not well accepted, in part because Granger causality cannot rule out potential common causes (e.g., Witt and Meyerand 2009).
[18] Voxels are a value on a regular grid in three-dimensional space, commonly representing a 3x3x3 cm brain area in neuroimaging data.

networks, researchers use predetermined anatomical atlases to define the nodes in the network and use tractography method or probabilistic measures of water diffusion down white fiber tracts (e.g., Diffusion Tensor Imaging) on magnetic resonance imaging data to determine the edges.

### 4.4.1 Network Basics

Network models are composed of nodes that are connected by edges, which determine the organization, or 'topology', of the system as a whole. The organization of the nodes and edges has first-order properties such as the path length or the clustering coefficient. In a network, the path length is the lowest number of edges that must be traversed to get from one node to another. The average path length of any network is determined with the following equation:

$$L_{net} = 1/(n \bullet (n\text{-}1)) \bullet \Sigma d(v_i, v_j)$$

Given a network with $n$ edges, $d(v_1, v_2)$ represents the shortest path length between $v_1$ and $v_2$. If there is no path between $v_1$ and $v_2$, then $d(v_1, v_2) = 0$. The average path length of a network of interest, compared to a random network, can be calculated with this formula: $\lambda = L_{net}/L_{ran}$, where $L_{net}$ designates the average path length of the network and $L_{ran}$ designates the average path length of a random network with a similar number of nodes and edges.[19]

The clustering coefficient of a network is the extent to which nodes form clusters with nearby nodes. When a network has a high clustering coefficient, it is likely that when two nodes are both connected to a third node, the first two nodes are also connected to each other. The average clustering coefficient of any network model can be determined by the following equation:

[19] For more discussion about the comparison of path length of a network of interest to a network representing the null hypothesis, in this case a random network, see Hadi Hosseini and Kesler (2013).

$C_{net}$ = (number of closed triplets)/(number of open and closed triplets)

Open triplets are defined as three nodes that are connected by two undirected edges and closed triplets are defined as three nodes that are connected by three undirected edges. The clustering coefficient of some network of interest, compared to a random network, is calculated with the following formula: $\gamma = C_{net}/C_{ran}$, where $C_{net}$ designates the average clustering coefficient and $C_{ran}$ designates the average clustering coefficient of a random network. Regular (or lattice) networks, in which each node is connected to nearby nodes, have high clustering coefficients and high average path lengths. Random networks, which have random connections between nodes, have low clustering coefficients and low average path length.

The modularity of a network is a measure of the extent of the division of a network into modules.[20] Networks with high modularity compared to random network models appear to have densely connected clusters with relatively sparse connections to other clusters. The modularity of a network can be defined as the fraction of edges that fall within a cluster minus the expected fraction, given a random distribution of edges. To calculate modularity, a network model is compared to a random network model where the degree of each node is preserved. Formally, the modularity of a network ($Q$) can be calculated with the following formula:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v \times k_w}{2m}\right]\frac{(s_v \times s_w) + 1}{2}$$

In the equation, $m$ designates the total number of edges, $A_{vw} = 1$ or $A_{vw} = 0$ designate the presence or absence (respectively) of an edge between nodes $v$ and $w$ in the network, $k$ designates for node degree, and $s_v = 1$ or $s_v = -1$ designate that node $v$ belongs to module 1 or module 2 (respectively)

[20] Note this is distinct from the notion of modularity in debates about cognitive architecture (e.g., Fodor, 1983), though see Colombo (2013) for an argument that the network science view of modularity is helpful in the former.

(Newman, 2006). For more than two communities, the formula can be run iteratively, or more complex formulae can be used.

The organization can also have second-order properties, which are the result of first-order properties.[21] One such property—small-worldness—was described by (Watts & Strogatz, 1998). Within small-world networks, each node can be reached from any other node in the network by traversing only a few edges. According to the typical definition of small-worldness, small-world networks have a high clustering coefficient but short path length. Small-worldness can be calculated with the following: $\sigma = \gamma/\lambda$.[22] For most biological networks, the path length is around 1 and clustering coefficient is between 2 and 3 (Bassett & Bullmore, 2006).

Another second order property—scale-free—involves the degree distribution of a network. The degree of a node ($k$) is determined by the number of edges adjacent or connected to it. Researchers can then characterize the degree distribution of the network by determining the distributions of degrees of all the network's nodes. According to the typical definition of scale-freeness, networks are scale-free when they have degree distributions that follow a power law, $Pr(k) \sim k^{-\alpha}$. These networks are called 'scale-free' because the degree distribution has a very gradual power law decay that indicates the network lacks a characteristic scale and so the network "looks the same" at every scale (Bullmore & Sporns 2009).

[21] Both first-order and second-order properties are called topological properties in Huneman's (2010) terms since they are likely to be represented in a graph or network.
[22] For other definitions of 'small-worldness', see Humphries and Gurney (2008) and (Muldoon et al., 2016).

### 4.4.2 From Network Models to the Brain

I propose that neuroscientists use network models to explain at least the following explanandum: *the macroscale human brain's distinct pattern of functional robustness and vulnerability to damage*. Let me first provide some necessary background on the relevant sense of robustness in this explanandum. Then I will describe the empirical evidence supporting this pattern of robustness and vulnerability in the macroscale human brain.

Functional robustness is "a property that allows a system to maintain its functions despite external and internal perturbations" (Kitano, 2007, 826). Functional robustness can be defined with respect to particular functions or to general functioning. Retaining specific functions, such as the timing of initiating DNA replication in yeast, can occur by having heterogeneous redundant mechanisms, such as the *Clb5* and *Clb6* genes (Kitano, 2007, 828). Specific functions can also be maintained by modularity, in the sense that a system can be decomposed into relatively distinct functional units. Often in cognitive and clinical neuroscience, researchers are interested in the processes, maintenance, and breakdown of particular cognitive functions.

However, in network neuroscience, researchers are interested primarily in the brain's capacity to maintain overall information transfer. Efficient information transfer is the primary function of interest. It is a general function in that it does not predict any particular cognitive activity or resulting behavior, but it is a prerequisite for both. Abstracting away from particular cognitive functions allows researchers to think about how the brain is organized to enable (and perhaps to retain) efficient information transfer. Network neuroscientists are interested in understanding how the human brain works as an informational system and where the transfer of information might break down. So, functional robustness here is the continued transfer of information throughout the brain after some internal or external change to the normal organization

of the system (e.g., a lesion, developmental differences). When there is functional robustness, changes in the underlying organizational structures (i.e., brain areas and the white fiber tracts connecting them) do not significantly reduce the efficient flow of information in the system. Functional vulnerability is when any change to the underlying organization of the normal system causes a significant decrease or complete failure to efficiently transfer information throughout the system. Researchers are interested in explaining the particular observed pattern of functional robustness and vulnerability, that is, explaining why efficient information transfer is preserved under some internal or external perturbations, but not others.

The pattern of robustness and vulnerability is based on observations of clinical patients with brain damage. While some lesions (e.g., in Broca's area) have relatively strong and restricted types of functional deficits, other lesions are often quickly compensated for (Stromswold, 2000; Young, Hilgetag, & Scannell, 2000; for applications to *C. elegans*, see Towlson & Barabasi 2020). Further, some lesions disproportionately affect a wide range of functions (Damasio & Damasio, 1989; Mesulam, 2000). Neuroscientists introduce the concept of diaschisis to describe the cases where lesions in a particular cortical area can modulate neuronal responses and undermine functions based in distant and otherwise viable cortical regions (Price, Warburton, Moore, Frackowiak, & Friston, 2001). Evidence of diaschisis has been exhibited in a wide variety of brain areas associated with a wide range of functions like attention, cognitive control, and language (Carter et al., 2010; Nomura et al., 2010; Price et al., 2001).

This fits more generally with a trend in research on psychiatric disorders to emphasize the commonalities among many mental disorders rather than the unique differences of each disorder. While much work on psychiatric conditions in neuroscience continues to search for predictors of diagnosis and prognosis, some recent work has sought to understand psychiatric disorders as

characterized in part by connectivity changes (e.g., dysconnectivity or hyper-connectivity relative to normal populations) (van den Heuvel & Sporns, 2019). For example, the normal modular structure of the human connectome is disrupted in autism, depression, and epilepsy (Alexander-Bloch et al., 2010; Lord, 2012; Rudie, 2012; Vaessen, 2013). Work on identifying common disruptions of the connectome for individuals with psychiatric and brain disorders is consistent with emphasizing symptom dimensions rather than discrete categories of mental disorder (see recent work on psychiatric disorders, e.g., Borsboom, 2010).

Researchers have recently turned to network models of the brain to model these cases of nonlocal functional deficits (e.g., He et al. 2007). On my view, this research supports the following topological explanation:

(1) Small-worldness and modularity as opposed to scale-free degree distributions entail a distinct pattern of functional robustness and vulnerability to perturbation. (demonstrated by simulations)
(2) The macroscale human brain realizes a small-world and modular network as opposed to a scale-free network. (supported by empirical evidence)
(3) Therefore, the macroscale human brain exhibits a distinct pattern of functional robustness and vulnerability to damage in virtue of its small-worldness and modularity.

