UNIVERSITY OF CALIFORNIA,
IRVINE

A Middle-Range Account of Analogical Inference

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Jessica Lauman-Lairson

Dissertation Committee:
Professor Lauren Ross, Chair
Professor Jeffrey Barrett
Professor Kyle Stanford

2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible without all of the help and support provided by my advisors, Dr. Lauren Ross and Dr. Jeffrey Barrett. The generosity of their time, encouragement, feedback, and support was invaluable to me. I also want to thank my committee member, Dr. Kyle Stanford, whose ideas contributed substantially to the dissertation. I am also grateful to the two anonymous reviewers who gave feedback on Chapter 2 when it was submitted to *Synthese* as a journal paper. Their careful reading and detailed feedback substantially improved the chapter. I am also grateful to my mom, Dr. Beatrice Lauman, who was a constant source of support; my mare, Mel, who played a formative role in my life; and my cat, Little Cat, who appeared outside of my door at the beginning of the doctorate and made Southern California feel like home. Finally, I would like to thank the coffee shops of Ojai, particularly Ojai Coffee Roasters, where much of this dissertation was written.

The text of Chapter 2 of this dissertation is an adaptation of the material as it appears in Lauman-Lairson, J. *A middle-range account of analogical inference: Target Domain Salience. Synthese* **206**, 42 (2025). https://doi.org/10.1007/s11229-025-05136-x, used with permission from Springer Nature.

VITA

Jessica Lauman-Lairson

EDUCATION

Doctor of Philosophy in Logic and Philosophy of Science **2025**
University of California, Irvine *Irvine, California*

Master of Arts in Logic and Philosophy of Science **2022**
University of California, Irvine *City, State*

Master of Arts in Philosophy **2019**
Durham University *Durham, United Kingdom*

Bachelor of Arts in English and Philosophy **2018**
Durham University *Durham, United Kingdom*

TEACHING EXPERIENCE

Teaching Assistant **2019–2025**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

A Middle-Range Account of Analogical Inference: Target Domain Salience **2025**
Synthese, Springer Nature

REFEREED CONFERENCE PUBLICATIONS

Questioning Relevant Similarity Accounts of Analogical Inference **June 2023**
'Workshop on Causality and Explanation', Ludwig Maximilan University of Munich

Questioning Relevant Similarity Accounts of Analogical Inference **October 2023**
'Science, etc', University of Washington

Target Domain Salience **November 2023**
'PSA Around the World', Philosophy of Science Association

Reverse Analogical Inferences: Gerrymandering the Source Domain Representation **June 2024**
SURe (Scientific Understanding and Representation) London School of Economics and University of Cambridge

Reverse Analogical Inferences: Gerrymandering the Source Domain Representation **July 2024**
British Society for the Philosophy of Science, University of York

SERVICE

Institute for Ascertaining Scientific Consensus (IASC) Spokesperson **2023**
I surveyed UC Irvine scientists as a spokesperson for the IASC, a global network that aims to measure scientific consensus and fight misinformation by surveying an international network of scientists.

GSR/DTEI course design **2023**
I assisted Professor Lauren Ross with designing a Winter 2024 course on structural explanations in social science.

Conference co-organizer **2024**
Co-organizer of 'New Directions in Analogical Reasoning' hybrid conference in November 2024. https://philevents.org/event/show/126882

Deputy Editor at Critique Journal **2018**
Critique is an open-access philosophy journal run by students at Durham University

AWARDS/GRANTS

GSR funding from NSF Career Award **2023**
Causal Explanation in Biology and the Diversity of Causal Concepts. National Science Foundation (NSF); Science, Technology, and Society Research Grant; 1945647

Funding for DTEI course design work **2023**
UC Irvine

ABSTRACT OF THE DISSERTATION

A Middle-Range Account of Analogical Inference

By

Jessica Lauman-Lairson

Doctor of Philosophy in Philosophy

University of California, Irvine, 2025

Professor Lauren Ross, Chair

Analogical inference is an important form of scientific reasoning, playing a role in theory choice, experimental design, hypothesis generation, and domain unification. Consequently, recent work in philosophy and cognitive science has aimed to develop normative standards for analogical inference. Current accounts fall into two main categories. The first type, characterized by Paul Bartha's articulation model, Dedre Gentner's systematicity principle, and Mary Hesse's two- dimensional account, aims to provide a definitive framework for what characteristics make an analogical inference plausible. The implicit goal of these accounts is to give necessary and sufficient conditions for prima facie plausibility. The second type of account, characterized by John Norton's material account, disputes the relevant similarity framework's goal of providing universal norms for analogical inference. Norton claims that the plausibility of an analogical inference is local and context-dependent. The material account dispenses with the normative project and argues that attempts to develop domain-general normative standards are misguided. Recognizing that enquiry is shaped by local norms, Norton veers toward a strong relativism in which he discounts the usefulness of metascientific norms. This dissertation works on constructing a middle ground between the formal, relevant similarity framework accounts and material accounts of analogical inference. The norms I postulate are not necessary and sufficient conditions for plausibility, but rather are values that scientists must weigh according to their local epistemic commitments.

Having these epistemic values in their toolkit provides scientists tools for debating about plausibility, examining their background assumptions, incorporating considerations from social epistemology, and using the natural science literature to extrapolate epistemic norms for evaluating relevant similarity.

Chapter 1

Introduction

1.1 Introduction

Analogical reasoning is a cognitive process in which perceived similarities between a source and target domain motivate the inference that some further similarity holds between the source and target domain. Analogical reasoning is used by scientists for a variety of purposes including hypothesis generation and theory choice. Some examples discussed in this dissertation include Einstein's analogical inference between two windowless elevators to argue that inertial mass and gravitational mass are equivalent (Einstein et al., 2015/1917, Section 2.20), ethnographic analogies used by archaeologists to interpret the material remains of settlement sites by comparing them to historically-linked contemporary sites [Huffman 2001], astrobiology research [Schulze-Makuch 2020], and early chemists' investigations of the relationship between the elements [Ross 2018].

Analogical reasoning in science often takes the explicit form of analogical inference, which is a kind of inductive argument. In an analogical inference, postulated similarities between a source and target domain are used to argue that some further feature of the source domain can be generalized to the target domain. Analogical inferences are often used for justifying theory choice. An analogical inference may be used to motivate allocating research funding towards a particular question, or to convince other members of a scientific community that a hypothesis is *prima facie* plausible. Sometimes, analogical inferences are used to make stronger *pro tanto* claims about plausibility. This particularly occurs in contexts in which no further empirical investigation is feasible, such as cases of archaeologists interpreting archival records from a fully excavated site, or instances of analogical transfer between mathematical domains where there is no empirical data to consult to justify the conjecture.

Given the important role of analogical inference in science, it is desirable to have epistemic norms that guide fruitful applications of analogical inference. To meet this need, there has been recent work in philosophy and cognitive science to develop normative accounts of

analogical inference. This dissertation investigates existing normative accounts, identifies several challenges for these accounts, and proposes a middle ground that resolves some of these challenges.

Current accounts fall into two main categories. The first type, characterized by Paul Bartha's articulation model, Dedre Gentner's systematicity principle, and Mary Hesse's two-dimensional account, aims to provide a once-and-for all justification and framework for what characteristics make an analogical inference plausible. The implicit goal of these accounts is to give necessary and sufficient conditions for *prima facie* plausibility. Bartha notes that Hesse and Gentner's accounts fall short of predictive adequacy, as the criteria they posit do not successfully differentiate between plausible and implausible analogical inferences. Thus, Bartha aims to refine their criteria by taxonomizing analogical inferences into five different types and providing distinct normative criteria for each type. However, even Bartha's more fine-grained framework does not meet his goal of constituting necessary and sufficient conditions for *prima facie* plausibility. In Chapter 2, I show that Bartha's two conditions, *Source Domain Salience* and *Overlap*, are neither necessary nor sufficient for *prima facie* plausibility. I call the formal normative framework postulated by these accounts the *relevant similarity framework* because they share a set of assumptions which I discuss in Chapter 2.

The second type of account, characterized by John Norton's material account, disputes the relevant similarity framework's goal of providing universal norms of analogical inference. Norton claims that the plausibility of an analogical inference is local and context-dependent. When evaluating the plausibility of an analogical inference, scientists rely on domain-specific background assumptions which cannot be captured by a universal normative framework. Epistemic norms for analogical inference also change as a domain's theories evolve. Under Norton's material account, the relevant similarity framework accounts are unrealistic because they fail to recognize that an analogical inference's plausibility is evaluated locally and empirically. Norton proposes that analogical inferences are justified by the existence of an

empirically verified material fact that unites the source and target domain, which Norton calls the 'fact of analogy'.

Norton's criticism of the relevant similarity framework is valid—evaluation of plausibility is local and context-dependent. However, Norton's account dispenses with the normative project altogether. Recognizing that enquiry is shaped by local norms, Norton endorses a strong relativism in which he discounts the usefulness of metascientific norms. The fact that there is no universal characterization of a plausible analogical inference does not establish that any normative account of analogical inference is unproductive. While epistemic norms for making and evaluating plausible analogical inferences are constructed and refined locally, a normative account of analogical inference can identify epistemic values that refine this process. Norton's account also faces several other challenges, which I discuss.

My dissertation constructs a middle ground between the formal relevant similarity framework accounts and the relativism of Norton's material account. My aim is to construct a scaffold upon which accepting that norms for analogical inference are locally determined does not mean collapsing into an uncritical endorsement of whichever local norms happen to be operative in a domain. Many resources for this project can be found in the relevant similarity framework accounts. Though these accounts fail at their project of offering necessary and sufficient conditions for plausibility, they capture many patterns that tend to characterize fruitful analogical inferences. Having these patterns in mind as epistemic values that characterize fruitful analogical inference is of benefit to scientists.

I also propose several other epistemic values that are not recognized in current accounts. In Chapter 2, I describe an essential characteristic of plausible analogical inferences, *Target Domain Salience*. Scientists use *Target Domain Salience* as a criterion when making and evaluating analogical inferences, but this criterion fails to be acknowledged in the relevant similarity framework accounts.

In Chapter 3, I challenge one of the epistemic norms of the relevant similarity framework accounts, which claims that gerrymandering of the source domain is unfruitful and epistemically unjustified. I show that this gerrymandering, which I call *reverse analogical inference*, is justified or unjustified for exactly the same reasons that any analogical inference is justified. Finally, in Chapter 4, I show that existing accounts of analogical inference and inductive reasoning have failed to provide adequate naturalistic investigation of the similarity judgments that underlie inductive reasoning. I show that while these similarity judgments are justified locally, there is useful work that can be done to refine the reliability of these similarity judgments by considering the natural science literature on how humans make such judgments. I give several examples of the sorts of epistemic norms that would be useful.

In summary, my dissertation works on constructing a middle ground between the formal, relevant similarity framework accounts and material accounts like Norton's. The norms I postulate are not necessary and sufficient conditions for plausibility, but rather are values that scientists must weigh according to their local epistemic commitments. Having these epistemic values in their toolkit provides scientists tools for debating about plausibility, examining their background assumptions, incorporating considerations from social epistemology, and using the natural science literature to extrapolate epistemic norms for evaluating relevant similarity.

Chapter 2

Target Domain Salience

2.1 Abstract

Current influential normative accounts of analogical inference—like Hesse's model, Bartha's articulation model, and Gentner's structure-mapping model—assume that an analogical inference's plausibility is determined by it having two essential characteristics, which I will call *Source Domain Salience* and *Overlap*. I show that these two characteristics are neither sufficient nor necessary for plausibility. First, these two characteristics do not capture all of the criteria that scientists use to evaluate analogical inferences. Second, some analogical inferences can be plausible because they *lack* these two characteristics. Using examples from physics, archaeology, and astrobiology, I show that an analogical inference's plausibility depends on a further essential characteristic, *Target Domain Salience*, that is not captured by Hesse, Bartha, and Gentner's two essential characteristics. I argue that a successful normative account of analogical inference should establish criteria for this further characteristic.

2.2 Introduction

An analogy is a type of reasoning that identifies similarities between two domains. Reasoning by analogy has a variety of uses in science; it is used in science education, as a heuristic tool in the context of discovery, and for informing theory choice in the context of justification. A particularly important type of analogical reasoning in science is *analogical inference*. Generally, an analogical inference is an inductive inference that identifies similarities between a source and target domain, and uses those similarities to motivate the conclusion that some further feature of the source domain can be generalized to the target domain. Usually the source domain is a more familiar domain and/or there is a certain degree of scientific consensus about the theory governing the source domain's phenomena. In contrast, the target domain may be less familiar or lack an agreed-upon theory. Analogical inferences are used

in science to guide research into understudied areas, to justify existing theories, to unify distinct theoretical entities, and to differentiate distinct entities that have previously been treated as indistinct.

There are many prominent examples of analogical inferences used in science. Charles Darwin used analogical inferences from artificial selection and from Malthus' economic theory to explain and justify his theory of natural selection [Darwin 2008/1859, pg. 54; pg. 7]. Einstein used an analogy between two windowless elevators to argue that inertial mass and gravitational mass are equivalent [Einstein 1917, section 2.20]. Analogical inferences are ubiquitous in paleoanthropology, where paleoanthropologists draw conclusions about our early human ancestors by making analogical inferences from familiar source domains like living primates or contemporary hunter-gatherers [Hrdy 2009].

Because analogical inferences play such an important role in science, there is a need for normative standards for analogical inference. We want to have criteria for evaluating the analogical inferences we encounter and guidelines for using analogical inferences in research in a fruitful way. In response to this need, the analogical inference literature has seen a number of proposed normative accounts of analogical inference.

These normative accounts fall into two main categories: philosophical accounts and cognitive science accounts. Many of the former accounts are informed by Mary Hesse's [1966] influential work on analogical inference and refine Hesse's criteria for what essential characteristics an analogical inference must have in order to be considered plausible. Paul Bartha's articulation model [2009] is one such account. Cognitive science accounts of analogical inference like Dedre Gentner's [1983] account focus on the importance of there being structural relations in the source domain that can be mapped to the target.