Researchers demonstrate that the topological properties of interest entail the relevant trait, property, or propensity to behave in the explanandum. In doing so, researchers contrast the topological properties of interest with other topological properties and look for differences in the entailment of the relevant trait, property, or propensities to behave. Achard et al. (2006) demonstrate that small-world networks have a distinct pattern of robustness and vulnerability to particular types of perturbation. To test for robustness in a network context, researchers delete nodes (Achard et al. 2006; sometimes nodes representing individual neurons for the *C. elegans*, see Towlson & Barabási, 2020), edges (Kaiser & Hilgetag, 2004), or simulate lesions by deleting multiple nodes in a spatially adjacent area (e.g., Alstott, Breakspear, Hagmann, Cammoun, &

Sporns, 2009; Kaiser, Martin, Andras, & Young, 2007). Small-world networks are explicitly compared to scale-free networks to determine the distinctive patterns of robustness and vulnerability for each type of network.[23] Both small-world and scale-free networks are robust to deletions of randomly selected nodes. However, small-worlds are more robust to deletions than scale-free networks when nodes with high degree are targeted. In fact, scale-free networks are particularly vulnerable to these degree-targeted attacks. Small-world networks are also vulnerable to deletions of nodes with high betweenness centrality in the network (Alstott et al. 2009). In particular, it is the modular structure of some small-world networks that limits the spread of dysfunction and neurodegeneration (van den Heuvel & Sporns, 2019).

Premise 1 receives further support from simulations of lesions in network models of the brain. When modeling lesions (e.g., deletions of spatially close nodes) in a network, lesions to some areas of the network showed more non-local effects than others (Alstott et al. 2009). That is, the functional deficits could not be accounted for by predicting functional deficits of the spatial area of the nodes along with disruptions of the immediate connections between that area and other areas. From this evidence, the researchers can conclude that small-world networks exhibit a distinctive pattern of robustness and vulnerability of function to perturbation and that this distinct pattern is different from the pattern seen in networks with other topological properties (e.g., non-small-world but scale-free networks).

To support premise 2, we can draw on the empirical evidence that many functional and structural network models of the macroscale brain find that the brain is a small-world (Hagmann et al. 2007; He, Chen, and Evans 2007; Iturria-Medina et al. 2007; Achard and Bullmore 2007;

[23] Although it is possible for a network to be both a small world and scale-free (e.g. the *C. elegans* neural system), the two properties are distinct. Recent evidence suggests that the human brain is not appropriately modeled as a scale-free network (Achard et al. 2006).

Achard et al. 2006; Stam 2004). Both structural and functional network models of the macroscale brain also indicate it is modular, in the sense that it can be divided into communities of networks that are relatively disconnected from each other (Bassett et al., 2013; Bullmore & Sporns, 2012; Meunier, Lambiotte, & Bullmore, 2010; Meunier et al., 2009).[24] Other evidence suggests that the macroscale human brain is not appropriately modeled as a scale-free network (Achard et al. 2006).[25] These findings provide initial evidence that the macroscale brain realizes a network model that is within the class of small-world but not scale-free models. Given the overall current empirical support described here, the macroscale brain can be said to realize the topological properties small-worldness and modularity.

Network neuroscientists have thus explained observed patterns of robustness and vulnerability of the brain in virtue of the mathematical relationship between small-worldness and such patterns.[26] Recall that they were initially accounting for clinical observations concerning the general pattern of robustness and vulnerability of the human brain to lesions. Modeling the macroscale brain as a network emphasizes the topological property explains this particular pattern.

[24] The sense of modularity in network science is distinct from the evolutionary sense of modularity and the Fodorian sense of modularity, contra (Colombo, 2013).

[25] However, both He, Zempel, Snyder, and Raichle (2010) and van den Heuvel, Stam, Boersma, and Hulshoff Pol (2008) model the brain as a scale-free network.

[26] Structurally, topological explanations resemble deductive-nomological (DN) explanations, with relationships between topological properties and behaviors playing a role analogous to that of laws of nature in the classic DN model. Lange (2013) has argued that mathematical explanations (of which topological explanations are a subset) describe the results of mathematical necessity and thus, are more necessary than ordinary laws of nature. I take no stand on whether such mathematical relationships should be considered genuine laws of nature.

### 4.4.3 Objections

One objection to my argument is that the "pattern of robustness and vulnerability" used in premise (1) is not the same as its use in the explanandum, and thus my argument equivocates. The motivation for this objection is that the robustness and vulnerability in network science is assessed involves manipulating a model (e.g., deleting a node or edge), whereas the robustness and vulnerability of the human brain as observed through clinical and psychiatric studies is assessed due to traumatic or developmental deviations from the normal macroscopic human brain.

This objection rests on a misunderstanding of the sense of "functional robustness" used in the argument. As I emphasized earlier, the argument takes place within the context of network neuroscience and thus focuses on a very minimal sense of 'function', which is not specific to any particular underlying mechanisms. In network neuroscience, the brain is represented as an efficient information processing system and the explanation is of general observed patterns robustness and vulnerability of the ability to process information through neural communication. All other details about the system are abstracted away, including specific cognitive functions. Robustness in this sense does not require that there are different underlying mechanisms that can carry out specific cognitive functions. Instead, it requires that there are multiple routes to get information from one part of the system to another. The empirical evidence can support both this information transfer functional robustness and claims about the robustness or vulnerability of specific cognitive functions. In the case of information transfer abilities, the details of particular cognitive functions are abstracted away.

Another possible objection to my argument here is that premise (2) is false because recent empirical work does not support the claim that the macroscopic human brain has small-world organization. There is significant controversy over whether neural systems can be appropriately

said to have small-world organization. For example, Hilgetag & Goulas (2015) argue that the brain may not be best modeled as a small-world network. First, different levels of organization (e.g., nodes as neurons, anatomical brain areas) may have different topological organizations. Second, if the actual density of connections were included in the network, then the brain might have a large-world organization (Moretti & Muñoz, 2013). Large-world networks are characterized as networks where some subparts of the network may be relatively inaccessible from other parts. More formally, we can define a "topological dimension, D, that measures how the number of neighbours of any given node grows when moving 1, 2, 3,..., r steps away from it: $N_r \sim r_D$ for large values of r" (Moretti & Muñoz 2013, 2). For (classic) small-world networks, as the network size grows exponentially, D approaches ∞. For large-world networks, $0 < D < \infty$.

The problem is that Hilgetag and Goulas premise their argument on the distinctions between the classical small-world organization introduced in Watts and Strogatz (1998), which are not modular or hierarchical, and large-world networks, which are both modular and hierarchical. So, as the name of their article asks: are brains really small-worlds? Perhaps the most reasonable answer is that they are not *classic* small-worlds (in Watts and Strogatz's sense), but there are more complex and biologically plausible definitions of small-worldness that are also compatible with the fact that the brain is expected to be modular and hierarchical.

New definitions of small-worldness aim to compare the organization of a given brain network with a more appropriate null model comparison classes that control for the density of connections in the network, which has been shown to artificially increase small-world estimates (Muldoon, Bridgeford, & Bassett, 2016). Our understanding of the density of networks has also progressed as new tract tracing methods reveal more density in connections; for example, macaque visual cortex is now thought to be 66% dense as opposed to previous estimates of approximately

20-40% density (Markov et al., 2013).[27] On these new metrics, most neural systems are still considered small-worlds, but the quantitative estimates of small-worldness are reduced (Bassett & Bullmore, 2017; Muldoon, Bridgeford, & Bassett, 2016). When examining human, Macaque, and *C. elegans* neural networks, only the *C. elegans* nervous system is no longer considered to be a small-world. Further, even groups that critique the frequentist approach to estimating small-worldness and propose a Bayesian alternative conclude that while the values of the small-worldness metric is uncertain, hold that it is still sufficiently informative to conclude that the human brain has a small-world structure (Zanin, Belkoura, Gomez, Alfaro, & Cano, 2018). Thus, Hilgetag and Goulas's strong conclusion that the brain is not a small-world does not follow from the evidence they have provided.

### 4.5 Responding to Mechanistic Objections

I will now argue that understanding this use of network models is best understood as a case of topological explanation. To do so, I present three options for the mechanist: (a) argue the case is not mechanistic but also not explanatory, (b) argue the case is explanatory but mechanistic, or (c) give up the scope claim that all explanation in neuroscience is mechanistic. I will consider objections for both (a) and (b), which follow from the mechanists' responses to other alleged counter-examples. I argue that none of these objections are successful and that the mechanist should take option (c).

[27] It is worth pointing out that density likely has been underestimated in early network neuroscience studies, which focused on bidirectional and unweighted edges, and were built from DTI data that cannot resolve small fiber tracts (Hilgetag & Goulas, 2015).

### 4.5.1 Giving Up the Scope

In section 4.5, I argue that the mechanist who defends either (a) or (b) in response to my case of purported explanation has to make some undesirable concessions. The mechanist can pursue these familiar objections to the mounting number of alleged counter-examples—including my own—but only by endorsing undesirable consequences: denying that cases which fulfill the difference-making and asymmetry norms are genuinely explanatory (a) or rendering the mechanistic account of explanation superfluous to accounts of causal explanation (b).

As a result, I argue that giving up the wide scope of mechanistic explanation in neuroscience (c) is the best move for the mechanist. A limited scope for mechanistic explanation is consistent with the mechanist account of explanation accurately describing a large number of cases of explanation in neuroscience as well as providing important normative guidance for mechanistic explanations. Acknowledging that the mechanistic account of explanation is limited in scope better accommodates the classic cases of mechanistic explanation and retains the distinctive norms of mechanistic explanation. To argue for limiting the scope of mechanistic explanation, I will now reply to a number of objections one might have to the case presented in Section 4.4.2 as a genuine counter-example to the scope claim.

### 4.5.2 Topological Explanations are Genuinely Explanatory

Craver argues that alleged counter-examples to the wide scope claims about mechanistic explanation are not in fact explanatory (e.g., Kaplan and Craver 2011; Piccinini and Craver 2011). I will address Craver's objection first and argue that the case meets a number of explanatory norms

(endorsed by the mechanists), including answering what-if-things-had-been-different (w-) questions and maintaining that explanation is asymmetrical.

### 4.5.3 Meets the Norms of Explanation

In this section, I will examine the ways in which the case described in section 4.4.2 is genuinely explanatory by appealing to one norm of explanation accepted by some mechanists (e.g., Craver 2007; Zednik 2014). In particular, I argue that the case fulfills difference-making norms by providing answers to w-questions. While the case fulfills a number of norms for explanation, it does not meet the criteria to count as a causal explanation. In light of this, I argue that topological explanations are genuine, but can be non-causal explanations.[28]

One norm of scientific explanation is to emphasize systematic patterns of counterfactual dependence that make a difference to the explanandum. Woodward (2003) describes these patterns of counterfactual dependence as answers to w-questions. On Woodward's view, "a successful explanation should identify conditions that are explanatorily or causally relevant to the explanandum: … those that 'make a difference' to the explanandum in the sense that changes in these factors lead to changes in the explanandum" (Woodward 2013, 5). While Huneman does not explicitly endorse Woodward's view, he does argue that topological explanation is a distinct form of explanation because in its instances, topological properties make a difference to the presence of the behavior of interest (Huneman 2010, 218).

[28] I will not discuss any potential examples of causal topological explanation, but such cases may still be non-mechanistic.

The mechanists who accept the wide scope claim also accept the ability to provide answers to w-questions as a norm of explanation (Craver, 2007; Kaplan, 2011; Zednik, 2014). In fact, in some of Craver's most recent work (Craver & Kaplan 2020, 311) he appeals to Woodward's account of w-questions in his account of the explanatory completeness norm of mechanistic explanation. One caveat is that Craver & Kaplan also introduce a subset of w-questions called how-does-that-work (h-) questions. H-questions are focused on constitutive explanations, whereas w-questions more broadly leave room for other kinds of explanations (e.g., aetiological causal explanations).

However, w-questions do not strictly require causal explanations and sometimes call for non-causal explanations. Both causal and non-causal explanations provide answers to w-questions, but they do so in different ways. Woodward (2003, 221) writes:

> the common element in many forms of explanation, both causal and noncausal, is that they must answer what-if-things-had-been-different questions. When a theory tells us how Y would change under interventions on X, we have (or have the material for constructing) a causal explanation. When a theory or derivation answers a what-if-things-had-been-different question but we cannot interpret this as an answer to a question about what would happen under an intervention, we may have a noncausal explanation of some sort.

An explanation is causal, on Woodward's view, when answers to w-questions can be interpreted according to an ideal intervention.[29] An explanation is non-causal when answers to w-questions cannot be given by an ideal intervention, namely, when constitutive or mathematical relationships are also changed in the process of an ideal intervention (see also Chirimuuta, 2018). In Woodward's (2003) framework, an intervention is an ideal, unconfounded experimental manipulation of the values of some upstream variable X such that the values of Y are affected only through this manipulation. On an ideal intervention, the changes in the values of Y can only be

[29] Other definitions of causal explanation exist in the literature (e.g., Machamer, 2004; Reutlinger & Andersen, 2016). I will not argue here for the preference of this distinction between causal and non-causal explanation.

through X and its causal path and not through other causal paths that do not involve X. Antecedents to answers of w-question can fail to be appropriate targets of an intervention when an unconfounded intervention is impossible according to the relationships among the variables (e.g., constitution, supervenience, mathematical entailment). For example, if an ideal manipulation of X is confounded by also manipulating Z, both of which contribute to changes in the values of Y, then an ideal unconfounded intervention is impossible.[30] If we can still answer w-questions about how changes in X would affect changes in Y, these answers will be non-causal explanations.

Topological explanations fulfill difference-making norms of explanation. Scientists can and do answer w-questions to learn about relationships between topological properties of the model and the behavior of interest. In particular, scientists can answer the following w-question: if the topological properties of the network model of the macroscale human brain had been different, would the observed pattern of robustness and vulnerability be different (and how)? To provide answers to this question, scientists "lesion" (sets of) nodes with different topological properties. Manipulations of any node in the network will not always affect the higher-order topological properties, but in some cases they will. For example, the deletion of nodes with the highest betweenness-centrality—a node through which a high number of shortest paths between other nodes in the network must pass—might result in a scale-free network model. In that case the observed pattern of robustness and vulnerability would be different, and the system would be robust to random perturbations but not targeted attacks on the nodes with high degree. However, in cases where a less crucial node was deleted, the difference between the two models would not

[30] An ideal intervention may be impossible in situations where intervening on one causal relationship also ideally intervenes on another causal relationship in the system. For example, given the supervenience of mental states on physical states, an *unconfounded* ideal intervention on mental states is impossible because by supervenience, any intervention on mental states will also be an intervention on physical states.

make a difference to the presence of the particular pattern of robustness and vulnerability. It is the topological property that makes a difference to the behavior of the system. The deletion of particular nodes in a network model make a difference to the behavior of the system only when they have certain topological properties themselves (e.g., centrality), which are dependent on the global state of the network.

While providing answers to this w-question is informative for the presence or absence of the particular observed pattern of robustness and vulnerability of the macroscale human brain, such answers do not follow from ideal interventions. At least in this case, we cannot intervene on the second-order topological properties (e.g., small-worldness) of the brain without simultaneously intervening on other relationships between first-order topological properties. The failure of an ideal intervention is due to the compositional relationship between second-order topological properties and first-order topological properties. According to the distinction between causal and non-causal explanations adopted in this chapter, the case of topological explanation presented in section 4.4.2 is non-causal. However, if some network model could be ideally intervened upon at the level of first-order properties to answer w-questions, then there may be cases of causal topological explanation.

One possibility open to the mechanist, however, is that better explanations provide answers to more w-questions (e.g., Craver & Kaplan 2020). This claim about degrees of explanatory power fits well with views of explanatory "depth" (Lange, 2015). Perhaps the mechanist would suggest that answering fewer w-questions means that topological explanations are less explanatory than mechanistic explanations.

Even if it is true that the case I present in this chapter answers fewer w-questions than a hypothetical mechanistic explanation of some particular traumatic brain injury, developmental or

psychiatric disorder, it does not follow that the case I present has no explanatory power. The ability to answer any w-questions comes with explanatory power. It may be the case that answering further w-questions is indicative of increased explanatory power, but that general claim does not undermine the explanatory power of answering some w-questions. Combined with the claim that we should be clear and careful about our explananda so that we mistakenly do not shift between alternative explananda when evaluating an explanation, this objection does not succeed in showing how these cases are not explanatory. It may be true that some mechanistic explanations answer more w-questions (and different types of w-questions, like h-questions), but that does not make them better explanations relative to the particular explanandum.

### 4.5.4 Craver's Asymmetry Objection

Craver (2016) has raised an objection to the understanding of network models in neuroscience as providing genuine explanations.[31] Craver argues that scientific explanation adheres to the norm of asymmetry. He has in mind the flagpole case (Salmon, 1989): the flagpole's height and position of the sun explain the shadow's length, but the shadow's length does not explain the flagpole's height. He argues that accounts of non-mechanistic explanation do not provide an account of why explanation should be asymmetrical. For example, paradigmatic cases of potential non-mechanistic explanation like the Königsberg bridge problem can be problematically run "reversed", and so these cases do not provide genuine explanations.

[31] In fact, Craver's objection is to all mathematical explanation, including topological explanation. While I agree that topological explanation is best understood as a type of mathematical explanation, I do not have space here to address whether all cases of topological explanation would count as a distinctively mathematical explanation. For more on different ways in which an explanation may be mathematical, see (Baker, 2009; Batterman, 2002; Lange, 2013).

On Craver and Povich's view, topological explanation cannot account for the asymmetry of explanation. One version of the solution to the Königsberg problem is a proper explanation, but in the other direction it is not. Consider the explanation Euler provides:

(1) *Empirical Premise*: Königsberg can be represented as a network with four nodes where three nodes have three edges and one node has five.
(2) *Mathematical Premise*: Of the networks composed of four nodes, only those containing either zero or two nodes of odd degree have Eulerian paths.
(3) *Conclusion*: There is no Eulerian path across the bridges of Königsberg.

Yet, according to Craver (2006, 701; Craver & Povich, 2017, 35), this explanation of the solution to the Königsberg problem can be reversed by switching the empirical premise and the conclusion while negating both to give the following *reversed case*:

(1) *Empirical Premise*: Marta walks an Eulerian path through Königsberg (i.e., an Eulerian path is possible).
(2) *Mathematical Premise*: Among connected networks composed of four nodes, only networks containing zero or two nodes with odd degree also contain an Eulerian path.
(3) *Conclusion*: Therefore, either zero or two of Königsberg's landmasses have an odd number of bridges in 1756.

Craver and Povich take the reversed case to be a bad argument, whereas *ex hypothesi* on the topological explanation account, the original argument is good. This constitutes a problem for topological explanation because, according to Craver, there are no norms of topological explanation that can explain this asymmetry.

Similarly, Craver argues that the same reversed argument can be given for the mathematical explanation of the brain's robustness in terms of its small-worldness. The empirical premise that the brain is a small-world together with the mathematical premise that small-worlds are more robust to random attack than random networks allow us to conclude that the brain is robust. He suggests the following small-world argument:

(1) *Empirical Premise*: System S is a small-world network.
(2) *Mathematical Premise*: Small-world networks are more robust to random attack than are random or regular networks.

(3) *Conclusion*: Therefore, system S is more robust to random attack than random or regular networks.

Alternatively, according to Craver's reversed cases, the reversed small-world argument would be:

(1) *Empirical Premise*: System S is less robust to random attack than random or regular networks.
(2) *Mathematical Premise*: Small-world networks are more robust to random attack than are random or regular networks.
(3) *Conclusion*: Therefore, system S is not a small-world network.