I will show that though Hesse, Bartha, and Gentner's accounts differ from each other on a fine-grained level, they all see an analogical inference's plausibility as being determined by

two essential characteristics which I call *Source Domain Salience* and *Overlap*, and they aim to give normative criteria that capture these two characteristics.

I will show that *Source Domain Salience* and *Overlap* are neither necessary nor sufficient for plausibility; there is a further essential characteristic, which I call *Target Domain Salience*, that scientists take into account when evaluating analogical inferences. Hesse, Bartha, and Gentner's accounts lack criteria for *Target Domain Salience*. In section 3, I consider how a successful normative account of analogical inference could go about capturing criteria for it.

2.3 Two essential characteristics

2.3.1 What is relevant similarity?

Analogical inferences rely on similarities between the source and target domain to motivate the conclusion that some further feature of the source domain is potentially generalizable to the target domain. But not just any similarities will do: a plausible analogical inference depends on relevant similarities. Consider the following example.

Example 2.3.1. The position of turtles in the phylogenetic tree has been a longtime puzzle in biology [Sato 2023; Shaffer 1997; Evers 2019; Fujita 2004]. Turtles share few similarities with other groups of vertebrates, probably partly due to their development of a shell altering selection pressures for other anatomical features. There are two main schools of thought when it comes to the position of turtles in the phylogentic tree. The first school of thought places turtles as the sister clade to the extinct subclass of reptiles, *Anapsida* [Lee 1997, pgs. 1-2]. Anapsid reptiles lacked temporal fenestrae (openings in the skull adjacent to the eye socket for jaw muscles), and this is thought to be an ancestral trait of all anapsid reptiles. Turtles too lack temporal fenestrae, and on the basis of this similarity, some phylogeneticists think that turtles must share a common ancestor with anapsid reptiles. The second school of

thought places turtles as the sister taxa of archosaurs (crocodiles and birds). This second school of thought uses DNA sequencing and cites molecular similarities between archosaurs and turtles [Crawford 2012].

Debate about the phylogenetic positioning of turtles involves debating which of these similarities is more relevant to turtles' ancestry. The archosaur school argue that turtles' lack of temporal fenestration is not sufficient evidence for them sharing a recent common ancestor with anapsid reptiles. They argue that while turtles do lack temporal fenestrae, in turtles it may not be an ancestral trait as it is in anapsid reptiles. Turtles could have come from an ancestral lineage that did have temporal fenestrae and lost them at some point in their evolutionary history. In other words, it is not lacking temporal fenestrae that is relevant to membership in *Anapsida*, but coming from an evolutionary lineage that lacked temporal fenestrae. There is not sufficient evidence in the fossil record to support this stronger condition for turtles' membership in the anapsid reptile family, so the relevance of turtles' morphological similarity to anapsid reptiles to the possibility of them being a sister taxa of anapsids is called into question.

In contrast, the anapsid reptile school of thought questions the relevance of the molecular similarities between turtles and archosaurs. They point out that while turtles do share molecular similarities with archosaurs, turtles also share a similar degree of molecular similarity with lepidosaurs (lizards, snakes, and tuatara) [Lyson et. al. 2012]. The molecular similarity between turtles and archosaurs cannot be highly relevant to the possibility of turtles being a sister taxa of archosaurs given that turtles share a similar amount of molecular properties with non-archosaurs.

In summary, the debate over turtles' phylogentic position is not a debate about whether turtles do share the aforementioned similarities with archosaurs and anapsid reptiles. The debate is about how *relevant* these similarities are to the possibility of turtles being related to archosaurs or anapsids. When scientists debate the plausibility of an analogical inference,

the debate is usually not about whether the similarities that a given analogical argument hinges on *exist*, but rather whether the similarities that the analogical inference hinges upon are *relevant* to the conclusion.

An essential goal of a normative account of analogical inference is to help us differentiate between relevant similarities and irrelevant similarities, such that we can decide whether the similarities cited in an analogical inference warrant the inference's conclusion.

As such, Hesse, Bartha and Gentner's normative accounts of analogical inference all attempt to capture criteria for relevant similarity. Under all three accounts, relevant similarity (and thus an analogical inference's plausibility) is evaluated by judging the strength of an analogical inference's *vertical relations* and *horizontal relations*. The *vertical relations* are the relations holding within the source domain between the features of that domain. The *horizontal relations* are the relations holding between the two domains.

Hesse was the first to develop a tabular representation of analogical inference, where an analogical inference consists of vertical and horizontal relations. This representation has continued to be widely used in the analogical inference literature, and is used by Bartha and Gentner. A tabular representation of analogical inference with horizontal and vertical relations is illustrated below.

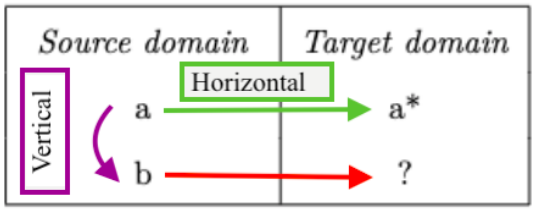


Figure 2.1: The tabular representation.

The vertical relation is a theory about what features of the source domain are relevant to the further property or proposition b holding in the source domain, and what the nature of that

relevance is[1]. The horizontal relation is a similarity relation between these relevant features of the source domain and some features of the target domain. An analogical inference argues that it is plausible that some analogue to b, $b*$, holds in the target domain because the properties of the source domain that are relevant to b holding in the source domain are relevantly similar to properties holding in the target domain. This analogical step—the analogical inference's conclusion—is illustrated by the red arrow in Figure 1.

The following example describes an analogical inference that was employed by early microbiologists, and shows how this inference can be expressed in the tabular representation framework.

Example 2.3.2. By the mid-19th century, vaccination had become a common practice[2]. For instance, in the American Civil War, both sides had regulations requiring that all soldiers be vaccinated against smallpox [Grabenstein 2006]. Though vaccination was widely practiced, scientists did not understand *why* vaccination worked. There was not yet any theory of the immune system and white blood cells had not yet been discovered. In an 1880 paper, Louis Pasteur proposed an explanation for the efficacy of vaccination by making an analogy from microbes cultured in a Petri dish [Pasteur 1880b]. Petri dishes were a relatively new innovation, and Pasteur had discovered that by allowing an organism to culture for an extended period of time, he could reduce its virulence. Supposedly, Pasteur made this discovery by chance when he accidentally left a dish of *P. multocida* (chicken cholera) culturing for a month while he was away on holiday [Barranco 2020]. Pasteur found that if at regular intervals one transferred a sample of the microbe to a fresh culture, then the virulence of *P. multocida* would remain constant over time. However, if one refreshed the culture medium less frequently, then virulence would decrease as the microbe used up the nutrients in the culture [Pasteur 1880b, pg. 128].

[1]For instance, the vertical relation could be a causal relation; the property a in the source domain could be relevant to b holding in the source domain because a causes b.

[2]Though inoculation was only introduced to Western countries in the 18th century, inoculation had already been a well-developed and common practice in Africa and Asia for thousands of years.

Describing the way that *P. multocida*'s growth stalled when placed in a culture medium lacking suitable nutrients, Pasteur says,

> N'est-ce pas l'image de ce qu'on observe quand un organisme microscopique se montre inoffensif pour une espèce animale à laquelle on l'inocule? (Isn't this the image of what we observe when a microscopic organism appears harmless to an animal species into which it is inoculated?) [Pasteur 1880b, pg. 129].

Pasteur argued that the mechanism of nutrient exhaustion that eventually inhibits microbe growth in a Petri dish could also explain the efficacy of vaccination. He argued that vaccination worked because infection by a particular microbe would eventually deplete the body's supply of the nutrients needed for sustaining that microbe [Smith 2005]. At the time, there was not yet a theory of the immune system, and the body was thought to be a passive environment like a Petri dish. Pasteur says,

> It seems as if the initial microbe inoculations have depleted a certain element that healing does not reconstitute and that the absence of which hinders the development of this small organism. This explanation will without doubt, become general and applied to all infectious diseases [Pasteur 1880b, pg. 135].

However, as medicine advanced, the strength of the horizontal relations began to come into doubt—evidence had still not been found of vaccinations depleting the body of certain nutrients. Additionally, white blood cells were discovered, and the body began to be conceived of as active, rather than a passive environment like a Petri dish. The final nail in the coffin came when veterinarian Henri Toussaint invented a vaccine that used dead, inert microbes yet achieved the same prophylactic effect as Pasteur's live attenuated vaccines. Because Toussaint's vaccine was inert, it was not possible that its efficacy was due to the vaccine metabolizing nutrients in the body [Smith 2012]. This suggested that a different mechanism

was at work in the target domain than in the source domain, and thus that there was not a strong horizontal relation between source and target domain.

Figure 2 shows a tabular representation of Pasteur's analogical inference.

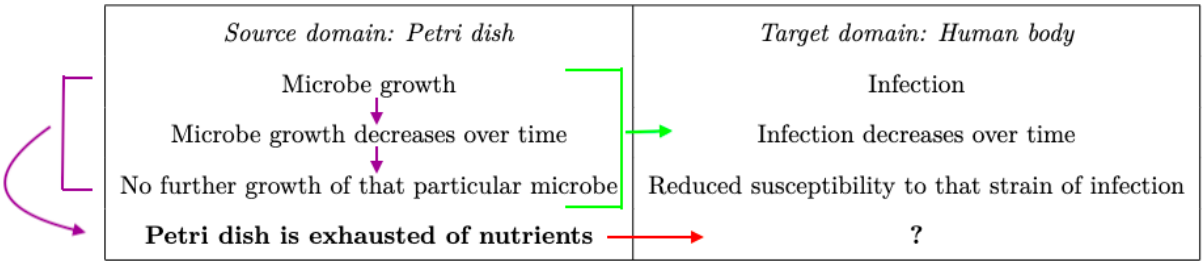


Figure 2.2: Pasteur's analogical inference

In the source domain, there is a vertical explanatory relationship between the explanans (the Petri dish being exhausted of nutrients) and the explicandum (the observed correlation between initial microbe growth, microbe growth decreasing over time, and finally no further growth). The observed phenomena are explained by the microbe growth depleting the Petri dish's supply of the nutrients necessary to support microbe growth. Pasteur presumed that there was also a strong horizontal relation between source and target domain—that the body was a passive environment like a Petri dish and that when a microbe grew in the body, it would deplete the body of the nutrients necessary to support that microbe. However, the strength of the horizontal relation came into doubt as critical differences between the source and target domain were discovered, such as white blood cells, and this led to the plausibility of the analogy being questioned.

How would Hesse, Bartha, and Gentner's normative accounts evaluate the plausibility of Pasteur's analogical inference? Their accounts agree that there are two essential characteristics that determine an analogical inference's plausibility. First, is there a relevant relationship

between the features of the source domain such that finding some features in the source domain is salient to the possibility of finding a further feature in the source domain? I call this *Source Domain Salience*. Second, is there a relevant horizontal similarity relation such that some of the salient features of the source domain have analogues in the target domain? I call this *Overlap*. Where the three accounts vary is in their criteria for answering these two questions. They agree that plausibility depends on these two essential characteristics, *Source Domain Salience* and *Overlap*, but disagree about the conditions under which these two conditions are satisfied. I will briefly outline these differences.

2.3.2 Source Domain Salience

Source Domain Salience is the requirement that there must be a relevant vertical relationship in the source domain such that finding some features is salient to the possibility of finding a further feature in the source domain. What kind of relationship is required? Bartha, Gentner, and Hesse vary in their answers to this question.

For Hesse, the vertical relation must be a causal relation. Under her account, an analogical inference is only plausible if "the vertical relations in the model are causal relations in some acceptable scientific sense, where there are no compelling *a priori* reasons for denying that causal relations of the same kind may hold between terms of the explicandum [Hesse 1966]." Some philosophers consider this requirement to be too strict [Bartha 2009].

In contrast, Bartha's account greatly refines and expands Hesse's criteria for the vertical relations in the source domain. First, Bartha's model allows that there can be other meaningful vertical relations besides causality that can hold in the source domain, such as mathematical dependencies or statistical relations [Bartha 2009]. Second, Bartha proposes finer-grained criteria for the vertical relations. Bartha specifies four different types of vertical relations (functional, predictive, correlative, and explanatory) and provides different criteria for sat-

isfying *Source Domain Salience* for each type.

Finally, under Gentner's account, an analogical inference consists of nodes (objects or features in the source and target domain) and two kinds of predicate. First, single-place predicates that predicate some property of a node. And second, two-place predicates that describe a relation between nodes. For Gentner, the salient features of the source domain are the relations between nodes, rather than node attributes. Gentner gives the example of a battery and a reservoir. The strength of the analogy between a battery and a reservoir depends on the battery and the reservoir being similar structurally; both have a mechanism by which they store potential energy, flow can be turned on to release energy to power systems, etc [Gentner 1983, pg. 156]. In contrast, non-structural attributes are irrelevant to the strength of the analogy; plausibility does not increase because batteries and reservoirs are both cylindrical, nor decrease because they are different sizes [Gentner 1983, pg. 156].

Under Gentner's account, *Source Domain Salience* is satisfied if an analogical inference's source domain has relations holding between its nodes, and especially if these relations are higher-order relations that cause and constrain lower-order relations[3].

2.3.3 Overlap

The second condition for an analogical inference's plausibility, *Overlap*, is the requirement that there must be a horizontal similarity relation holding between relevant features of the source domain and corresponding features of the target domain.

Hesse's account includes two criteria for *Overlap*. First, her material requirement: there must be observable or "pre-theoretic" similarities between the objects in the source and

[3]A higher-order relation is one that predicates relations between relations, rather than relations between objects. For example, a higher-order causal/explanatory relation holds between the relation 'revolve-around' holding between the planets and the sun and the relation 'gravitationally-attracts' holding between the sun and the planets [Gentner 1983].

target domain. For example, there are observable similarities between fluid waves, sound, and light that exist independently of there being any detailed scientific theory about each phenomenon— they all appear to be produced by movement and have properties of reflection and diffraction [Hesse 1966, pg. 11].

Second, Hesse proposes that none of the essential features of the causal relationship in the source domain should be known to fail to hold in the target domain [Hesse 1966]. It must be the case that "the essential properties and causal relations of the model have not been shown to be part of the negative analogy between model and explicandum [Hesse 1966, pg. 91]."