If it is true that the topological explanation concerning small-worldness and robustness can be run in both directions, then topological explanation would be symmetrical and thus, fail to satisfy one norm of explanation.

Proponents of non-mechanistic explanation have given two general responses to this objection: either explaining how non-mechanistic explanation can adhere to the asymmetry norm of explanation or by denying the asymmetry norm. Accepting the asymmetry norm has been much more popular than rejecting it, though the latter strategy could be motivated on an epistemic view of explanation. If what counts as an explanation is primarily epistemic, then as van Fraassen (1980) argued, the shadow of the flag pole does give an explanation of its height insofar as it provides reason to believe how tall the flagpole is, provides understanding about how tall the flagpole must be given the height of its shadow, and so on. However, here I will not consider the recent distinction between ontic and epistemic views of explanation. First, this distinction need not track the distinction between proponents of mechanistic explanation and critics. For example, some mechanists subscribe to the ontic view (Craver, 2013) while others hold the epistemic view (Wright & Bechtel, 2007). Second, the arguments here are neutral with respect to these views.

Instead, I will appeal to recent work by Lange (2013; 2018) arguing that the asymmetry norm is upheld in cases of mathematical explanation. Lange's response is based on the claim that the explanatory question or problem often appears to be ambiguously defined. There are features

of an explanatory question or problem that are constitutive. These features are assumptions required in order to begin to answer the explanatory question or problem that is posed. As Lange (2017, 33) says, they are "fixed parameters of the cases with which those why questions are concerned." To illustrate, Lange suggests that were the Burgermeister of Königsberg to have built additional bridges in order to walk an Eulerian path, then we would still not consider the Königsberg bridge problem to be solved. Such a move would violate the assumptions on which the explanatory question or problem is premised. As part of the Bridges of Königsberg problem, it is an assumption that the network of bridges and landmasses remains fixed. On this view, we only need to slightly revise the explanatory schema for these cases in order to avoid the symmetrical responses to both original and reversed cases.

Take the Königsberg case: we begin with the explanatory question "Why is there no possible Eulerian path in historic Königsberg, which is represented as a network with four nodes where three nodes have three edges and one node has five?" So, to change the network layout of Königsberg is to avoid answering the original explanatory question. But in the reversed cases, the explanatory question does not presuppose the empirical claim. Consider the reversed Königsberg case, which would have the following explanatory question: "Why does either zero or two of Königsberg's landmasses have an odd number of bridges in 1756, given that Marta walks an Eulerian path across the bridges?" The presence of an Eulerian path is not constitutive of the explanatory task set out by the question. In the original cases but not in the reversed cases, the explanatory question requires a graph theoretic perspective and assumes the network layout of the bridges. Similar arguments can address the small-world case.

If the mechanist persists in arguing that topological explanation does not meet any norms of explanation, then she denies that non-mechanistic strategies fulfill the difference-making and

asymmetry norms of explanation. In this section, I have argued that the case of topological explanation I present does fulfill the difference-making and asymmetry norms, which are accepted by the mechanists who endorse the wide scope claim. Unlike canonical cases of mechanistic explanation, the case of topological explanation fulfills difference-making norms non-causally. That is, they provide answers to some w-questions, but not by utilizing ideal interventions. Further, topological explanations are not symmetrical. I have argued that current objections to the explanatory value of topological explanations are unsuccessful and I have outlined the ways in which the case I described in section 4.4.2 meets some difference-making norms of explanation.

### 4.5.5 Topological Explanations are Not Mechanistic

Other mechanists might agree that the case discussed in section 4.4.2 is genuinely explanatory but argue that the case is adequately accounted for on the mechanistic account of explanation, namely, as a mechanism sketch. I will reply that the case is not adequately considered as a mechanism sketch, as this would (falsely) imply that including, e.g., more causally directed connections in the network would improve its explanatory power. Another way to pursue this objection involves arguing that better explanations involve answering more w-questions (e.g., Kaplan 2011, 354). Further, I argue that the case does not fit a reasonable account of mechanistic explanation.

### 4.5.6 Additional Causal Detail is Irrelevant and Detrimental

The mechanist might respond to my arguments above that network models are mere mechanism sketches with causal details to be filled in after more research has been completed.

Mechanism sketches are relatively incomplete models, which either leave gaps for "bottom-out" entities or "black boxes" for some parts and activities in the system (e.g., Craver 2007).[32] According to this objection, the explanation of the distinct pattern of robustness and vulnerability of the macroscale human brain will be better explained if more causal information were included in the network model(s). The network models discussed in this chapter do not include any causal information about the direction of causal influence between two connected nodes nor any causal detail about the nodes themselves (e.g., whether the node is a set of inhibitory neurons). As more causal detail is added, the objection goes, these models become filled out as mechanism schemata. However, mechanism schemata need not include every possible detail about a system (Kaplan & Craver 2011, 610). These models represent all and only the relevant causal detail about the system for the purposes of explaining the phenomenon of interest (recall 3M and 3M* above).

Many philosophers of science, including Craver writing elsewhere (Craver, 2010), agree with the principle that only relevant information should be included in an explanation. We might also characterize the further principle that irrelevant details make an explanation worse. When combined, these principles reflect some elements of 3M* that details should be included in an explanation if and only if they are relevant to the explanandum. One view is that only details that make a difference to the *presence of the phenomenon* in the explanandum should be included in an explanation (e.g., Strevens, 2009). I will interpret this principle in the following way: details are relevant to an explanation only if those details decide between the members of the implicit contrast class (i.e., make a difference to) the presence of some trait, property, or propensity in an

[32] For a different version of the distinction see Craver and Darden (2013) and Machamer, Darden, & Craver (2000).

explanandum. Casual details that are relevant for some explanandum, may be irrelevant for another explanandum.[33]

One potential objection to my argument here is that in order to provide an explanation of any particular mental or brain disorder, neuroscientists need to appeal to further causal information not provided by the putative case of topological explanation. For example, neuroscientists will need to account for the role of tau in any explanation of Alzheimer's disease. Indeed, some promising evidence of the progression of Alzheimer's disease suggests that tau may spread along the major white fiber pathways in the brain (e.g., Zhou et al., 2012). These integrative practices of network models with molecular and cellular evidence fits best with the view that sometimes topological properties and, what might be considered, mechanistic details work together to provide an explanation (Huneman, 2015).

In response, I claim that such an objection does not hold the explanandum fixed. Craver & Kaplan (2020) recently emphasize that explananda need to be clearly distinguished when evaluating potential explanations and so should accept that it could be a particular failure of this objection. It is important to note that if some further mechanistic details are necessary to provide a sufficient explanation for the progression of Alzheimer's disease, it does not count against the particular purported explanation I have presented here. The objection mixes up two different explananda: the explanandum concerning Alzheimer's disease (focused on explaining the development of a particular disease) compared to the explanandum we began with (focused on explaining a general pattern of resilience and vulnerability to disease in general). In the case of explaining *the distinct pattern of robustness and vulnerability of the macroscale human brain*, the topological properties of the brain are explanatorily sufficient. If the brain had different topological

[33] One could hold that causal details are always relevant, but see Andersen (2018) for arguments against this position.

properties, such as scale-free organization rather than small-worldness, then the distinct pattern of robustness and vulnerability would be different.

According to the norms of explanation described above (e.g., answering w-questions), the presence of this topological property of the brain is explanatorily sufficient to explain the observed pattern of robustness and vulnerability. Filling out causal details of network models of the brain (e.g., determining causally directed edges) are irrelevant for the purposes of this explanandum. The mathematical consequence of the robust functional pattern is due to the contrast between small-world and modular networks from scale-free networks. The addition of these causal details to the network model will not help the neuroscientists provide a better explanation of the distinct pattern of robustness and vulnerability exhibited by the macroscale brain. Given the claim that the addition of irrelevant detail in an explanation is detrimental to its explanatory power, the addition of irrelevant causal detail for this explanandum would provide a worse explanation.

If the mechanist still considers these models mechanism sketches, she is suggesting that causal detail added to current functional or structural network models (regardless of the impact on topological properties) will be explanatorily relevant because such causal information is relevant for any explananda. Alternatively, the mechanist could be slipping between different explananda. Either possibility is not a defensible position, given the commitments Craver and Kaplan have made to: (1) include all and only relevant details in an explanation and (2) be precise about what explanandum is being considered in debates about specific cases. I have argued here and in the previous section that understanding the relationship between the topological property and the behavior of system is sufficient for answering w-questions and so for providing a genuine explanation of this particular explanandum. The addition of causal information in network models of the brain will be relevant only for some other explananda.

### 4.5.7 Fails to Fulfill Distinctive Norms of Mechanistic Explanation

Another objection to alleged counter-examples to the wide scope claim pursued by mechanists involves widening the account of mechanistic explanation. Some mechanists allow a wide variety of explanatory practices to count as mechanistic in virtue of fulfilling even one or two norms of mechanistic explanation. For example, Zednik (2014, 2019) claims that as long as models are organized and emphasize points of intervention in a system then those models provide mechanistic explanations. Such a permissive view of what strategies count as mechanistic accommodates nearly all counter-examples, but with the consequence that the distinctive norms of mechanistic explanation, such as the emphasis on finding entities and activities, are abandoned. The original account of mechanistic explanation (e.g., Machamer, Darden, and Craver 2000) was less permissive and better described scientific explanatory practices in molecular biology and cellular neuroscience.