In contrast, Bartha sees Hesse's material requirement as being too strict–he points out that there are plausible analogical inferences where there are not many material similarities between source and target domain, and implausible analogical inferences where there are a substantial number of material similarities between source and target [Bartha 2009]. Instead, Bartha proposes the less restrictive *potential for generalization condition*, which says that (1) some of the elements of the vertical relation in the source domain must have analogues in the target domain, and (2) it must be the case that none of the elements that are critical to the vertical relation holding in the source domain are known to fail to hold in the target domain [Bartha 2013, 3.5.2]. Bartha has a detailed framework for what counts as a 'critical' element of the vertical relation for each of the four types of vertical relations specified in his model.

For Gentner, satisfaction of *Overlap* is determined by the quantity of higher-order relations that can be mapped from source to target domain. Gentner calls this 'systematicity'. Bartha points out that Gentner's account has a shortcoming; sometimes the crucial drivers of an analogical inference's plausibility are object attributes rather than higher-order relations between objects [Bartha 2009]. For instance, in many ethnographic analogies, physical similarities between a contemporary tool and artifact are used to infer the function of the

artifact from the function of the contemporary tool [Bartha 2009]. In this case, the strength of the analogy depends on the *properties of nodes* being mappable to the target domain, not higher-order structural relations between nodes. Another challenge for Gentner's account is that sometimes higher degrees of structural similarity can be irrelevant or decrease the plausibility of an analogical inference, as I will show in section 3.

Though their accounts vary in how they go about capturing *Source Domain Salience* and *Overlap*, Hesse, Bartha, and Gentner's accounts all treat these as being the two essential characteristics of a plausible analogical inference. As I discussed in Section 1, the essential goal of a normative account of analogical inference is to capture criteria for relevant similarity. Bartha, Gentner, and Hesse claim that relevant similarity can be captured by giving the right criteria for these two essential characteristics, *Source Domain Salience* and *Overlap*. I will call the conjunction of these two conditions the *relevant similarity framework*.

One important caveat, before going further, is that these accounts do not claim that these two essential characteristics are necessary and sufficient conditions for strong plausibility. Rather, they treat them as necessary and sufficient conditions for modest, *prima facie* plausibility.

For Bartha, this means that if an analogy satisifes the prior association and potential for generalization conditions, then "(1) It has epistemic support: an appreciable likelihood of being true (or "successful") and (2) It has pragmatic importance: it is worth investigating, granted the feasibility of and interest in investigation [Bartha 2009, pg. 18]."

Gentner claims that there are three components of evaluating an analogical inference's plausibility: structural evaluation, factual evaluation, and evaluation of whether the analogy aligns with scientists' goals. Her research is on structural evaluation, and she claims that of these three components, structure plays the most significant role in determining plausibility. In this sense, my account is not in opposition to Gentner's theory, since she does acknowledge that factors not captured by her systematicity principle do affect plausibility.

2.4 Questioning the relevant-similarity framework

The relevant-similarity framework works well for correctly evaluating many analogical inferences. However, I will show that it fails to capture a third essential characteristic that scientists use to evaluate analogical inferences, *Target Domain Salience*. In section 3.1, I will show that satisfying *Source Domain Salience* and *Overlap* is not a sufficient condition for an analogical inference having *prima facie* plausibility. Next, in section 3.2, I will show that relevant *dissimilarity* can sometimes increase an analogical inference's plausibility, such that failing to satisfy *Source Domain Salience* and *Overlap* can sometimes increase the plausibility of an analogical inference.

2.4.1 The relevant-similarity framework is not sufficient for plausibility

In this section I will show that *Source Domain Salience* and *Overlap* fail to capture all of the criteria that scientists consider when determining whether an analogical inference is plausible. Sometimes scientists agree that there is a relevant relationship in the source domain and agree that the essential features of that relationship can be found in the target domain, yet do not find the analogical inference plausible.

I will use two examples to demonstrate this. The first example is the Central Cattle Pattern, a prominent ethnographic analogy from archaeology, and the second example is Einstein's famous elevator thought experiment.

Example 2.4.1. The Central Cattle Pattern (CCP) is a type of spatial layout found in southern African herding villages that has been widely discussed in the archaeology literature and has been used by archaeologists as a model for interpreting Iron Age village sites in the same region. Archaeologists identified a strong correlation between the social organization

and spatial organization of Tswana herding villages in southern Africa [Huffman 1982, 1990]. Spatially, the villages are divided up into wards which each consist of dwellings arranged in a circle. The spatial position of each family's dwelling in the ward is determined by their degree of relatedness to the village headman. The spatial center of the ward, at the nucleus of the ring of dwellings, is also the social center of the ward, and is the location for marriages, trades, and legal and political matters. Also located at the center of the ward is a *kraal* of cattle that have a symbolic role as a representation of the status and wealth of the ward, and also the functional role of being used in trades or as bridewealth in the event of a marriage [Fewster 2006, pg. 64]. The wards have inner and outer sections which each correspond with status and gender. The center of the ward is occupied by men and contains high-status graves, communal storage bins for grain, and cattle. Women are relegated to the outer section of the ward, which is the location of their sleeping places, the ward's kitchens, and low-status burials [Huffman 2001, pg. 3].

It is a common principle of archaeology that there is a structural-functional relationship between a society's beliefs, norms, and cultural practices and its spatial organization. The layout of a society's physical spaces constrains what social behaviours manifest, and a society's social organization shapes the layout of its physical spaces. As a result, information about a society's social organization can be inferred from its spatial organization, and vice versa. Inferences about social organization from spatial organization are thought to be especially convincing in cases where the society in question is historically linked to a contemporary society that has similar spatial organization. In these cases, archaeologists sometimes infer that the ancient society had a similar social organization to the contemporary, historically-linked society. This is known as the direct historical approach. Huffman summarizes this principle, saying: "Since social and settlement organisation are different aspects of the same thing and products of the same view of the world, a continuity in settlement pattern is evidence for continuity in social organisation and worldview [Huffman 2014, pg. 657]."

Huffman argues that there is a distinct symmetry between the social and spatial organization in Central Cattle Pattern (CCP) villages that can be justified as a uniformity in villages across the region. The spatial layout of the village reflects its social characteristics, such as a patrilineal social hierarchy, a bridewealth system centered around cattle, and the exclusion of women from the political activities of the village.

Because the CCP model demonstrates such a strong uniformity in the relationship between social and spatial organization in both contemporary villages and in nineteenth and twentieth century historical records of CCP-type villages, Huffman argued that this relationship can be generalized to assume that a bridewealth system revolving around cattle existed in all historically linked settlement sites that have a distinct CCP layout. Historical records show that the symmetry between social organization and physical organization articulated by the CCP is known to hold in all villages in the region for at least the past 150 years [Lane 2005, pg. 31].

Huffman and other archaeologists have used this structural-functional uniformity to make inferences about the social organization of ancient CCP-type settlement sites in the area. This includes early sites like Toutswe, a large tenth-to-fourteenth century site that is located in the same region as the contemporary CCP villages. Huffman argues that the cattle-centered bridewealth social order associated with the CCP layout in contemporary settlements can be generalized to these Iron Age CCP-type sites. "If the structural relationship between language, culture and spatial organization is valid, it follows that archaeological evidence for the Central Cattle Pattern is sufficient to demonstrate the existence of cattle-based bridewealth in the past [Huffman 1993, pg. 220]."

Figure 3 shows a tabular representation of the CCP analogical inference:

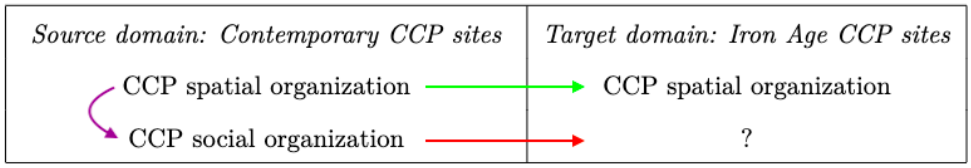


Figure 2.3: The Central Cattle Pattern analogical inference

This analogical inference has a source domain of contemporary CCP-type villages and a target domain of Iron Age CCP-type sites in the same region that have a direct historical connection to the contemporary villages. The argument is that in the source domain there is a strong structural-functional relationship between CCP-type village layout and the function of the layout reflecting and facilitating the patrilieal, cattle-based bridewealth social organization of the villages. This is thought to be a robust, stable relationship because it holds for all contemporary CCP-type villages across the region. Additionally, it appears to be a stable relationship across time because historical records from the past 150 years show that all recorded CCP-type villages in the region within that time range were patrilineal and had a cattle-based bridewealth system.

There are two main horizontal similarity relations: (1) the spatial organization of contemporary CCP-type sites can be generalized to Iron Age CCP-type villages in the region and (2) the contemporary sites and Iron Age sites belong to the same lineage. On the basis of these similarities, Huffman infers that the social organization of contemporary CCP-type sites could also plausibly be generalized to these ancient Iron Age CCP-type sites. In the source domain, the physical characteristics of the CCP-type settlements and the region where the settlements are located both explain and are explained by the function of the CCP layout, which is facilitating the settlement's patrilineal, cattle-based bridewealth social organization. On the basis of the similarities between the features in the source and target domain, archaeologists hypothesize that a similar explanatory relationship holds in the target domain. That is, the analogous physical characteristics and location of the Iron Age CCP settlements

could be explained by the settlements also having a patrilineal, cattle-based bridewealth social organization.

Do *Source Domain Salience* and *Overlap* successfully capture the criteria that archaeologists use when evaluating an analogical inference like this one? Under Hesse, Bartha, and Gentner's accounts, evaluation would consist of two steps. The first step would be to check whether the Central Cattle Pattern analogical inference satisfies *Source Domain Salience*, by determining whether the relationship holding between CCP structural layout and social organization is a strong, uniform, and widespread relationship in the source domain. The second step would be to evaluate *Overlap*—the degree to which the salient features of the source domain can be found in the target domain.

But there is a puzzle. Archaeologists *agree* that there is a robust structural-functional relationship between spatial organization and social organization in contemporary CCP-type villages. They also agree that an identical spatial layout can be found in Iron Age CCP settlements. However, they *disagree* about whether the structural similarities between contemporary and Iron Age CCP-type villages justify the conclusion that the social organization of contemporary CCP-type settlements can be generalized to Iron Age CCP-type settlements [Fewster 2006; Lane 2005; Reid 1997]. In other words, they agree that the CCP analogical inference satisfies *Source Domain Salience* and *Overlap*, but they disagree about the analogical inference's plausibility. Thus, the source of their disagreement is not captured by Bartha, Gentner, and Hesse's normative accounts; they disagree about a criterion that falls outside the scope of the relevant-similarity framework.

So what is the source of their disagreement? Archaeology underwent a paradigm shift between the processual (1960s-1980s) and post-processual (post 1980s) eras. The processual era of archaeology was characterized by a positivist approach including the view that archaeologists discover explanations rather than creating them, the idea that archaeology should and can be impervious to non-epistemic values, and the notion that changes in human culture

are largely reducible to environmental change rather than human agency.

Starting in the 1980s, there was a shift to the post-processual ideology, which was a collection of viewpoints that critiqued many of the principles held by processualists. Post-processual archaeology is characterized by the view that human culture is irreducible to changes in environmental conditions, awareness that non-epistemic values can influence archaeologists' interpretations of material remains, and rejection of the idea that human behaviour can be objectively inferred from the archaeological record. Post-processualists employ a more careful approach towards evaluating ethnographic analogies, particularly looking for ethnographic analogies that are reductionist about human culture or founded on androcentric or ethnocentric biases. This has resulted in some ethnographic analogies that were widely accepted by the processualists being rejected by some post-processualists. The Central Cattle Pattern is one such example.

Post-processual archaeologists have two main objections to the CCP analogical inference. First, Reid et al.[1997] argue that colonialism could have destabilized the structural-functional relationship identified in the source domain because social changes resulting from colonialism would not necessarily be reflected by material changes in CCP village layout. Reid argues that spatial layout is salient to social organization in the source domain, but colonialism is a factor that might destabilize this relationship in the target domain. Even though analogous spatial features hold in Iron Age CCP-type settlements, these features are not necessarily salient to the possibility of a contemporary CCP-type social organization.

The second objection to the CCP analogical inference is that the similarities between the source and target domain may only seem salient to the conclusion in the context of a problematic background assumption. Lane [1994] argues that the CCP analogy promotes the stereotype that African societies are passive and merely respond to environmental changes rather than actively changing through human creativity and innovation [Lane 1994, pg. 271]. The concern is that the veracity of the analogical inference's conclusion could be distorted

by scientists' non-epistemic values—in particular, the stereotype of African societies being passive and unchanging.

In cases of underdetermination, social values can bridge the gap between theory and evidence [Douglas 2014]. One recent example is a warrior chamber grave in Birka, Sweden which since its excavation in 1878 has been assumed to be the grave of a high-status male. In 2017, a team led by a woman archaeologist conducted genomic research which revealed that the Viking warrior was actually female [Price 2019]. This discovery prompted reexamination of other graves which has revealed other instances of incorrect identification of sex. Before this study, contemporary gender norms had bridged the gap between theory and evidence in sexing graves. The Birka study, and other subsequent studies, suggest that contemporary gender roles may not be a fruitful background assumption to employ in interpreting Viking graves.

In the CCP example, Lane's concern is that the salience of the structural similarities between contemporary and Iron Age CCP-type sites may be hinging on the background assumption that African societies are unchanging and non-generative, and this background assumption is not one which is epistemically fruitful [Lane 1994; Fewster 2006].

Both of these objections to the CCP analogical inference fall outside the scope of Hesse, Bartha, and Gentner's two essential characteristics. Processualists and post-processualists agree that the CCP analogical inference satisfies the conditions *Source Domain Salience* and *Overlap*. They agree that spatial organization is salient to social organization in the source domain. They agree that the salient features of the source domain have analogues in the target domain. What they disagree about is whether the features that are salient to social organization in the source domain are also salient to social organization where they occur in the target domain. I call this *Target Domain Salience*.

Definition 2.1. *Target Domain Salience:* The features of the target domain are salient

to the possibility of the source domain's further feature being generalizable to the target domain.

Bartha, Hesse, and Gentner's accounts assume that *Target Domain Salience* can be established by evaluating *Source Domain Salience* and *Overlap*. This is a problem for their models because scientists use additional criteria to evaluate *Target Domain Salience*.

Bartha (and to a lesser degree, Hesse and Gentner) acknowledge that there is variation in scientific communities' criteria for evaluating analogical inferences. However, they assume that this variation occurs at the stage of evaluating the salience of the relationship in the source domain, and that the salience of the relationship in the *target domain* is invariant across the different background assumptions of different scientific communities. Bartha says:

> Standards for what counts as an acceptable causal or explanatory relationship have varied greatly over time. By contrast, the evaluation of whether or not we have justification for transferring such a relationship to a new setting can be relatively independent of historical milieu [Bartha 2009].

Bartha's assumption is that evaluation of salience occurs at the stage of evaluating the source domain, and that once *Source Domain Salience* is established, evaluation of plausibility consists of establishing *Overlap*. Hesse and Gentner's accounts rely on the same assumption.

This assumption is incorrect. Even within a given scientific community, the criteria for salience in the target domain can be independent from the criteria for salience in the source domain.

Consider another example in which the plausibility of an analogical inference is influenced by *Target Domain Salience* and so falls outside the scope of the relevant similarity models.

Example 2.4.2. Einstein makes an analogical inference between two elevators. One elevator

is at rest on a planet in a gravitational field, such as on Earth. The second elevator is outside of a gravitational field in outer space, accelerating uniformly upwards. Inside of the zero-gravity, accelerating elevator, there is a man, and he inspects his circumstances inside the elevator. To stay upright, he must continually exert force with his legs. He finds that if he drops an object that was previously held in his hand, the object will approach the floor with accelerated motion relative to the elevator and himself. He repeats this experiment with a few objects of different mass and finds that the objects' rate of acceleration towards the floor is the same for all of the objects. Drawing on his knowledge of gravitational fields, he concludes that it must be the case that he and the windowless elevator are in a gravitational field and that the dropped objects accelerated towards the floor because of their gravitational mass.

Einstein says,

> Ought we to smile at the man and say that he errs in his conclusion? I do not believe we ought to if we wish to remain consistent; we must rather admit that his mode of grasping the situation violates neither reason nor known mechanical laws. Even though it is being accelerated with respect to the "Galileian space" first considered, we can nevertheless regard the chest as being at rest [Einstein 1917/2015].

All of the man's observations are consistent with his elevator being at rest in a gravitational field [Einstein 1917/2015]. Standing up feels the same, dropping an object produces an identical effect, etc. The conclusion of Einstein's analogical inference is that given the empirical indistinguishability of the phenomena in the gravitational and accelerating elevators, gravitational mass is a plausible explanation for the inertial phenomena the man in the accelerating elevator is experiencing. In the at-rest elevator, the explanation for the phenomena is gravitational mass, and in the accelerating elevator the explanation for the

phenomena is inertial mass. Einstein concludes that in these circumstances in which the two explanations exhibit a physical equivalence, we can move between the two explanations.

Figure 4 shows a tabular representation of Einstein's analogical inference.

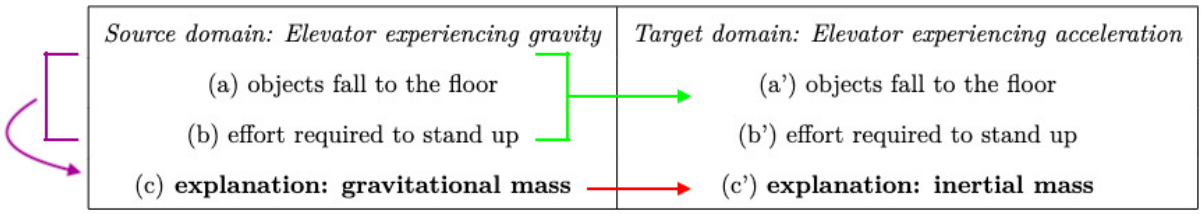


Figure 2.4: Einstein's analogical inference

Above, we can see Einstein's analogical inference represented in a table. The empirical phenomena that someone in an elevator at rest in a gravitational field would observe are *a* and *b*. The observational phenomena that the man is experiencing in the zero-gravity, accelerating elevator are *a'* and *b'*. The gravitational effects *a* and *b* are empirically indistinguishable from the inertial effects *a'* and *b'*. As a result, Einstein infers that we can plausibly generalize the explanation for the gravitational phenomena to explain the inertial phenomena. Under the empirically equivalent circumstances in the source and target domain, the phenomena that the man experiences in the target domain can interchangeably be explained by *c*, inertial mass, or *c'*, gravitational mass.

At first glance, it seems that this is a straightforward case of *Source Domain Salience* and *Overlap* powering the analogical inference's plausibility. *Source Domain Salience* is satisfied by the well-established explanatory relationship in the source domain between the phenomena *a* and *b* and the explanation *c*, gravitational mass. *Overlap* is easily satisfied because there is not merely similarity between the phenomena {*a, b*} and {*a', b'*}; there is complete indistinguishably. As a result, at first glance it appears that this is a textbook case of ana-

logical inference in which its plausibility is powered by it satisfying *Source Domain Salience* and *Overlap*.

However, the analogical inference is actually a bit more complex than that. Prior to Einstein's analogical inference, it was already known that gravitational mass and inertial mass are numerically equivalent. The fact that an object's gravitational mass and inertial mass are always equal was first demonstrated by Newton, who did an experiment with pendulums to show that the period of oscillation is independent of the mass of the object hanging from the pendulum's string. This showed that the inertial mass and gravitational mass of the swinging object always cancelled each other out, establishing that inertial and gravitational mass are always equal [Newton 1687/1846, book III, theorem VI]. At the time of Einstein's analogy, physicists already knew and accepted that an object's inertial and gravitational mass are always equal, but the reason for this numerical equivalence had not yet been explained and so scientists still treated inertial and gravitational mass as distinct theoretical entities.

Einstein's thought experiment did not aim to establish that inertial and gravitational mass are equivalent theoretical entities merely by reiterating their numerical equivalence. His thought experiment added an extra ingredient that established the *salience* of their numerical equivalence to the possibility of them being theoretically equivalent. In Einstein's analogical inference, a domain-specific background assumption does the heavy lifting in making *Source Domain Salience* and *Overlap* relevant to the analogical inference's conclusion. Norton [1991] discusses the fact that many of Einstein's famous thought experiments were made during the height of logical positivism and that many positivist assumptions show up in his thought experiments. Though Norton does not frame Einstein's thought experiment as an analogical inference, he argues that Einstein's thought experiment assumes something like the logical positivist's verifiability principle: that the only theoretical entities that are meaningful are those that are empirically verifiable [Norton 1991, pg. 135].

Under the verifiability principle, a theoretical distinction between gravitational mass and

inertial mass is only meaningful if we can observationally distinguish between them. The argument goes: since we cannot empirically distinguish between gravitational mass and inertial mass (as demonstrated by the thought experiment) it is not meaningful to theoretically distinguish between them.

This positivist principle is one that plays a role in some of Einstein's other work from that period of his career. In a letter to Moritz Schlick in 1915 responding to Schlick's paper on relativity, Einstein says, "Your argument that positivism suggests the theory of relativity...is also very right...This line of thought had a great influence on my efforts [Einstein 1915/2005]." Schlick describes the role that the verifiability principle played in Einstein's rejection of the concept of absolute simultaneity:

> Einstein said to the physicists (and to the philosophers): you must first state what you *mean* by simultaneity, and you can do this only by showing how the proposition "two events are simultaneous" is verified. But with this you have *completely* determined its meaning...If one should say that it did contain something more he must be able to say what more this is, and to do this he would have to tell us how the world would differ if he were mistaken. But this cannot be done, because by assumption all the observable differences are already included in the verification [Schlick 1948, pg. 90].

How does the role that the verification principle played in Einstein's analogical inference fit into the framework described by Bartha, Gentner, and Hesse? Under their accounts, an analogical inference's plausibility is determined by it satisfying *Source Domain Salience* and *Overlap*. So under their accounts, if Einstein's analogical inference established a convincing case for the equivalence of gravitational and inertial mass, it must have done so by strengthening *Source Domain Salience* or *Overlap*. But Einstein's analogical inference did not help establish *Source Domain Salience* or *Overlap*; both were already established prior

to Einstein's analogy[4]. Instead, Einstein's analogy provided an argument for *why* the satisfaction of *Source Domain Salience* and *Overlap* should be taken as evidence for theoretical indistinguishability.

Another way of saying this is that the verifiability principle established *Target Domain Salience*. It provided an answer to the question of why the similarity between the source and target domain is salient to the possibility of the source domain's explanation being generalizable to the target domain.

In summary, the first moral of the Einstein elevator example is that *Source Domain Salience* and *Overlap* do not give sufficient conditions for plausibility. In Einstein's elevator analogy, the analogical inference satisfying *Source Domain Salience* and *Overlap* was not sufficient for its conclusion being considered plausible. Instead, on Norton's account, it was a background assumption, the verifiability principle, that convinced Einstein of the plausibility of the analogical inference [Norton 1991].

A further moral to be drawn from these two examples is that the background assumptions that do the heavy lifting in making an analogical inference plausible are often local and context-dependent. In the CCP example, processual and post-processual archaeologists disagree about the plausibility of the CCP analogical inference because they disagree about the legitimacy of the underlying social and epistemic values powering the analogy. In the Einstein elevator example, the verifiability principle background assumption is of course not invariant across scientific communities; it is very much one which is local to the time and context in which the analogy was presented. Einstein's commitment to this background assumption weakened over the course of his career. One illustration of Einstein's departure from positivism can be seen in his endorsement of hidden variable theories in his later work

[4]Though perhaps someone could could make the argument that the verifiability principle helped establish a stronger case of *Overlap*. This would be a challenging argument to make, because it would be difficult to argue that there can be a stronger similarity relation than physical equivalence/indistinguishability. Perhaps someone could argue that the verifiability principle strengthened *Overlap* because it established a theoretical similarity relation in addition to the already-existing physical similarity relation.

in quantum mechanics, where his commitment to the concept of an objective reality superseded his commitment to the idea that a theory's only meaningful postulates are those which are in-principle verifiable. In a 1948 letter to British physicist Donald Mackay, Einstein says "At that time [of developing the theory of special relativity] my mode of thinking was much nearer positivism than it was later on...My departure from positivism came only when I worked out the general theory of relativity [Isaacson 2017]."

2.4.2 Relevant dissimilarity can increase plausibility

A second challenge for Bartha, Hesse, and Gentner's accounts is that *Source Domain Salience* and *Overlap* are not necessary conditions for *prima facie* plausibility. Sometimes scientists judge the strength of an analogical inference to be greater as the target domain becomes increasingly dissimilar to the source domain. Scientists' background assumptions can weigh more heavily than *Source Domain Salience* and *Overlap* in their evaluations of plausibility. Furthermore, this seems to be an epistemic virtue. I will argue for this claim using examples from stimulus generalization research and astrobiology.

Example 2.4.3. Stimulus generalization research looks for patterns in the conditions under which humans, non-human animals, and plants will generalize learned behaviour to novel situations and stimuli. One study from Hanson [1959] involves training pigeons to peck at a button when a positive stimulus (550 nanometer light) is displayed, and to not peck at the button when a negative stimulus (570 nanometer light) is displayed. The pigeons are rewarded if they peck at 550nm, and are not rewarded (but also are not punished) if they peck at 570nm. After this initial training period, researchers expose the pigeons to novel, untrained wavelengths to see whether the pigeons generalize the pecking behaviour to wavelengths similar to the positive stimulus. What the researchers found was surprising. We would expect that the pigeons would display the pecking behaviour most often for the positive stimulus they had been trained to, 550 nanometer light. But instead, the pigeons consistently

preferred to peck when exposed to 540 nanometer light. This finding is illustrated in Figure 5.

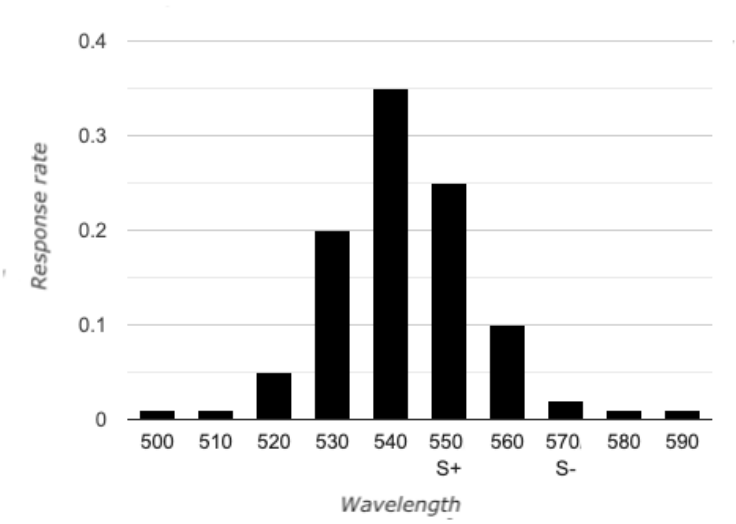


Figure 5: Response bias [redrawn from Ghirlanda et al. 2003; original data from Hanson 1959].

The explanation for this phenomenon is that the pigeons had developed a response bias from being exposed to the negative stimulus [Ghirlanda et al. 2003]. The pigeons inferred that pecking is likely to be a successful behaviour when the stimulus is similar to 550nm, but is unlikely to be a successful behavior when the stimulus is similar to 570nm. Thus, they are more likely to find a generalization plausible if it is slightly further away from S- than S+ is. This response bias was also found in a range of other animals, including humans, in other studies [Ghirlanda et al 2003].

This response bias seems to be a virtue rather than a vice; we want to be careful in our inferences and only act when our action has a high probability of being successful. So it makes sense that we weigh relevant *dissimilarity* as well as relevant similarity when deciding whether a generalization is plausible.

Consider an example of a case where scientists exhibit this kind of epistemic virtue—where relevant dissimilarity between the source and target domain increases scientists' credence in an analogical inference's plausibility.