Consider an alternative to Zednik's permissive account. Woodward (2013) offers a reasonable elaboration of the mechanistic account of explanation as a cluster-concept and illuminates a set of three criteria (though not complete) that systems should have for mechanistic strategies of explanation (e.g., decomposition and localization, the search for functionally differentiated entities) to apply: stability of causal relationships, fine-tunedness of spatio-temporal organization, and modularity of causal relationships. Stability requires that the causal relations among the components be more stable over interventions than the system's overall behavior. The fine-tunedness of organization requires that the specific causal properties of the components and their spatio-temporal organization makes a difference to the system's overall behavior. Modularity of causal relationships requires that components' causal interactions remain relatively unaffected when interventions are performed on some other component.

As cases fulfill fewer of the criteria, the cases will be less explicable by a mechanistic strategy. The case of the pattern of robustness and vulnerability of the macroscale brain to perturbation fails two of these criteria. Note that these criteria apply to the system and not the models. However, we can restrict our assessment of the system to features that are relevant to the explananda of interest. First, the macroscale brain case fails the stability criterion. Causal connections between some brain areas are less stable under interventions than the efficient transfer of information across the whole brain. So, the causal relationships between the components (i.e., brain areas) are less stable under perturbation than the whole system behavior. Information can be efficiently transferred from point A to point B in the brain through different routes when causal relationships among components are disrupted. The only exception to this case is when a particularly well-connected component is perturbed; in those cases, the global behavior (i.e., information transfer) is less stable than many of the causal relationships among components.

Second, the macroscale human brain viewed in terms of its efficiency of information transfer, fails to fulfill the spatio-temporal fine-tunedness criterion. A relatively limited amount of causal properties of the components make a difference to the systems' ability to efficiently transfer information. Researchers can abstract away from causal details about how information is transferred, whether it is electrical or chemical, or how it is modulated by neurotransmitters. Also, the spatio-temporal organization of the system is similarly relatively limited. Spatial arrangement among the components matters only insofar as the connections among components are represented. However, these connections do not need to be to scale and in fact, many of the details of these connections can be abstracted away, such as how many neurons have axonal projects through a particular white fiber tract or whether there are reflexive connections.

As this case looks quite different from stereotypical examples of mechanistic explanation, it fulfills fewer distinctive norms of mechanistic explanation. I have provided a number of reasons for rejecting the claim that this case is mechanistic—it fails two of the three normative criteria distinctive of mechanistic explanation.

Mechanists like Zednik may wish to argue that even cases that meet a minimal number of criteria of causal explanation should count as mechanistic and that the distinctive criteria for mechanistic explanation described in this section are too restrictive. Consider the weakened criteria of condition (a) from 3M to 3M* discussed in section 4.3.1. One problem with this permissive view of mechanistic explanation is that accepting cases at the periphery of the cluster-concept of mechanistic explanation will broaden the account and ultimately render it trivial. The mechanistic account of explanation would be superfluous if it turned out that mechanistic explanation is merely synonymous with our best account of causal explanation. The proponents of this permissive account of mechanistic explanation give up the distinctive content and norms of mechanistic explanation (such as spatio-temporal fine-tunedness) highlight features of the account that gave it traction in the canonical cases of mechanistic explanation. For these reasons, mechanists should reject the permissive account of mechanistic explanation.

However, I've argued that if the mechanist accepts the restrictive account of mechanistic explanation—including the three criteria mentioned above—then the case of explanation I've described in this chapter is not explicable by mechanistic strategies. If we retain the more restrictive view of mechanistic explanation presented in Woodward (2013), then as I've argued here the distinctive pattern of robustness and vulnerability of the macroscale human brain is a phenomenon that is not appropriately explained by the mechanistic account. Rather the case is an instance of non-mechanistic topological explanation in neuroscience.

## 4.6 Conclusion

By shifting the debate between mechanists and proponents of alternative theories of explanation to focus on distinct types of explanatory questions, we can better articulate the limits of the mechanistic explanation. I have argued that the mechanist cannot maintain the wide scope claims about explanation in neuroscience in the face of cases of topological explanation in neuroscience (like the case in section 4.4.2). Instead, the mechanist should give up the wide scope claim and concede that some explanatory projects in neuroscience do not fit the mechanistic account.

# 5.0 Conclusion

In this dissertation, I have analyzed the two major senses of robustness in scientific research. Chapters 2 and 3 deal with robustness reasoning, or the idea that robust results from our scientific modeling and experimentation practices are more trustworthy.

In Chapter 2, I argued that robustness analysis has utility during scientific discovery and the pursuit of hypotheses rather than contexts of confirmation of results or robust theorems. I developed and defended a new type of robustness analysis called 'scope robustness analysis', which is a helpful research strategy when researchers have some knowledge about the target system that can be leveraged to constrain other hypotheses that are explored via modeling. I drew connections to the modeling of possibilities in existing accounts of how-possibly explanations and perspectival modeling. I demonstrate scope robustness analysis by analyzing a case of generative modeling in network neuroscience. Researchers use the strategies I described to identify the likely range of trade-off between two developmental principles for the system in question.

In Chapter 3, I argued for two major claims. First, the sense of diversity among the methods required for triangulation is not explanatory diversity, but rather diversity in the ways each method might fail. My major argument for this claim was that the explanatory diversity criterion is based on an analysis of abductive eliminative inferences underlying the design of experiments on Brownian motion. However, triangulation relies on a different type of underlying inference: common cause inductive inferences. For this reason, Perrin's account of the estimation of Avogadro's number is an instance of triangulation, but Perrin's experiments on Brownian motion are not. So, the explanatory diversity criterion does not identify "sufficiently diverse" methods for triangulation, but may play some role in how researchers typically design a series of experiments

to eliminate alternative explanations for their initial observed results. Given the problems with other diversity criteria, I hold that the failure diversity criterion, as given by Wimsatt, remains our best understanding of "sufficiently diverse" methods in triangulation.

My second claim is that philosophical accounts of triangulation cannot explain why triangulation can fail in practice. We need to explain why triangulation can fail in practice, in addition to why it succeeds, in order for our account of triangulation to be descriptively adequate and to normatively guide scientific practice. Even extending these accounts, which are based on successful cases of triangulation, by considering failures of the two major success criteria do not explain some cases of failure to triangulation. Likewise, bringing in resources from the literature on inductive risk is insufficient to address the problem. Thus, I provide a new account of triangulation based on epistemic risk and highlight particular ways and places in research practice where triangulation can fail. Importantly, my account also requires contributing to the types of epistemic risk that can arise in all kinds of knowledge practices in addition to triangulation in particular.

Finally, I applied my account of triangulation to explain the failure of the triangulation argument for implicit attitudes. While most of the philosophical criticism of implicit attitude research has focused on its issues with characterizing the phenomenon or its lack of predictive validity, I demonstrated that implicit attitude research relies on an assumption that all indirect measures are measuring the same phenomenon. This triangulation argument for implicit attitudes has not been the focus of most critiques in the social psychology literature either. I argued that there are two main problems with this triangulation argument, which can be understood using my account of triangulation. First, there is an epistemic risk in deciding whether data should serve as evidence for the same hypothesis. Second, there is a type of inductive risk that is particularly salient

in triangulation: the risk that there is insufficient evidence to rule out the hypothesis that there is more than one phenomenon.

In Chapter 4, I turned to consider explanation of phenomena that are robust to many lower-level details. Rather than assessing how best to explain robust phenomena in general, I examined this topic in the context of mechanistic explanation. Much of the debate about the scope of mechanistic explanation can be considered in the context of identifying whether some phenomena require other explanatory strategies. In this chapter, I argued that topological explanation is distinct from mechanistic explanation and for some cases of robust explananda in network neuroscience, topological explanation is more appropriate. Specifically, I considered the explananda of the pattern of macroscopic human brain's robustness and vulnerability to damage to its parts. For this particular explanandum, we can provide a topological explanation citing the topological properties of the system (i.e., small-worldness and modularity) and their mathematical consequences (i.e., the pattern of robustness and vulnerability in a network).

Considering recent responses to purported cases of non-mechanistic explanation, I argued that the mechanist should take this case to be both explanatory and non-mechanistic. It is explanatory because it fulfills norms of explanation that the mechanists also accept, namely, difference-making norms and asymmetry norms. The case ought to be seen as non-mechanistic because more details about the parts, such as their spatio-temporal locations, would not improve an explanation of this particular explanandum. While the mechanist could reject that all mechanistic explanations require, e.g., appeal to spatio-temporal locations of the parts, I have argued that it would be an unsavory move for the mechanist. In rejecting norms specific to mechanistic explanation, the mechanist loses the distinct content of the account and instead relies on more general norms of explanation (especially causal explanation). This result would retain the

wide scope claim concerning mechanistic explanation, but risks making the mechanistic account of explanation (with the distinctive norms elaborated) superfluous to existing accounts of explanation.

## Bibliography


Achard, S., Salvador, R., Whitcher, B., Suckling, J., & Bullmore, E. (2006). A Resilient, Low-Frequency, Small-World Human Brain Functional Network with Highly Connected Association Cortical Hubs. *Journal of Neuroscience*, *26*(1), 63–72. https://doi.org/10.1523/JNEUROSCI.3874-05.2006

Achard, Sophie, & Bullmore, E. (2007). Efficiency and cost of economical brain functional networks. *PLoS Computational Biology*, *3*(2), 0174–0183. https://doi.org/10.1371/journal.pcbi.0030017

Adler, J. (1976). The sensing of chemicals by bacteria. *Scientific American*, *234*, 40–47.

Alexander-Bloch, A. F., Gogtay, N., Meunier, D., Birn, R., Clasen, L., Lalonde, F., … Bullmore, E. T. (2010). Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. *Frontiers in Systems Neuroscience*, *4*(147), 1–16. https://doi.org/10.3389/fnsys.2010.00147

Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., & Sporns, O. (2009). Modeling the impact of lesions in the human brain. *PLoS Computational Biology*, *5*(6). https://doi.org/10.1371/journal.pcbi.1000408

Andersen, Holly. (2018). Complements, Not Competitors: Causal and Mathematical Explanations. *British Journal for the Philosophy of Science*, *69*(2), 485–508. https://doi.org/10.1093/bjps/axw023

Ankeny, R. A. (2001). The natural history of Caenorhabditis elegans research. *Nature Genetics*, *2*(June), 474–479.

Ankeny, R. A. (2006). Wormy Logic : Model Organisms As Case-Based Reasoning, (07).

Baker, A. (2009). Mathematical Explanation in Science. *British Journal for the Philosophy of Science*, *60*, 611–633. https://doi.org/10.1093/bjps/axp025

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger & J. S. Nairne (Eds.), *The Nature of Remembering: Essays in Honor of Robert G. Crowder* (pp. 117–150). Washington, DC: American Psychological Association.

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*, 509–512. https://doi.org/10.1126/science.286.5439.509

Bassett, D. S., & Bullmore, E. (2006). Small-World Brain Networks. *The Neuroscientist*, *12*(6), 512–523. https://doi.org/10.1177/1073858406293182

Bassett, Danielle S., & Bullmore, E. T. (2017). Small-World Brain Networks Revisited. *Neuroscientist*, *23*(5), 499–516. https://doi.org/10.1177/1073858416667720

Bassett, Danielle S., Porter, M. A., Wymbs, N. F., Grafton, S. T., Carlson, J. M., & Mucha, P. J. (2013). Robust detection of dynamic community structure in networks. *Chaos*, *23*(1). https://doi.org/10.1063/1.4790830

Batterman, R. W. (2002). Asymptotics and the Role of Minimal Models. *British Journal for the Philosophy of Science*, *53*(1), 21–38.

Batterman, R. W., & Rice, C. C. (2014). Minimal Model Explanations. *Philosophy of Science*, *81*(3), 349–376. https://doi.org/10.1086/676677

Bechtel, W. (2013). From Molecules to Networks: Adoption of Systems Approaches in Circadian Rhythm Research. In H. Andersen, D. Dieks, W. Gonzalez, T. Uebel, & G. Wheeler (Eds.), *New Challenges to Philosophy of Science: The Philosophy of Science in a European Perspective* (pp. 211–223). Springer, Dordrecht.

Bechtel, W. (2015a). Can Mechanistic Explanation be Reconciled with Scale-free Constitution and Dynamics? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *53*, 84–93.

Bechtel, W. (2015b). Generalizing Mechanistic Explanations through Graph-theoretic Perspectives. In P.-A. Braillard & C. Malaterre (Eds.), *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences* (Vol. 11, pp. 199–225). https://doi.org/10.1017/CBO9781107415324.004

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*. John Wiley & Sons Inc.

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A*, *41*(3), 321–333. https://doi.org/10.1016/j.shpsa.2010.07.003

Biddle, J. B. (2016). Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease. *Perspectives on Science*, *22*(3), 397–417. https://doi.org/10.1162/POSC

Biddle, J. B. (2018). "Antiscience zealotry"? values, epistemic risk, and the GMO debate. *Philosophy of Science*, *85*(3), 360–379. https://doi.org/10.1086/697749

Biddle, J. B., & Kukla, R. (2017). The Geography of Epistemic Risk. In K. C. Elliott & T. Richards (Eds.), *Exploring Inductive Risk: Case Studies of Values in Science* (pp. 215–238). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190467715.003.0011

Bogen, J., & Woodward, J. F. (1988). Saving the phenomena. *The Philosophical Review*,

*XCVII*(3), 303–352.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*, 631–643.

Bullmore, E, & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci.*, *10*(3), 186–198. https://doi.org/10.1038/nrn2575

Bullmore, Ed, & Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, *13*(MAY), 336–349. https://doi.org/10.1038/nrn3214

Campbell, D.T. (1969). Definitional Versus Multiple Operationalism. In E.S. Overman (Ed.), *Methodology and Epistemology for Social Science*. Chicago: University of Chicago Press.

Campbell, Donald T, & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105.

Carter, A. R., Astafiev, S. V., Lang, C. E., Connor, L. T., Rengachary, J., & Stube, M. J. (2010). Resting Interhemispheric Functional Magnetic Resonance Imaging Connectivity Predicts Performance After Stroke. *Annals of Neurology*, *67*, 365–375.

Cartwright, N. (1991). Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy*, *23*(1), 143–155. https://doi.org/10.1215/00182702-23-1-143

Cartwright, Nancy. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Chen, B. L., Hall, D. H., & Chklovskii, D. B. (2006). Wiring optimization can relate neuronal structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12), 4723–4728. https://doi.org/10.1073/pnas.0506806103

Chen, Y., Wang, S., Hilgetag, C. C., & Zhou, C. (2013). Trade-off between Multiple Constraints Enables Simultaneous Formation of Modules and Hubs in Neural Systems. *PLoS Computational Biology*, *9*(3), 31–33. https://doi.org/10.1371/journal.pcbi.1002937

Cherniak, C, Changizi, M., & Won Kang, D. (1999). Large-scale optimization of neuron arbors. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *59*, 6001–6009.

Cherniak, Christopher. (1995). Neural component placement. *Trends in Neurosciences*, *18*(12), 522–527. https://doi.org/10.1016/0166-2236(95)98373-7

Cherniak, Christopher, Mokhtarzada, Z., Rodriguez-Esteban, R., & Changizi, K. (2004). Global optimization of cerebral cortex layout. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(4), 1081–1086. https://doi.org/10.1073/pnas.0305212101

Chirimuuta, M. (2018). Explanation in computational neuroscience: Causal and non-causal.

*British Journal for the Philosophy of Science*, *69*(3), 849–880. https://doi.org/10.1093/bjps/axw034

Chirimuuta, Mazviita. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese*, *191*(2), 127–153. https://doi.org/10.1007/s

Chklovskii, D., & Stepanyants, A. (2003). Power-law for axon diameters at branch point. *BMC Neuroscience*, *4*, 18.

Colaço, D. (2018). Rip it up and start again: the rejection of a characterization of a phenomenon.

Colombo, M. (2013). Moving Forward (and Beyond) the Modularity Debate: A Network Perspective. *Philosophy of Science*, *80*(July), 356–377. https://doi.org/10.1086/670331

Colombo, Matteo. (2013). Moving Forward (and Beyond) the Modularity Debate: A Network Perspective. *Philosophy of Science*, *80*(July), 356–377.

Corkin, S. (2002). What's new with the amnesic patient H.M.? *Nature Reviews Neuroscience*, *3*(2), 153–160. https://doi.org/10.1038/nrn726

Cramer, A. O., Waldorp, L. J., van der Maas, H. L., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2–3), 137–193.

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

Craver, C. F. (2010). Prosthetic Models. *Philosophy of Science*, *77*, 840–851.

Craver, C. F. (2013). The Ontic Conception of Scientific Explanation. In A. Hutteman & M. Kaiser (Eds.), *Explanation in the Biological and Historical Sciences.* Springer.

Craver, C. F. (2016). The Explanatory Power of Network Models. *Philosophy of Science*, *83*(December), 698–709.

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries Across the Life Sciences*. The University of Chicago 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*, *71*, 287–319. https://doi.org/10.1093/bjps/axy015

Craver, C. F., & Povich, M. (2017). The directionality of distinctively mathematical explanations. *Studies in History and Philosophy of Science Part A*, *63*, 31–38. https://doi.org/10.1016/j.shpsa.2017.04.005

Culp, S. (1994). Defending Robustness: The Bacterial Mesosome as a Test Case. *Philosophy of Science*, *1*, 46–57.

Damasio, H., & Damasio, A. R. (1989). *Lesion Analysis in Neuropsychology*. Oxford: Oxford University Press.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*, 443– 451.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit Measures: A Normative Analysis and Review. *Psychological Bulletin*, *135*(3), 347–368. https://doi.org/10.1037/a0014211

Douglas, H. (2016). Values in Science. *Oxford Handbook in The Philosophy of Science*, 23.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive Racism and Selection Decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319. https://doi.org/10.7897/2277-4343.04136

Erdős, P., & Rényi, A. (1960). On the Evolution of Random Graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, *5*, 17–60.

Euler, L. (1741). Solutio problematic ad geometriam situs pertinentis [The solution of a problem relating to the geometry of position]. In *Commentarii academiae scientiarum Petropolitanae.*

Fazio, R. H., & Hilden, L. E. (2001). Emotional Reactions to a Seemingly Prejudiced Response: The Role of Automatically Activated Racial Attitudes and Motivation to Control Prejudiced Reactions. *Personality and Social Psychology Bulletin*, *27*(5), 538–549. https://doi.org/10.1177/0146167201275003

Fazio, Russell H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Fazio, Russell H., & Olson, M. A. (2003). Attitudes: Foundations, functions, and consequences. *The SAGE Handbook of Social Psychology*, (January), 123–145. https://doi.org/10.4135/9781848608221.n6

Fazio, Russell H, Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the Automatic Activation of Attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238.

Feest, U. (2005). Operationism in psychology: what the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences*, *41*(2), 131–149.

Feest, U. (2016). The experimenters' regress reconsidered: Replication, tacit knowledge, and the dynamics of knowledge generation. *Studies in History and Philosophy of Science Part A*, *58*, 34–45. https://doi.org/10.1016/j.shpsa.2016.04.003

Fitelson, B. (2001). A Bayesian Account of Independent Evidence with Applications. *Philosophy of Science*, *68*, S123–S140.