Example 2.4.4. One paradigmatic use of analogical inference in science is by astrobiologists hypothesizing about what other planets are capable of supporting life. Astrobiology has a long tradition of employing analogical inference; for instance, Kant used an analogical inference from deserted islands and deserts on Earth to argue that the apparent uninhabitability of planets like Jupiter does not imply that most of the other planets in the universe are uninhabitable [Kant 1755/2012, III]. Analogical inferences continue to be an important tool in contemporary astrobiology research. Some astrobiologists see analogical inference as essential for overcoming the N=1 problem, the problem of how we can make any generalizations about what factors give rise to life when our available sample consists of only Earth [Sullivan 2013].

I will show that relevant dissimilarity plays a key role in astrobiologists' analogical inferences. The Planetary Habitability Index (PHI), proposed by Schulze-Makuch et al. [2020], ranks planets according to how likely they are thought to be to support life [Bora et al. 2016]. On this index, some planets that lack the features that gave rise to life on Earth rank higher (are thought more likely to support life) than planets that are highly similar to Earth.

Consider two KOIs on the Planetary Habitability Index, KOI-4878.01 and KOI-5715.01. KOI stands for 'Kepler Object of Interest' and describes planetary objects that are presumed to exist on the basis of periodic blinking of stars observed through the Kepler space telescope. KOI-4878.01 is thought to be highly similar to Earth: it orbits a spectral-G type dwarf star, has a similar mass and volume to Earth, has a similar orbital period, and has a global mean surface temperature similar to Earth's [Chawda 2024]. On the Earth Similarity Index (ESI), KOI-4878.01 has a score of 0.98 where a score of 1.0 would signify a maximally Earth-like planet. Chawda et al. say "If confirmed, KOI-4878.01 would be one of the most

Earth-like planets found so far [Chawda 2024]." In contrast, KOI-5715.01 has some marked dissimilarities to Earth; it orbits a spectral-K type dwarf star, is thought to be older than Earth, and is approximately twice the size of Earth [Schulze-Makuch et al. 2020]. On the ESI, KOI-5715.01 has a score of 0.75, meaning that it falls well below the threshold of 0.9 for potentially Earth-like planets.

Of these two candidate planets, KOI-4878.01 has close analogues for the factors that contributed to life existing on Earth, whereas KOI-5715.01 lacks close analogues and has some features that substantially differ from Earth's features. Yet on the Planetary Habitability Index (PHI), KOI-4878.01 has a lower ranking than KOI-5715.01, meaning that astrobiologists hypothesize that KOI-4878.01 has a lower probability of being habitable than KOI-5715.01. Schulze-Makuch et al. argue that when allocating research resources, planets like KOI-5715.01 that are different from Earth "may deserve higher priority for follow-up observations than most Earth-like planets [2020]."

Why does KOI-5715.01 rank higher than KOI-4878.01 on the PHI yet rank lower than KOI-4878.01 on the ESI? For one, KOI-5715.01 has a spectral-K type dwarf star. Spectral-K type dwarf stars tend to have greater longevity than spectral-G type dwarf stars, so a planet with a spectral-K type dwarf star would have more time for life to develop before its star dies [Schulze-Makuch et al. 2020, pg. 1396]. Spectral-K type dwarf stars are also more stable and emit less radiation and solar flares, so a planet with a spectral-K star would provide a more gentle environment for fragile early lifeforms. Likewise, its other features like larger mass and slightly warmer temperature than Earth's better satisfy astrobiologists' theories about what factors are conducive to life [Schulze-Makuch et al. 2020]. These properties—spectral-K dwarf star, age, warmer temperature, larger mass—that satisfy astrobiologists' background assumptions about what properties are conducive to life, are properties that make KOI-5715.01 *different* from Earth. Given that Earth is the only available source domain for analogical inferences about what factors give rise to life, how can it be that a planet that is

different from Earth (KOI-5715.01) is thought to be more habitable than a planet which is highly similar to Earth (KOI-4878.01)?

Earth is our only reference point for what factors contribute to making a planet habitable to life; this is astrobiology's N=1 problem. As a result, our theories about what factors make a planet habitable necessarily have an Earth-centric bias [Schulze-Makuch 2020, pg. 1395]. Many of the conditions that contributed to life existing on Earth could be irrelevant in the scope of the universe. For instance, Schulze-Makuch [2020] points out that "Habitability in a wider sense is not necessarily restricted to water as a solvent or to a planet circling a star. For example, the hydrocarbon lakes on Titan could host a different form of life...If molecules other than water (e.g., ammonia) are considered as a solvent for an alien biochemistry, the types of possible habitats become highly speculative and increase enormously in quantity [pg. 1395]."

Given that Earth is our only source domain for making analogical inferences about the potential habitability of candidate exoplanets, some degree of Earth-centric bias is unavoidable. However, using degree of similarity to Earth as our metric for evaluating the habitability of candidate exoplanets amounts to the assumption that the factors that gave rise to life on Earth are the *ideal* conditions for life to originate. This is clearly wrong—we do not want to assume that out of all of the planets in the universe, Earth is the most suitable possible planet for supporting life [Schulze-Makuch et al. 2020].

Accordingly, astrobiologists do not treat Earth as the ideal. Instead, they extrapolate hypotheses about what factors could make a planet habitable, and they evaluate potential exoplanets by whether they satisfy these hypotheses. Astrobiologists recognize that the factors that gave rise to life on Earth could hinder life on another planet, and vice versa. For instance, water is essential to life on Earth, but on another planet with a different biochemistry, water could impair life and a less corrosive solvent like formamide might be preferable [Adam et al. 2018; Saladino et al. 2012]. In summary, even though Earth is the only

available source domain for analogical inferences about the habitability of other planets, it is possible for planets (like KOI-5715.01) to be plausible candidates for habitability *because* of their dissimilarities to Earth, not in spite of them.

Can Hesse, Bartha, and Gentner's accounts handle such cases where relevant dissimilarity increases the plausibility of an analogical inference?

Gentner's account cannot capture relevant dissimilarity because under her account, plausibility is determined by the number of relations (especially higher-order relations) that can be mapped from source to target domain [Gentner 1983]. The number of relations that can be mapped to the target domain decreases as the target domain becomes increasingly dissimilar to the source domain, so relevant dissimilarity would decrease plausibility under Gentner's account [Bartha 2009, pg. 71].

For instance, if a planet had formamide as a solvent instead of water, this would constitute a relational dissimilarity because formamide differs from water in how it dissolves molecules, which molecules it dissolves, and how it causally interacts with other aspects of a planet's habitability. For instance, formamide breaks down some ionic compounds more readily than water does, but is generally less corrosive than water and is able to keep the constituent molecules of RNA stable in a way that water cannot [Adam et al. 2018].

A planet's solvent also affects the habitable zone of a planet, the range of orbits around a star at which planets' surface temperatures allow their solvents to be in a liquid state. If a planet's orbit is close to the sun, the planet's solvent can experience higher levels of evaporation which causes a greenhouse effect, further increasing surface temperatures and reducing the planet's potential habitability. However, formamide has a lower vapor pressure than water and so will evaporate more slowly than water under the same temperature conditions. Thus, if two planets, A and B, both have a short orbital distance from a large star such that they are getting significant radiation, and planet A has water while planet B has formamide,

then Planet B will have a larger habital zone than Planet A. Then, all other factors being equal, Planet B will be a more plausible prospect for habitability than Planet A, by virtue of formamide creating less greenhouse effect than water would in the same temperature conditions.

Under Gentner's account, plausibility should decrease if the number of correspondences between the source and target domain decrease. The formamide example shows that sometimes lower systematicity results in greater plausibility.

Hesse's account encounters a similar challenge. Under her account, if the features of the source domain that were essential to life developing on Earth are not present in the target domain, then this would reduce the plausibility of habitability being generalizable to the target domain.

Bartha provides an astute account of how dissimilarities can increase plausibility [Bartha 2009, pg. 71], but only considers one type of dissimilarity. Bartha makes a distinction between contributing and counteracting causes, and points out that eliminating similarities between the source and target domain "contributes to plausibility when what is eliminated is a counteracting cause [Bartha 2009, pg. 71]."

Bartha illustrates this with the example of life on Mars:

> To illustrate, let us consider the analogical argument that life exists or has existed on Mars...The best source domains for this analogy are frozen lakes in Antarctica or glaciers in Greenland, where microbes have been found to thrive despite the cold. The word "despite" is crucial here. Freezing temperatures are preventive or counteracting causes; they are negatively relevant to the existence of life [Bartha 2009, pg. 71].

If Mars is warmer than glaciers in Greenland that sustain life, then this should not decrease

the plausibility of Mars having life. To handle this, Bartha's account only requires that the relevant features of the source domain that *contribute* to the further feature holding in the source domain need to be generalizable to the target domain. Irrelevant features of the source domain that *prevent* the further feature from holding in the source domain need not be found in the target domain for the analogical inference to be plausible. The claim is that reducing the degree of overlap between source and target domain by removing a preventative cause should not reduce the plausibility of an analogical inference. This is in line with findings from psychology studies investigating how people reason about preventative causes when making analogies [Holyoak 2008].

This makes Bartha's account able to handle certain types of dissimilarity–namely, *irrelevant* dissimilarity—cases where features of the source domain that are not salient to the further feature holding in the source domain are missing from the target domain. However, this does not help resolve the case of *relevant* dissimilarity: cases where either (1) the source domain has a contributing feature that is absent in the target domain without reducing plausibility, (2) the target domain has a feature that is absent in the source domain yet contributes to target domain salience, or (3) a feature that is a contributing cause in the source domain is a counteracting cause in the target domain or vice versa.

Some examples of these scenarios are:

(1) Liquid water was a contributing cause in the development of life on Earth. Suppose that liquid water is absent on a candidate exoplanet, X. This could increase the plausibility of X being habitable if an alternative solvent like formamide is present and formamide is more compatible with the biochemistry of X.

(2) Water activity (the presence of free water available for cells to absorb) is a crucial determiner of microbial life on Earth, particularly in saline environments [Fox-Powell et al. 2016]. This has led to the assumption that water activity on Mars is similarly relevant to the poten-

tial habitability of Mars, especially because Mars has saline waters. However, astrobiologists have found that the different histories of Mars and Earth have resulted in different water chemistry such that water activity does not determine microbial growth in simulated Martian environments [Fox-Powell et al. 2016]. While water activity is a contributing cause for life on Earth, in Martian environments water activity is only a contributing cause if certain other chemical preconditions are met [Fox-Powell et al. 2016].

In summary, relevant dissimilarity between source and target domain can increase scientists' credence in an analogical inference's conclusion. In the example I gave, astrobiologists' background assumptions about what factors could give rise to life weigh more heavily than just satisfaction of *Source Domain Salience* and *Overlap* alone. I claimed that this is an epistemic virtue. Judging an analogical inference by *Source Domain Salience* and *Overlap* alone would amount to assuming that Earth's properties are the ideal properties for giving rise to life [Schulze-Makuch 2020].

I will briefly outline the morals that can be drawn from the past two sections. First, in the Central Cattle Pattern and Einstein elevator examples, we saw that *Source Domain Salience* and *Overlap* are not sufficient conditions for an analogical inference's plausibility. Now in the stimulus generalization and astrobiology examples, we see that *Source Domain Salience* and *Overlap* are not necessary conditions for an analogical inference's plausibility. An analogical inference can be plausible because it fails to satisfy *Source Domain Salience* and *Overlap*—relevant dissimilarity can increase plausibility.

A possible objection to my argument could be that the aforementioned normative accounts, particularly Bartha's, aim to establish criteria for *prima facie* plausibility. It could be claimed that in the examples I give, *Source Domain Salience* and *Overlap* are sufficient for *prima facie* plausibility, and that *Target Domain Salience* only comes into play when evaluating stronger definitions of plausibility.

I have three responses to this. The first point is pragmatic: the goal of a normative account is to give scientists tools for making and evaluating analogical inferences. In several of the examples discussed, *Target Domain Salience* is used by scientists to evaluate whether an analogical conjecture has basic epistemic support and to decide whether to devote research and funding to further investigation. A satisfactory account of analogical inference should provide an epistemic toolkit for scientists to use when making these judgments. If an account with criteria for *Target Domain Salience* offers more epistemic tools, and makes more of the implicit structure of analogical inferences explicit, then it is preferable to an account that lacks criteria for *Target Domain Salience*.

My second response is that several of the examples discussed do align with Bartha's definition of *prima facie* plausibility. For instance, in the astrobiology example, the Planetary Habitability Index is not a metric of which planets are thought to have a high probability of supporting life; it is a metric of which ones could potentially be habitable. Even for the ones one that ranked higher on the PHI index, astrobiologists only claim that they meet minimal conditions for potential habitability.

Finally, it seems apparent that the very distinction between *prima facie* plausibility and stronger plausibility depends upon local background knowledge. *Prima facie* plausibility depends on judgments of inductive risk, epistemic values, how many alternate hypotheses exist in the target domain, and the representativeness of the target domain's sample size. These considerations are all local to the target domain. Suppose that one astrobiologist holds the background assumption that the probability of life forming if a planet is habitable is close to 1. A second astrobiologist holds the background assumption that life has a very low chance of originating even if a planet is habitable, and thus believes that very few planets in our galaxy are inhabited.[5] These two astrobiologists will hold different standards for the *prima facie* plausibility of a given planet harboring life. The second astrobiologist will have

[5]Disagreement about the probability of a habitable planet harboring life is a source of controversy amongst astrobiologists, and is one of the main objections to the Drake equation

more stringent criteria for *prima facie* plausibility than the first astrobiologist will because they think that the presence of life is a matter of chance and not just habitability.

Inductive risk is also a factor that is local to the target domain and plays a role in evaluations of what constitutes *prima facie* plausibility. For a given analogy, if finding an analogical conjecture plausible when it is incorrect would yield serious consequences, but finding it implausible when it is correct would do little harm, then the criteria for establishing *prima facie* plausibility will be stringent. If the reverse is true, then it is rational for the criteria for *prima facie* plausibility to be lenient such that even an analogy that has a weak source domain relationship or violates *Overlap* by having a critical difference between the source and target domain might be considered *prima facie* plausible. There is a history of scientists fruitfully using analogies that do not meet the conditions of *Source Domain Salience* and *Overlap*, such as Mendeleev's manipulation of elements' atomic masses in the periodic table [Ross 2018].