Fodor, J. A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Forber, P. (2010). Confirmation and explaining how possible. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *41*(1), 32–40. https://doi.org/10.1016/j.shpsc.2009.12.006

Franklin, A., & Howson, C. (1984). Why Do Scientists Prefer to Vary Their Experiments? *Studies in History and Philosophy of Science*, *15*(1), 51–62.

Gawronski, Bertram, Hofmann, W., & Wilbur, C. J. (2006). Are "Implicit" Attitudes Unconscious? *Consciousness and Cognition*, *15*, 485–499.

Gillett, C. (2010). Moving beyond the subset model of realization: The problem of qualitative distinctness in the metaphysics of science. *Synthese*, *177*(2), 165–192. https://doi.org/10.1007/s11229-010-9840-1

Green, S. (2013). Tracing Organizing Principles : Learning from the History of Systems Biology. *Hist. Phil. Life Sci.*, *35*, 553–576.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*.

Greenwald, A. G., McGee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press.

Hadi Hosseini, S. M., & Kesler, S. R. (2013). Influence of Choice of Null Network on Small-world Parameters of Structural Correlation Networks. *PLoS ONE*, *8*(6), e67354.

Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V. J., Meuli, R., & Thiran, J. P. (2007). Mapping human whole-brain structural networks with diffusion MRI. *PLoS ONE*, *2*(7). https://doi.org/10.1371/journal.pone.0000597

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

He, B. J., Snyder, A. Z., Vincent, J. L., Epstein, A., Shulman, G. L., & Corbetta, M. (2007). Breakdown of Functional Connectivity in Frontoparietal Networks Underlies Behavioral Deficits in Spatial Neglect. *Neuron*, *53*(6), P905-918. https://doi.org/10.1016/j.neuron.2007.02.013

He, B. J., Zempel, J. M., Snyder, A. Z., & Raichle, M. E. (2010). The Temporal Structures and Functional Significance of Scale-free Brain Activity. *Neuron*, *66*(3), 353–369. https://doi.org/10.1016/j.neuron.2010.04.020

He, Y., Chen, Z., & Evans, A. C. (2007). Small-world Anatomical Networks in the Human Brain Revealed by Cortical Thickness from MRI. *Cerebral Cortex*, *17*, 2407–2419.

Heesen, R., Bright, L. K., & Zucker, A. (2016). Vindicating methodological triangulation. *Synthese*, 1–15. https://doi.org/10.1007/s11229-016-1294-7

Hilgetag, C. C., & Goulas, A. (2015). Is the brain really a small-world network ? *Brain Structure and Function*. https://doi.org/10.1007/s00429-015-1035-6

Holroyd, J., Scaife, R., & Stafford, T. (2017). What is implicit bias? *Philosophy Compass*, *12*(10), 1–18. https://doi.org/10.1111/phc3.12437

Houkes, W., & Vaesen, K. (2012). Robust! Handle with Care. *Philosophy of Science*, *79*(July), 345–364.

Hudson, R. G. (1999). Mesosomes: A Study in the Nature of Experimental Reasoning. *Philosophy of Science*, *66*(2), 289–309.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*(640).

Humphries, M. D., & Gurney, K. (2008). Network "small-world-ness": A quantitative method for determining canonical network equivalence. *PLoS ONE*, *3*(4). https://doi.org/10.1371/journal.pone.0002051

Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese*, *177*(2), 213–245. https://doi.org/10.1007/s11229-010-9842-z

Huneman, P. (2015). Diversifying the picture of explanations in biological sciences: ways of combining topology with mechanisms. *Synthese*. https://doi.org/10.1007/s11229-015-0808-z

Iturria-Medina, Y., Canales-Rodriguez, E. J., Melie-Garcia, L., Valdes-Hernandez, P. A., Martinez-Montes, E., Aleman-Gomez, Y., & Sanchez-Bornot, J. M. (2007). Characterizing brain anatomical connections using diffusion weighted MRI and graph theory. *NeuroImage*, *36*(3), 645–660.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*(5), 349–364. https://doi.org/10.1037/h0031617

Jones, N. (2014). Bowtie Structures, Pathway Diagrams, and Topological Explanation. *Erkenntnis*, *79*, 1135–1155. https://doi.org/10.1007/s10670-014-9598-9

Jones, T. B., Bandettini, P. A., Kenworthy, L., Case, L. K., Milleville, S. C., Martin, A., & Birn, R. M. (2010). Sources of group differences in functional connectivity: An investigation applied to autism spectrum disorder. *NeuroImage*, *49*(1), 401–414. https://doi.org/10.1016/j.neuroimage.2009.07.051

Kaiser, M., & Hilgetag, C. C. (2004). Edge vulnerability in neural and metabolic networks.

*Biological Cybernetics*, *90*(5), 311–317. https://doi.org/10.1007/s00422-004-0479-1

Kaiser, M., & Hilgetag, C. C. (2006). Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS Computational Biology*, *2*(7), 0805–0815. https://doi.org/10.1371/journal.pcbi.0020095

Kaiser, M., Martin, R., Andras, P., & Young, M. P. (2007). Simulation of robustness against lesions of cortical networks. *European Journal of Neuroscience*, *25*(10), 3185–3192. https://doi.org/10.1111/j.1460-9568.2007.05574.x

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*(3), 339–373. https://doi.org/10.1007/s11229-011-9970-0

Kaplan, D. M., & Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, *78*(4), 601–627. https://doi.org/10.1086/661755

Kim, S., Putrino, D., Ghosh, S., & Brown, E. N. (2011). A Granger Causality Measure for Point Process Models of Ensemble Neural Spiking Activity. *PLoS Computational Biology*, *7*(3), e1001110.

Kinoshita, S., & Peek-O'Leary, M. (2005). Does the compatibility effect in the race Implicit Association Test reflect familiarity or affect? *Psychonomic Bulletin and Review*, *12*(3), 442–452. https://doi.org/10.3758/BF03193786

Kitano, H. (2004). Biological robustness in complex host-pathogen systems. *Nature Reviews Genetics*, *5*(November), 826–837. https://doi.org/10.1007/978-3-7643-7567-6_10

Klyachko, V. a, & Stevens, C. F. (2003). Connectivity optimization and the positioning of cortical areas. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(13), 7937–7941. https://doi.org/10.1073/pnas.0932745100

Knight, R. T. (2007). Neural Networks Debunk Phrenology. *Science*, *316*(5831), 1578–1579.

Knuuttila, T., & Loettgers, A. (2011). Causal isolation robustness analysis: the combinatorial strategy of circadian clock research. *Biological Philosophy*, *26*, 773–791.

Kostić, D. (2012). Mechanistic and topological explanations. *Synthese*, (1998).

Kostić, D. (2018). The topological realization. *Synthese*, *195*(1), 79–98. https://doi.org/10.1007/s11229-016-1248-0

Kostić, D. (2020). General theory of topological explanations and explanatory asymmetry. *Phil Trans Royal Society B: Biological Sciences*, *375*.

Kukla, R. (2019). Infertility, epistemic risk, and disease definitions. *Synthese*, *196*(11), 4409–4428. https://doi.org/10.1007/s11229-017-1405-0

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic Modelling as Robustness Analysis. *British Journal for the Philosophy of Science*, *61*(3), 541–567. https://doi.org/10.1093/bjps/axp049

Kuorikoski, J., & Marchionni, C. (2016). Evidential Diversity and the Triangulation of Phenomena. *Philosophy of Science*, *83*(2), 227–247. https://doi.org/10.1086/684960

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: What we know (so far) about the method. In B. Wittenbrink & B. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59–102). New York: Guilford.

Lange, M. (2013). What Makes a Scientific Explanation Distinctively Mathematical? *British Journal for the Philosophy of Science*, *64*, 485–511.

Lange, Marc. (2015). On "Minimal Model Explanations": A Reply to Batterman and Rice. *Philosophy of Science*, *82*(April), 292–305.

Lange, Marc. (2018). A reply to Craver and Povich on the directionality of distinctively mathematical explanations. *Studies in History and Philosophy of Science Part A*, *67*, 85–88. https://doi.org/10.1016/j.shpsa.2018.01.002

Levins, R. (1966). The Strategy of Model Building in Population Biology. *American Scientist*, *54*(4), 421–431.

Levy, A., & Bechtel, W. (2013). Abstraction and the Organization of Mechanisms. *Philosophy of Science*, *80*(2), 241–261. https://doi.org/10.1086/670300

Lloyd, E. A. (2010). Confirmation and Robustness of Climate Models. *Philosophy of Science*, *77*(5), 971–984.

Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science*, *49*, 58–68.

Lord, A. et al. (2012). Changes in community structure of resting state functional connectivity in unipolar depression. *PloS One*, *7*, e41282.

Machamer, P. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science*, *18*(1), 27–39. https://doi.org/10.1080/02698590412331289242

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking About Mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Machery, E. (n.d.). What is a replication?

Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, vol. 1*. Oxford: Oxford University Press.

Markov, N. T. et al. (2013). Cortical high-density counter- stream architectures. *Science*, *342*, 1238406.

Massimi, M. (2018). Perspectival modeling. *Philosophy of Science*, *85*(3), 335–359.

Mayo, D. (1996). The Experimental Basis from Which to Test Hypotheses: Brownian Motion. In *Error and the Growth of Experimental Knowledge*. Chicago: Chicago University Press.

Mesulam, M. M. (2000). *Principles of Behavioral and Cognitive Neurology* (2nd Editio). Oxford: Oxford University Press.