2.5 Conclusion

I showed that under Hesse, Bartha, and Gentner's accounts, there are two essential characteristics that an analogical inference must have in order to be plausible, *Source Domain Salience* and *Overlap*. Current normative accounts see these as necessary and sufficient conditions for *prima facie* plausibility. Bartha critiques Gentner's account on the grounds that her systematicity principle is neither necessary nor sufficient for plausibility [Bartha 2009, pg. 70], and frames his formulation of *Source Domain Salience* and *Overlap* as a refinement that does determine *prima facie* plausibility.

Using examples from archaeology, physics, and astrobiology, I argued that Bartha, Gentner, and Hesse's formulations of *Source Domain Salience* and *Overlap* all fail to constitute nec-

essary and sufficient conditions for *prima facie* plausibility. The main claim of this paper was that in addition to *Source Domain Salience* and *Overlap*, there is a further essential characteristic than scientists use to evaluate the plausibility of analogical inferences. *Target Domain Salience* plays a crucial role in scientists' plausibility judgments but is not captured by Bartha, Gentner, or Hesse's accounts.

I argue that a successful normative account should have criteria for *Target Domain Salience*. However, I do not claim that adding *Target Domain Salience* to *Source Domain Salience* and *Overlap* would make them jointly necessary and sufficient for *prima facie* plausibility. Rather, I see these three essential characteristics as being more similar to epistemic values. Just like with epistemic values, scientists must debate about which to prioritize in a given context according to their goals and domain-specific knowledge. Current normative accounts constitute an incomplete set of epistemic values because they lack criteria for *Target Domain Salience*, yet *Target Domain Salience* is a criterion that is used fruitfully by scientists in evaluating analogical inferences. A successful account should include normative standards for *Target Domain Salience*.

Additionally, I showed that often what powers the plausibility of an analogical inference is scientists' context-dependent background assumptions. The fact that an analogical inference's plausibility depends on local background assumptions has already been established by John Norton [2021].

Under Norton's account, analogical inferences are built upon a domain-specific, material law, the 'fact of analogy', which unites the source and target domain. The fact of analogy is an inductive generalization that renders the actual analogical conjecture deductive. For instance, in the Pasteur example illustrated in Figure 2, the fact of analogy would be something like: "The mechanisms that cause microbe growth to decrease in the body and in the Petri dish are the same". Under Norton's account, all of the steps of an analogical inference are deductive except for the evaluation of the fact of analogy—this is the inductive step.

Norton claims that the evaluation of plausibility should not depend on domain-general normative standards. Instead, evaluation of an analogical inference's plausibility should consist of scientists empirically investigating the strength of the fact of analogy.

There are three challenges that make Norton's account unsuitable as a normative account of analogical inference. First, Norton's account is not normative in the sense that it does not provide any criteria for making and evaluating analogical inferences. Under Norton's account, to evaluate the strength of an analogical inference, scientists "engage in empirical investigations of the fact of analogy [Norton 2021, pg. 14]." This suggestion dismisses a key goal of normative accounts of analogical inference; normative accounts aim to provide an epistemic tool-kit for scientists to refine their empirical investigation and interpretation of evidence. Often, scientists' goal in evaluating an analogical inference is to decide whether to allocate funding and research time towards a conjectured hypothesis, or invest it elsewhere. There are also many cases in which further empirical investigation is not helpful or possible; such as in cases where the available evidence is permanently finite[6].

Second, Norton's 'fact of analogy' is too narrow to describe the full scope of background assumptions that scientists use in their analogical inferences. Norton claims that the fact of analogy is a material fact that can be examined empirically. Not all background assumptions fit this description. In the Einstein elevator example, the background assumption powering the analogy is not an empirical claim; it is a claim about epistemic values—that we should not differentiate theoretically between entities that we cannot differentiate empirically. What kind of empirical investigation could determine the strength of this background assumption? It falls outside the scope of Norton's 'fact of analogy' because it cannot be examined empirically. Norton's account also fails to handle cases where the fact of analogy is an abstract structural relation [Genta 2020, pg. 21]. Third, Bartha points out that there are some instances where the 'fact of analogy' is just the conclusion of the analogical inference, in

[6]Permanent underdetermination sometimes occurs in archaeology, such as in cases where a site was fully excavated in the past and inferences have to be drawn on the basis of archival records [Bartha 2009].

which case Norton's account is circular—the proposition that we would empirically examine to determine the strength of the analogical inference is just the conclusion of the analogical inference [Bartha 2020, Genta 2020].

What I have said so far may seem pessimistic about the feasibility of devising a successful normative account of analogical inference. However, contrary to Norton's claim, I suggest that it is possible to provide useful epistemic norms for analogical inference even though many of the criteria scientists use are local and context-dependent. I suggest that a successful model of analogical inference will uncover some general and domain-specific patterns that tend to characterize fruitful use of analogical inference in those domains. These should include criteria for how background assumptions are used (for instance one suggestion might be that there should be a limit on how many times we reuse a background assumption if it doesn't eventually yield independent supporting evidence), identification of innate cognitive biases (like those discussed in the stimulus generalization literature) that can influence our similarity and salience judgments, and guidelines for how to identify and constructively mediate cognitive biases and social values when they inevitably influence scientists' *Target Domain Salience* judgments. These would be epistemic tools rather than necessary and sufficient conditions.

That positive account is the topic of another chapter, and the more modest claim that this chapter aims to establish is that *Target Domain Salience* is an essential component of plausibility that is not captured by current normative accounts of analogical inference. Including *Target Domain Salience* as a criterion in normative accounts will have several important consequences.

First, it will resolve a mismatch between normative accounts and descriptive accounts of analogical reasoning in science. Looking at analogical reasoning in scientific practice, we see that scientists often break the question of plausibility down into three sub-questions: (1) "Is there a strong relationship in the source domain?", (2) "Do the critical features of

that relationship have analogues in the target domain?", and (3) "Are the features of the target domain salient to the possibility of the further feature being generalizable to the target domain?" Even when scientists debate about basic *prima facie* plausibility, there are examples where their debate hinges on disagreement about *Target Domain Salience*, not just *Source Domain Salience* and *Overlap*. This third criterion is not reducible to the first two; scientists use some criteria for evaluating *Target Domain Salience* that are not captured by *Source Domain Salience* or *Overlap*. Despite this, only the former two questions are included in current normative accounts, which creates a puzzle.

Second, including *Target Domain Salience* as a criterion for plausibility and working to provide epistemic norms for evaluating *Target Domain Salience* would be a first step towards bridging the gap between the two-dimensional accounts (which aim to give domain-general, necessary-and-sufficient conditions for plausibility), and Norton's material account which sees evaluation as a local, empirical affair. *Target Domain Salience* judgments are a key place where local background assumptions and domain-specific knowledge affect evaluations of analogical inferences. A successful account of analogical inference will aim to provide a collection of epistemic norms for scientists to have as resources for evaluating *Target Domain Salience*. These will not take the form of necessary and sufficient conditions, but rather will be fallible guidelines that characterize fruitful patterns and practices.

Finally, including criteria for *Target Domain Salience* in normative accounts will help make the influence of social values and background assumptions explicit and may provide an ideal locus for bringing considerations from social epistemology into current normative accounts of analogical inference.

In summary, a successful normative account of analogical inference should go beyond the scope of *Source Domain Salience* and *Overlap*; it should make the implicit background assumptions we employ in our analogical inferences more explicit, and it should give us both general and domain-specific tools for evaluating *Target Domain Salience*.

Chapter 3

Reverse Analogical Inferences: Gerrymandering the Source Domain Representation

3.1 Introduction

Analogical inferences play an important role in science; they are used to guide research into understudied areas, generate new hypotheses, justify existing theories, unify distinct theoretical entities, and differentiate distinct entities that have previously been treated as indistinct. Because analogical inferences play an important role in science, the analogical inference literature—which spans philosophy and cognitive science—has aimed to identify normative criteria for analogical inference. These normative accounts consist of an evaluation of the vertical (often causal) relations holding in the source domain, and of the horizontal similarity correspondences holding between the source and target domain. Under current influential accounts, an analogical inference's plausibility is determined by it having (1) a meaningful representation in the source domain and (2) relevant similarities between the source and target domain [Hesse 1966, Bartha 2009, Bartha 2015, Gentner 1983].

However, there is a puzzle. The degree of similarity between source and target domain is determined by *how* the source and target domain are represented, and in cases of underdetermination there are multiple possible ways for the source and target domain to be represented. Thus, there are two possible ways of interpreting dissimilarity when evaluating an analogical inference. First, dissimilarity between source and target domain could indicate that the analogical inference is implausible (i.e. that the further feature of the source domain cannot be generalized to the target domain). This first type of interpretation is well discussed in the analogical inference literature. *Alternatively*, dissimilarity could be taken as justification for altering the representation in the source or target domain in order to increase the degree of similarity between source and target domain. This latter interpretation of dissimilarity has not been examined in the analogical inference literature.

Using two examples from biology and chemistry, I demonstrate instances where scientists respond to a lack of similarity between source and target domain in this latter way—not by

questioning the plausibility of the analogical inference, but by altering the representation in the source domain. This gerrymandering of the source domain—which I call *reverse analogical inference*—can entail reevaluating what level of description is meaningful, what causal relations are important, or what the boundaries of different entities are. Such gerrymandering of the source domain can be fruitful in a number of ways: it can amend the causal relationships described in the source domain to make them more stable under changes in background conditions, identify higher-level relations that are more generalizable, etc.

These reverse analogical inferences are used by scientists across a variety of fields but have not been acknowledged or examined in the analogical inference literature. This chapter examines how scientists decide whether to interpret dissimilarity as evidence for changing their source and target domain representations. I show that the decision about whether to conclude that an analogical conjecture is implausible or make a reverse analogical inference is dependent on scientists' local epistemic commitments and cannot be captured by domain-general criteria. Finally, I propose a refinement of current influential accounts of analogical inference in order to take into account this unacknowledged type of analogical inference.

3.2 The Standard Picture

Under current *relevant similarity framework*[1] accounts, evaluating an analogical inference is a linear process consisting of two steps, (1) evaluating the strength of the vertical relationship in the source domain and identifying the critical features of that relationship *(Source Domain Salience)*, and (2) identifying the degree of similarity between those features and the features in the target domain *(Overlap)* [Bartha 2009, pg. 109; Lauman-Lairson 2025, pg. 6].

Under these accounts, steps (1) and (2) are epistemically prior to scientists' conclusion about the plausibility of the analogical conjecture. For Bartha, Hesse, and Gentner, the

[1]This category of accounts, relevant similarity framework accounts, is defined on pg. 2 and pg. 13

satisfaction of *Source Domain Salience* and *Overlap* is necessary and sufficient for the *prima facie* plausibility of an analogical conjecture. Justification does not flow in the opposite direction; scientists' intuitions about the plausibility of an analogical conjecture should not be a factor in their evaluation of whether *Source Domain Salience* and *Overlap* are satisfied [Bartha 2009].

If scientists conduct steps (1) and (2) and discover that a feature identified as critical in the source domain is absent in the target domain, then they should conclude that the analogical conjecture is implausible [Bartha 2009, pg. 40]. Consider the following example, which is consistent with the standard accounts.

Example 3.2.1. Anthropologists have aimed to establish a theory about the origins of horse domestication. The Botai archaeological site in Kazakhstan drew attention as a plausible location for the origin of domestication. Archaeologists found a significant quantity of horse remains and horse bone tools situated near the remains of the Copper Age village [Olsen 2006]. The site and horse remains bore similarities to other archaeological sites that were known to have housed domesticated horses. First, some soil at the Botai site had chemical properties that was compatible with horse manure. Second, remains of wooden posts suggested the possibility of paddocks. Pottery at the site contained possible remains of mare milk. Finally, some of the horse remains showed wear patterns on their teeth that were consistent with damage from a primitive bit [Olsen, 2006; Gaunitze 2018]. The Botai site was also chronologically and geographically situated in a way that made it a plausible location for early domestication. None of these features constituted strong evidence, but were thought to be potential markers of domestication. These features constituted the positive analogy.

A negative analogy was also present. One group of archaeologists who specialized in ethnographic studies of animal husbandry were not convinced that the horses at Botai had been domesticated [Taylor 2021]. They argued that in most communities that practiced pastoral-

ism, it was consistently the case that slaughtered animals were primarily young males and aging males and females [Taylor 2021]. This was an ideal practice for maintaining herd numbers and avoiding having too many breeding-age males. This slaughter pattern could be seen consistently in the archaeological record at other sites, invariant across different times and regions. At Botai, they conducted analyses of the horse remains and found that a majority of the horse remains were breeding age adults with an equal distribution of males and females [Taylor 2021]. Slaughtering breeding age adults of both sexes would be a surprising practice if the animals were a domesticated breeding herd.

There was also a more critical difference. When genetic testing had previously been conducted at known sites of horse domestication, the remains had always been genetically related to contemporary domesticated horses [Orlando et al. 2013]. Anthropologists tested the remains at Botai to see if this similarity held. However, genetic testing established that the horse remains at Botai are not genetically related to modern equids, and instead are ancestors of wild Przewalski horses, a species of equid that diverged from the last common ancestor of modern domesticated horses at least 30,000 years prior to the first instance of equine domestication [Orlando et al. 2013]. Przewalski horses have more chromosomes than domesticated horses and vary in conformation and temperament [Gaunitze 2018]. For decades there has been scientific consensus that Przewalski horses "represent the last remaining true wild horses [Sarkissian 2015]" whose ancestors were never domesticated.

If the horses at Botai were domesticated, this would overturn two firmly entrenched background assumptions: (1) that Przewalski horses were truly wild, not feral horses, and (2) that horse domestication originated in one location and spread elsewhere rather than originating independently in two different places. As a result, an influential 2021 paper concluded that the archaeological record at Botai was not evidence of domestication, and was simply evidence of wild horses being hunted [Taylor 2021].

The analogical inference is illustrated in Table 3.1.