Meunier, D., Lambiotte, R., & Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience*, *4*(December), 1–11. https://doi.org/10.3389/fnins.2010.00200

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K. D., Bullmore, E. T., & Valdes-sosa, P. (2009). Hierarchical modularity in human brain functional networks. *Frontiers in Neuroinformatics*, *3*(October), 1–12. https://doi.org/10.3389/neuro.11.037

Middendorf, M., Ziv, E., & Wiggins, C. H. (2005). Inferring network mechanisms: The Drosophila melanogaster protein interaction network. *PNAS*, *102*(9), 3192–3197.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*, 455–469.

Moretti, P., & Muñoz, M. A. (2013). Griffiths phases and the stretching of criticality in brain networks. *Nature Communications*, *4*. https://doi.org/10.1038/ncomms3521

Muldoon, S. F., Bridgeford, E. W., & Bassett, D. S. (2016). Small-world propensity and weighted brain networks. *Scientific Reports*, *6*(June 2015), 1–13. https://doi.org/10.1038/srep22057

Munafò, M. R., & Smith, G. D. (2018). Repeating experiments is not enough. *Nature*, *553*, 399–401.

Newman, M. E. J. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8696.

Nomura, E. M., Gratton, C., Visser, R. ., Kayser, A., Perez, F., & D'Esposito, M. (2010). Double Dissociation of Two Cognitive Control Networks in Patients with Focal Brain Lesions. *Proceedings of the National Academy of Sciences*, *107*, 12017–12022.

Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: robustness analyses in economics and biology. *Biology & Philosophy*, *24*(1). https://doi.org/10.1007/s10539-011-9278-y

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science*, *14*, 636– 639.

Orzack, S. H., & Sober, E. (1993). A Critical Assessment of Levins's The Strategy of Model

Building in Population Biology (1966). *The Quarterly Review of Biology*, *68*(4), 533–546.

Parker, W. S. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science*, *78*(4), 579–600.

Parker, W. S. (2013). Getting (even more) serious about similarity. *Biology and Philosophy*, (November 2013), 1–10. https://doi.org/10.1007/s10539-013-9406-y

Payne, B. K., & Gawronski, B. (2015). A History of Implicit Social Cognition: Where Is It Coming From? Where Is It Now? Where Is It Going? In B. Gawronski & B. K. Payne (Eds.), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford.

Pérez-Escudero, A., & de Polavieja, G. G. (2007). Optimally wired subnetwork determines neuroanatomy of Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 17180–17185. https://doi.org/10.1073/pnas.0703183104

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Power, J. D., Schlaggar, B. L., & Petersen, S. E. (2014). Recent progress and outstanding issues in motion correction in resting state fMRI. *NeuroImage*, *105*, 536–551. https://doi.org/10.1016/j.neuroimage.2014.10.044

Price, C. J., Warburton, E. A., Moore, C. J., Frackowiak, R. S., & Friston, K. J. (2001). Dynamic diaschisis: Anatomically Remote and Context-Sensitive Human Brain Lesions. *Journal of Cognitive Neuroscience*, *13*(4), 419–429.

Raerinne, J. (2013). Robustness and sensitivity of biological models. *Philosophical Studies*, *166*(2), 285–303. https://doi.org/10.1007/s11098-012-0040-3

Reutlinger, A., & Andersen, H. (2016). Abstract versus Causal Explanations? *International Studies in the Philosophy of Science*, *30*(2), 129–146. https://doi.org/10.1080/02698595.2016.1265867

Ross, L. N. (2015). Dynamical Models and Explanation in Neuroscience. *Philosophy of Science*, *82*(1), 32–54. https://doi.org/10.1086/679038

Rothermund, K., & Wentura, D. (2004). Underlying Processes in the Implicit Association Test: Dissociating Salience From Associations. *Journal of Experimental Psychology: General*, *133*(2), 139–165. https://doi.org/10.1037/0096-3445.133.2.139

Rudie, J. D. et al. (2012). Altered functional and structural brain network organization in autism. *NeuroImage: Clinical*, *2*, 79–94.

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

Salmon, W. (1989). Four Decades of Scientific Explanation. *Minnesota Studies in the Philosophy of Science, 13*, 3–219.

Schickore, J., & Coko, K. (2013). Using multiple means of determination. *International Studies in the Philosophy of Science*, *27*(3), 295–313. https://doi.org/10.1080/02698595.2013.825498

Schupbach, J. N. (2015). Robustness, Diversity of Evidence, and Probabilistic Independence. In V. Maki & S. Ruphy (Eds.), *Recent Developments in the Philosophy of Science: EPSA13 Helsinki* (pp. 305–316).

Schupbach, J. N. (2018). Robustness Analysis as Explanatory Reasoning. *British Journal for the Philosophy of Science*, *69*(1), 275–300.

Seung, S. (2012). *Connectome: How Our Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt Trade.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Silberstein, M., & Chemero, A. (2013). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences. *Philosophy of Science*, *80*(5), 958–970. https://doi.org/10.1086/674533

Solé, R. V., Pastor-Satorras, R., Smith, E., & Kepler, T. B. (2002). A Model of Large-Scale Proteome Evolution. *Advances in Complex Systems*, *5*(1), 43–54.

Soler, L. (2012). Introduction: The Solidity of Scientific Achievements: Structure of the Problem, Difficulties, Philosohpical Implications. In L. Soler, T. Nickles, E. Trizio, & W. C. Wimsatt (Eds.), *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science* (pp. 1–60). New York: Springer.

Stam, C. J. (2004). Functional Connectivity Patterns of Human Magnetoencephalographic recordings: a "small-world" network? *Neuroscience Letter*, *355*(1–2), 25–28.

Stegenga, J. (2009). Robustness, Discordance, and Relevance. *Philosophy of Science*, *76*, 650–661.

Strevens, M. (2009). *Depth: An Account of Scientific Explanation*. Cambridge: Harvard University Press.

Stromswold, K. (2000). The Cognitive Science of Language Acquisition. In M. Gazzaniga (Ed.),

*The new cognitive neurosciences* (2nd Editio, pp. 909–932). MIT Press.

Towlson, E. K., & Barabási, A.-L. (2020). Synthetic ablations in the C. elegans nervous system. *Network Neuroscience*, 1–17. https://doi.org/10.1162/netn_a_00115

Vaessen, M. J. et al. (2013). Abnormal modular organization of functional networks in cognitively impaired children with frontal lobe epilepsy. *Cerebral Cortex*, *23*, 1997–2006.

van den Heuvel, M P, Stam, C. J., Boersma, M., & Pol, H. E. H. (2008). Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. *NeuroImage*, *43*(3), 528–539. https://doi.org/10.1016/j.neuroimage.2008.08.010

van den Heuvel, Martijn P., & Sporns, O. (2019). A cross-disorder connectome landscape of brain dysconnectivity. *Nature Reviews Neuroscience*, *20*(7), 435–446. https://doi.org/10.1038/s41583-019-0177-6

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Vázquez, A. (2010). Protein Interaction Networks. In O. Alzate (Ed.), *Neuroproteomics* (pp. 38–44). Boca Raton, FL: CRC Press/Taylor & Francis. https://doi.org/10.1159/000067642

Watts, D. J., & Strogatz, S. H. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Weisberg, M. (2006). Robustness Analysis. *Philosophy of Science*, *73*(5), 730–742.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy*, *104*(12), 639–659. https://doi.org/10.5840/jphil20071041240

Weisberg, M. (2013). *Similarity and Simulation: Using Models to Understand the World*. Oxford University Press.

Weisberg, M., & Reisman, K. (2008). The Robust Volterra Principle. *Philosophy of Science*, *75*(1), 106–131.

White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986). The Structure of the Nervous System of the Nematode Caenorhabditis elegans. *Philos.Trans.R.Soc Lond B Biol.Sci.*, *314*, 1–340.

Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Science Part A*, *40*(1), 92–101. https://doi.org/10.1016/j.shpsa.2008.12.005

Wimsatt, W. (2012). Robustness: Material, and Inferential, in the Natural and Human Sciences. In *Characterizing the Robustness of Science* (pp. 89–104). Boston: Boston Studies in the Philosophy of Science.

Wimsatt, W. C. (1981). Robustness, Reliability, and Overdetermination. In *Characterizing the Robustness of Science* (pp. 61–87). https://doi.org/10.1007/978-94-007-2759-5

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. University of Chicago Press.

Witt, S. T., & Meyerand, M. E. (2009). The Effects of Computational Method, Data Modeling, and TR on Effective Connectivity Results. *Brain Imaging Behavior*, *3*(2), 220–231.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*(2), 219–240. https://doi.org/10.1080/13501780600733376

Woodward, J. (2013). Mechanistic Explanation: Its Scope and Limits. *Proceedings of the Aristotelian Society, Supplement*, 39–65.

Wright, C. D., & Bechtel, W. (2007). Mechanisms and Psychological Explanation. In P. Thagard (Ed.), *Handbook of the Philosophy of Science.* Elsevier B.V.

Young, M. P., Hilgetag, C. C., & Scannell, J. W. (2000). On imputing function to structure from the behavioural effects of brain lesions. *Philos.Trans.R.Soc Lond B Biol.Sci.*, *355*(1393), 147–161. https://doi.org/10.1098/rstb.2000.0555

Zanin, M., Belkoura, S., Gomez, J., Alfaro, C., & Cano, J. (2018). Topological structures are consistently overestimated in functional complex networks. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-30472-z

Zednik, C. (2014). Systems, networks, and mechanistic explanations in neuroscience. *Philosophy of Science Assoc. 24th Biennial Mtg (Chicago, IL)*, 1–22.

Zednik, C. (2019). Models and mechanisms in network neuroscience. *Philosophical Psychology*, *32*(1), 23–51. https://doi.org/10.1080/09515089.2018.1512090

Zhou, J. et al. (2012). Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron*, *73*, 1216–1227.