Known sites of domestication	**Botai site**
Horse remains with bit wear	→ Horse remains with apparent bit wear
Manure heaps	→ Apparent manure heaps
Remains of corral posts	→ Apparent remains of corral posts
Genetic ancestors of modern domesticated horses	**X** Genetic ancestors of Przewalski's horses
Low proportion of breeding-age female remains	**X** High proportion of breeding-age female remains
Domesticated horses	**X Not domesticated horses**

Table 3.1: The Botai analogical inference

Here, scientists' evaluation of the analogical inference is consistent with current normative accounts of analogical inference. Current *relevant similarity framework* accounts say that when evaluating an analogical inference, scientists should first evaluate the strength of the vertical relationship represented in the source domain and decide which features of the source domain are critical to that relationship. Then, they should check whether those critical features hold in the target domain [Bartha 2009]. If there is a critical difference between the source and target domain representations, then scientists should conclude that the analogical conjecture is implausible. Under current accounts, the conditions *Source Domain Salience* and *Overlap* are epistemically prior to the assessment of an analogical conjecture's plausibility; they provide the conditions under which a conjecture may be entertained as a hypothesis.

At the time, there was wide agreement in archaeology that critical markers of a site of horse domestication include (1) having a low proportion of breeding-age females in horse remains and (2) the remains being genetically related to contemporary domesticated equines. These were considered critical features of the source domain. When these features were found to be absent at Botai, this constituted a violation of the *Overlap* condition. As prescribed by the relevant similarity framework, an influential [2021] paper concluded that the analogical conjecture was implausible—Botai must not be a site of early domestication [Taylor 2021].

3.3 Underdetermination of representations

However, there is a puzzle for current normative accounts. It is a common theme in philosophy that there exists a gap between theory and evidence that is filled by values and background assumptions [Kuhn 1962, Douglas 2014]. In the evaluation of analogical inferences, a gap exists between the available evidence and scientists' choice of representations in the source and target domain. The degree of similarity between source and target domain is determined by how the source and target domain are represented, and in cases of underdetermination, there are multiple possible representations. Which representation scientists choose determines the degree of similarity obtained.

In summary, under current normative accounts of analogical inference, the plausibility of an analogical inference is determined by the degree of similarity between the source and target domain. This poses a challenge for current accounts because the degree of similarity between the source and target domain is subject to underdetermination. The degree of similarity depends on the representational choices made in the source and target domain, and there is often a plurality of conceivably valid representations. Different representations that are equally coherent with the available data will yield different degrees of source-target overlap. Thus, there is no unique basis for assessing similarity.

Example 3.3.1. The underdetermination of similarity can be seen in the investigation of the Botai site. When making an analogical inference between known sites of domestication and Botai, there are multiple possible representations of the source and target domains that one could choose that would be consistent with the empirical data. Thus, scientists who had access to the same empirical evidence had different interpretations of the degree of similarity between Botai and known sites of horse domestication. Some representations of the source and target domain are consistent with a domestication hypothesis while others are consistent with a non-domestication hypothesis.

One example is in scientists' interpretation of the finding that the Botai horse remains contain a high proportion of breeding-age females. Some proponents of the theory that Botai was an early site of horse domestication argued that this disanalogy between Botai and known sites of horse domestication was not significant [Fages 2020]. They proposed several explanations to establish that the pattern of slaughtering mainly young males and post-breeding age animals is not an essential feature of domesticated horse herds. For instance, one group of archaeologists proposed that breeding age horses could have been ritually slaughtered at Botai, and that this would explain the disanalogy between Botai and sites with known pastoral horse management [Taylor et al. 2021]. Another attempt to minimize the significance of the negative analogy was to point out the difficulty of determining sex in the horse remains. A final argument was that the pattern of finding mainly young males and elderly horses in the horse remains was not a reliable indictor of domestication anyways. These anthropologists argued that even in societies that did not practice domestication and only hunted horses, there would have also been a high frequency of young males and older horses in the horse remains. This is due to juvenile males forming small bachelor bands that are an easy target for hunters, and older horses being easier to hunt [Olsen 2006].

These arguments all support an interpretation of the empirical data at Botai in which the dissimilarities between Botai and known sites of horse domestication are not *critical* differences. Some archaeologists took the high proportion of breeding-age female horse remains at Botai to be a critical disanalogy, while others minimized the disanology by positing the auxiliary hypotheses described above.

The underdetermination of similarity raises concerns about the efficacy and scope of current normative accounts of analogical inference because it undermines their ability to provide stable criteria for plausibility given that degree of similarity is flexible and indeterminate. However, it does not in itself constitute an inconsistency with their framework. Under current accounts, scientists' representational choices are treated as independent from and external to

the normative evaluation of analogical inference. Current accounts assume that the question of how to represent the source and target domains has already been resolved prior to the analogical inference being proposed.

At multiple points in his [2009] book, Bartha emphasizes that analogical inferences proceed from *established* representations and norms for similarity that are already fixed prior to the analogical inference being proposed. For instance, Bartha describes the source and target domains as "a set of objects, properties, relations, and functions, together with a set of accepted statements about those objects, properties, relations, and functions [Bartha 2009, pg. 13]." At another point, Bartha says, "An analogical argument is an explicit representation of analogical reasoning that cites accepted similarities between two systems in support of the conclusion that some further similarity exists [Bartha 2009, pg. 1]." Under this picture, the underdetermination of representations problem is outside the proper scope of a normative account of analogical inference. Representational choice is bracketed as background, not included as a subject of normative accounts of analogical inference.

It is entirely reasonable that normative accounts of analogical inference seek to limit their scope in this way. Their goal is to offer normative standards for analogical inference, not to offer normative standards for all aspects of scientific reasoning. From this perspective, the underdetermination of representations falls outside the domain of normative accounts of analogical inference. Then, the fact that normative accounts do not provide tools for solving the underdetermination problem does not directly threaten their normative adequacy, even if it does limit the meaningfulness of their predictions.

However, the science literature suggests that this boundary cannot be cleanly maintained. There are numerous cases in the scientific literature in which scientists' representational choices are made within the context of analogical reasoning itself. In such cases, the selection of a representation is not epistemically prior to the analogical inference but is shaped by it—scientists choose representations in part to facilitate overlap between a source and target

domain. Consider an example of this in the Botai enquiry.

When genetic testing of the remains at Botai revealed that the horses at Botai were ancestors of Przewalski horses, some scientists were willing to give up the firmly entrenched theory that Przewalski horses are a species whose ancestors were never domesticated. This theory had had scientific consensus for decades. However, faced with the evidence that the horses at Botai were ancestors of Przewalski horses, some archaeologists decided that they were more committed to the belief that the Botai horses were domesticated than to the widely-accepted theory that Przewalski horses were wild [Gaunitz 2018]. Currently, one faction of archaeologists claim that Botai was not a site of domestication and that Przewalski horses are truly wild [Taylor 2021] while another faction claim that Botai was a site of domestication and therefore Przewalski horses must be feral, not wild [Gaunitz 2018]. The available evidence does not determine which of these representations is most plausible.

The most significant aspect of this example is that archaeologists' epistemic commitment to a given representation of the source domain changed in the context of the analogical inference. Under current normative accounts of analogical inference, scientists should first evaluate *Source Domain Salience*, the strength of the relationship in the source domain and the critical features of that relationship. Next, scientists should evaluate *Overlap*, the degree to which those critical features hold in the target domain.

Instead, archaeologists' beliefs about how to represent the source domain appeared to be affected by the consideration of which representation would yield a high degree of source-target similarity and make the analogical inference satisfy *Overlap*. Their epistemic commitments prior to the analogical inference—which included a belief that Przewalski horses are a wild species whose ancestors were never domesticated—shifted in the context of the analogical inference, such that they decreased their credence in a representation that they have previously been committed to. For instance, some archaeologists initially claimed that genetic relatedness to contemporary domesticated horses carried evidential weight for the conjec-

ture that the Botai horses were domesticated, but later reduced their assessment of the evidential weight of this feature when it was found that Botai horses are not ancestors of contemporary domesticated horses [Olsen 2006]. It seems that the archaeologists developed intuitions about the plausibility of the analogical conjecture, and used this as a criterion in their representation choice.

The psychology literature also supports the claim that representation choice is not epistemically independent from analogical inference. For instance, Medin et al. [2002] argue that representational framing and analogical reasoning are interdependent processes, where the representation of the source and target is influenced by the goals and context of the analogy [Medin et al. 2002]. Similarly, Gentner et al. find that agents adjust their representations of the source and target during the process of analogical comparison in order to maximize structural correspondence between the source and target [Gentner et al. 1993]. I will discuss these psychology findings more in Sect 3.5.

The finding that analogical reasoning affects representation choice is inconsistent with the framework posed by current normative accounts of analogical inference and creates a puzzle. The fact that scientists gauge their degree of commitment to a representation within the context of analogical inferences, influenced by considerations of making the analogical inference go through, means that the underdetermination of representations poses a challenge that legitimately falls within the proper scope of normative accounts of analogical inference. It is the job of a normative account of analogical inference to offer some answer to the underdetermination of representations puzzle because scientists weigh their commitment to different representations within the context of analogical reasoning.

The relevant similarity framework accounts say that the plausibility of an analogical inference is determined by the degree of similarity between the source and target domain. I showed that the degree of similarity between source and target domain is determined by scientists' choice of representation of the source and target domain. Further, I showed that scientists'

choice of representation is shaped by considerations of what effect a representation choice will have on the degree of similarity between the source and target domain. This is a process that could be viciously or virtuously circular. In order to be normatively adequate, an account of analogical inference must provide guidelines for how scientists should choose between multiple underdetermined representations of the source and target domains.

In summary, current normative accounts of analogical face a challenge that arises from the underdetermination of representations. Under current normative accounts, an analogical inference's plausibility is determined by the degree of similarity between the source and target domain. However, the degree of similarity between source and target domain is dependent on how scientists represent the source and target domain, and there are often multiple possible representations that are consistent with the available empirical evidence. Which representation scientists choose determines the degree of similarity between source and target domain, and thus the plausibility of the analogical inference. In these cases of underdetermination, the degree of similarity between source and target domain is not discovered, but rather is created by the representational choices made. Thus, the plausibility of an analogical inference under current accounts is not dependent on intrinsic properties of the source and target domain, but on subjective framing choices made by scientists.

This worry about the underdetermination of representations is not an entirely new concern for philosophers of analogical inference. However, it has only been thought to be a problem for a subset of syntactic accounts of analogy. Hofstadter [1995], Abrantes [1999], and Bartha [2009] argue that Gentner's structure-mapping account suffers from an underdetermination problem. Bartha claims that under Gentner's account, the plausibility of the analogical inference is "pre-ordained [Bartha 2009, pg. 67]" by the representational choices made. This is because under Gentner's account, only the abstract, syntactic contents of a representation are mappable. Semantic content is not a consideration when deciding whether a source maps to a target. For example, in the Rutherford model of the atom, the mapping between

a source and target domain consists only of syntactic relational structure between nodes, and does not include any semantic content like the conceptual meanings of 'sun', 'electron' or 'revolves around' [Gentner 1983, pg. 1]. This means that degree of similarity depends on the syntactical representation choices made, rather than the semantic content of the domains. A source and target domain that are semantically dissimilar from one another can be represented as syntactically similar. Likewise, a source and target domain that are meaningfully similar can be syntactically represented in a way that causes them to fail to have a horizontal mapping. In the Rutherford example, if one represents the source as 'Sun GRAVITATIONALLY ATTRACTS planets' and the target as 'Nucleus ELECTRICALLY ATTRACTS electrons' then one will get an isomorphism between the source and target, and the analogy will be plausible under Gentner's account. If, however, one adds the property of valence to the target domain to represent the fact that electrical force can have a positive or negative charge, then you get a three-place relationship in the target domain such that there is no longer a mapping between the source and target and the analogy fails [Bartha 2009, pg. 69].

For any source and target, there are multiple possible syntactic representations, and in order to choose representations that will yield relevant similarities between the source and target domain, one has to already know what the relevant mapping between the source and target domain is. If one merely chooses a plausible representation of the source domain and then independently chooses a plausible representation of the target domain, then the result may be a failure to obtain any isomorphism between the source and target, or the result may be to obtain one of multiple, incompatible isomorphisms that support different analogical conjectures [Abrantes 1999].

Hofstader also points out that Gentner's account relies on a distinction between objects and relations, where the former are represented as nodes and the latter as vertical mappings. Yet there is no matter of fact about whether a real-world feature is an object or relation.

For instance, Gentner treats 'heat' as an object in the fluid flow model of heat transfer, but heat is treated as a property or relation in other analogies [Hofstadter 1994, pg. 94]. If heat were treated as a relation while fluid were treated as a node, then there would be no mapping between the source and target. To get the 'right' mapping, one must know in advance whether the fruitful mapping treats a given feature as a node or a relation.

In summary, a source and target domain may seem relevantly similar only because they have been represented in a particular way. Furthermore, the very act of representing the source and target already involves a kind of analogical creativity in which the two domains are brought into alignment. Thus, part of the useful epistemic work that analogical inferences do is done at the stage of choosing a representation of the source and target. Despite this, epistemic norms for representation choice are not included in current normative accounts.

Bartha and Nappo raise a similar concern about underdetermination for formal and mathematical analogies. Mathematical conjectures are often motivated by inductive reasoning from less familiar to more familiar mathematical domains [Nappo 2022]. In mathematical analogies, mathematicians find structural similarities or partial isomorphisms between the mathematical forms of two domains, and conclude that a conjecture that holds in the source domain also holds in the target domain. The puzzle is that there are multiple distinct mathematical formalisms that can represent a given physical phenomenon or yield a given theorem. For example, wave mechanics and matrix mechanics are two formalisms developed to describe quantum systems that vary substantially in their structural form yet make the same empirical predictions. The matrix mechanics formulation uses an algebraic method whereas wave mechanics uses differential equations [Perovic 2008]. Schrödinger [1926] and then von Neumann [1932] proved that these two formulations are predictively equivalent despite having different mathematical form.

There are many possible forms that a mathematical representation of a source domain can take, and which of these is chosen determines the degree of similarity or isomorphism that

is obtained in an analogy. Fruitful mathematical analogies often depend on mathematicians coming up with nonstandard or creative representations of the source or target domain in order to generate the similarities or partial isomorphisms that the mathematical analogy hinges on [Pólya 1954; Bartha 2009, pg. 169]. Consider the following example from Bartha [2009] and Nappo [2022], originally from Pólya [1954].

Example 3.3.2. Leonard Euler found that all polyhedra are governed by a shared law that dictates the relationship between the number of vertices, edges, and faces [Euler 1758]. Euler's formula for polyhedra is V + F - E = 2. Pólya [1954] rewrote Euler's formula to uncover a structural similarity between the formula for polytopes of different dimensions. Euler's formula for two-dimensional polytopes is simply V = E. The three-dimensional case is V + F - E = 2. Pólya rewrote these two formulas to create a structural similarity between them, where they both are represented as an alternating sum proceeding from number of elements of lowest dimension (vertices) to elements of highest dimension (faces for 2-D and solids for 3-D)[Bartha 2009, Pólya 1954]. For two-dimensional polytopes, there is one face, so 1 - F = 0. Substituting 1 - F for 0, Pólya obtained the formula V - E + F = 1. For three-dimensional polyhedra, V + F - E = 2 is numerically equivalent to V - E + F - C = 1 because the number of cells (number of three dimensional objects) equals 1.

By adding the variable F (number of two-dimensional objects) to the 2-D case and the variable C (number of three-dimensional objects) to the 3-D case and rearranging the terms, Pólya creates two new representations that exhibit a structural similarity that did not exist between the mathematical forms of the original 2-D and 3-D formulations of Euler's formula. From the pattern exhibited in the 2-D and 3-D case, one can make an analogy to predict a law for 4-D cases and larger dimensions. For a 4-D polytope, the largest dimension is polychora. The analogy from the 2-D and 3-D cases predicts that the formula for the 4-D case is V - E + F - C + P = 1, where P is the number of polychora.

The analogy makes a fruitful prediction; it successfully predicts the relationship between

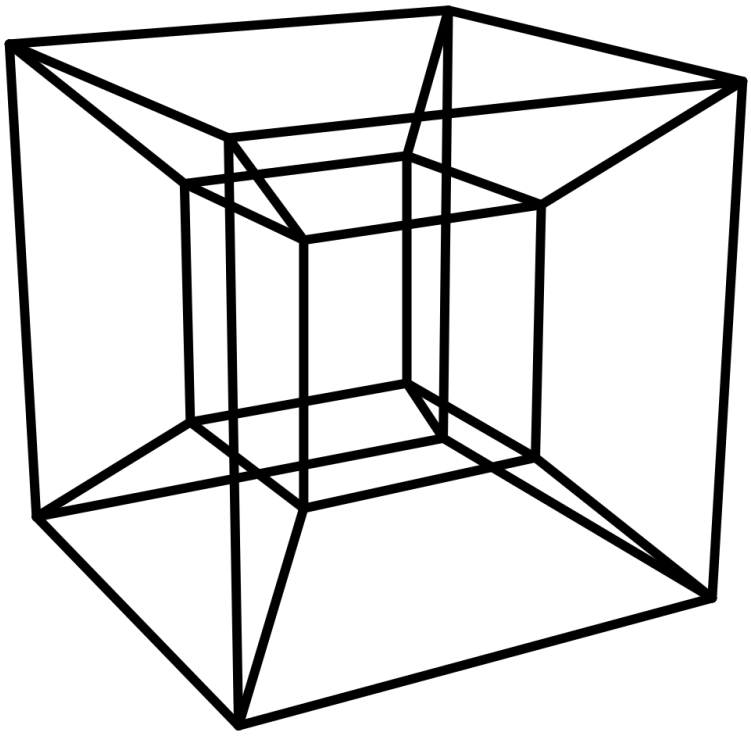


Figure 3.1: Tesseract (16 vertices, 32 edges, 24 faces, 8 cells, and 1 polychora).

vertices, edges, faces, cells, and polychora for 4-D objects. For example, a tesseract has 16 vertices, 32 edges, 24 faces, 8 cells, and 1 polychora. If we input these values into our conjectured formula for 4-D cases, V - E + F - C + P = 1, we find that the formula does hold. Manipulating the representation of Euler's formula to force a structural isomorphism between 2-D and 3-D cases yielded a fruitful conjecture about higher-dimension cases.

Nonstandard representations are a fruitful and essential component of mathematical analogy [Pólya 1954]. Mathematicians regularly form useful conjectures by manipulating mathematical forms to create isomorphisms with other mathematical domains. Of course, nonstandard representations can also generate structural similarities between a source and target domain that are not meaningful or informative. Consider the following example that Bartha [2009] gives.

Example 3.3.3. Among rectangles, the rectangle that maximizes area is the square. A plausible analogical inference results from using this as a source domain to conclude that the

three-dimensional box that maximizes volume is the cube. The successful analogy results from assuming that area is relevantly similar to volume, and it seems intuitively obvious that if a square maximizes area then a cube must maximize volume [Bartha 2009]. How should we express the analogical inference that provides inductive support for this conjecture?

The problem that arises is that it is easy to create an mathematical analogy that satisfies *Source Domain Salience* and *Overlap*, yet does not offer any inductive support for its conjecture. The formulas for area and volume could be written such that the two formulas are structurally analogous and satisfy *Source Domain Salience* and *Overlap*, but the similarity is a superficial one [Bartha 2009].

Area could be rewritten as $3x^{2-2} + x^{2-1}y^{2-1} - 3sin^{2-2}y$, which equals $x * y$.

Similarly, volume could be rewritten as $3x^{3-2} + x^{3-1}y^{3-1} - 3sin^{3-2}y$, which equals $x * y * z$.

In this case, there is a structural similarity in the formulas for area and volume; both are instances of the formula $3x^{n-2} + x^{n-1}y^{n-1} - 3sin^{n-2}y$ [Bartha 2009, pg. 155]. Does this similarity justify the conjecture that the box that maximizes volume is the cube? It seems clear that this is a manipulated similarity, in which superfluous terms have been added to the formulas for area and volume in order to make them exhibit a structural similarity [Bartha 2009]. Bartha calls this a case of *specious resemblance*, in which "two expressions with no meaningful similarity can be rewritten so as to satisfy the purely syntactic criterion (of similarity)[Bartha 2009, pg. 169]."

How should we distinguish between nonstandard representations that are fruitful, like that in 3.2, from those that are contrived, like example 3.3?

3.4 Constraints for appropriate representation

The existing analogical inference literature only recognizes that underdetermination of representations is a problem for a subset of syntactic accounts: Gentner's structure-mapping account and mathematical analogies under Bartha's account [Bartha 2009, pg. 169].

No attempts have been made to refine Gentner's account to bolster it against the underdetermination problem. However, Bartha [2009] and Nappo [2022] have proposed adding criteria to Bartha's [2009] account in order to resolve the underdetermination problem for mathematical analogies.

Bartha's solution is to propose an additional criterion, *Internal Coherence*, that must be met in order for *Source Domain Salience* to be satisfied for geometric analogies.[2]

Definition 3.1. Internal Coherence: For a geometric similarity to count as admissible, the relevant relations and functions should be expressed using standard representation, unless some justification internal to the domain can be given for a nonstandard representation. More specifically, any novel representation should be justified in terms of the proof that is the prior association for the analogical argument [Bartha 2009, pg. 169].

Bartha aims to distinguish between legitimate and contrived manipulations by appealing to the motivation for the manipulation. He argues that manipulating the source or target domain representation in order to influence the degree of similarity and make the analogical conjecture plausible is an unjustifiable practice, whereas creative representations can be admissible if the reasons for the choice of representation are internal to the source domain. In short, a manipulated representation is admissible if the motivation for changing the source

[2]Bartha makes a distinction between analogies based on geometric versus algebraic similarities and claims that underdetermination of representations is a only a problem for geometric analogies, which are based on more abstract structural similarities than algebraic analogies [Bartha 2009, pg. 162]. Nappo argues that algebraic representations are similarity underdetermined and require additional criteria as well [Nappo 2022, pg. 15].

domain is "independent of the analogy and its purposes [Bartha 2009]". However, the manipulation is unjustifiable if the motivation for the manipulation is "I want the [representation] to look like that (so as to provide analogical support for my conjecture) [Bartha 2009]."

Nappo worries that Bartha's *Internal Coherence* criterion is too weak, and describes several counterexamples in which mathematical analogies satisfy the *Internal Coherence* condition but are analogies that we should count as specious and uninformative [Nappo 2022]. Nappo proposes a stronger criterion, *Materiality*, which is borrowed from Hesse's [1963] account.

Definition 3.2. Materiality: The similarities mentioned in the analogical argument must be 'material' or 'pre-theoretic', i.e., they must be cases of sharing of features whose significance in a given context can be recognized before and independently of the analogical argument [Nappo 2022, pg. 10].

Materiality requires that the horizontal relations cited in the analogical inference hold epistemically prior to and independently of the causal, vertical relations that the analogical inference illuminates [Nappo 2022, pg. 13; Hesse 1963, pg. 63]. The similarities must be recognized prior to the analogical inference being made. An analogy that would not satisfy *Materiality* is the structural analogy 'Father is to son as state is to citizen' [Nappo 2022, pg. 13; Hesse 1963] because "there does not seem to be any horizontal relation of similarity between the terms, except in virtue of the fact that the two pairs are related by the same vertical relation [Hesse 1963, pg. 63]."

Nappo claims that *Materiality* resolves cases like the specious analogy described in example 3.3. In 3.3, the similarity between the source and target is that they are both instances of the formula $3x^{n-2} + x^{n-1}y^{n-1} - 3sin^{n-2}y$. Under *Materiality*, this resemblance is illegitimate because the similarity is not evident to mathematicians prior to the analogical inference being proposed. Nappo also claims that Pólya's analogical inference in Example 3.2 fails to satisfy *Materiality*, and thus does not provide inductive support for the conjecture and

should merely be seen as a heuristic tool [Nappo 2022]. Under Nappo's account, analogical inferences that satisfy *Internal Coherence* but not *Materiality* are suitable as heuristic tools, but are not strong enough to provide inductive support [Nappo 2022].

Bartha worries that Nappo's criterion is too stringent and defeats the goal of a normative account of analogical inference, which is to give scientists and mathematicians a tool for evaluating the plausibility of an analogical inference without going too far down the path of confirmation.

Another challenge for Nappo's account is that his definition of *Materiality* for mathematical analogies is vague. Hesse's definition of *Materiality* for material analogies is also vague, and it describes a psychological state of affairs—an perceptual similarity that is salient prior to having any scientific theory about the phenomena. For instance, Hesse claims that the similarity of water waves and sound waves satisfies *Materiality* because both have pre-theoretic, observable reflection and refraction effects. What does it mean for one mathematical proof to be pre-theoretically similar to another mathematical proof? Some of Nappo's discussion suggests that *Materiality* is a temporal notion; a similarity satisfies *Materiality* if it is recognized prior to and independently of the analogical inference being proposed. But what about cases in which manipulating a domain to manufacture a similarity leads to a notion of similarity that mathematicians find more intuitive and salient than the temporally prior notion?

In contrast to Nappo's concern that Bartha's *Internal Coherence* condition is too weak and needs to be supplemented by a stronger condition like *Materiality*, I have the opposite concern about *Internal Coherence*. I worry that philosophers could interpret Bartha's condition too strongly, and in doing so rule out a class of fruitful analogical inferences.

The question I have raised in this paper is, 'Given that different representations yield different degrees of *Overlap*, and there are often multiple possible representations that satisfy *Source*

Domain Salience, how should scientists and mathematicians choose which representation to use in the analogical inference?' The only answers to this question in the analogical inference literature so far come from Nappo [2022] and Bartha [2009]. Bartha's *Internal Coherence* condition says that the justification for a representation choice must be domain-internal, and cannot be of the form "I chose representation A over representation B because choosing A makes the analogy satisfy *Overlap*". Nappo's *Materiality* says that in order for a similarity between a source and target domain to provide inductive support, the similarity must be observable and salient independently of and prior to the analogical inference being constructed.

In broad terms, Bartha and Nappo both claim that the underdetermination of representations problem should be resolved by disallowing representation choices that were motivated by the goal of increasing the degree of similarity between source and target and making the analogy go through.

I will argue that this is the wrong way to go about constraining representations, and that scientists should consider the implication a representation will have for *Overlap* and use this as a criterion when deciding how to represent the source and target domain.

3.5 Reverse analogical inferences

The analogical inference literature makes a distinction between two types of analogical inference, material and formal. Material analogical inferences rely on physical, causal similarities between source and target domains, while formal analogical inferences rely on similarities in abstract structure or mathematical isomorphisms. Current accounts only recognize *specious resemblance* as being a threat for the subset of formal analogical inferences described in the last section: syntactic analogies under Gentner's structure-mapping account and algebraic

(but not geometric) analogies under Bartha's account. Representational gerrymandering is seen as being a fringe problem for this subset of formal analogies. Current accounts assume that material analogical inferences are not susceptible to gerrymandering [Bartha 2009; Hesse 1969]. Under Bartha's account, this is because material analogies rely on accepted, established source domain representations and entrenched scientific standards for relevant similarity. Bartha says,

> We seldom see analogical arguments [in science] that depend upon similarities that must themselves be established and justified analogically...Analogical arguments in science depend upon basic similarities that are supposed to be unproblematic. The relevant concepts have a clear meaning over a range of cases including those under discussion; they are not open-textured [Bartha 2009].

Under Hesse's account, material analogies are not susceptible to manipulated similarity because they rely on observable, pre-theoretic similarities between domains. Pre-theoretic similarities are features that "appear to be alike to fairly superficial observation [Hesse 1969, pg. 11]." This guards against some types of manipulated, specious similarity, like the predicate 'grue' in Goodman's puzzle of induction. Hesse would say that 'green' is a suitable predicate to use for induction because it is a physical feature that is pre-theoretically observable, whereas 'grue' is abstract and structural.

In this section, I will challenge the assumption that only abstract, formal analogical inferences are susceptible to gerrymandering of the source and target domain to create similarity. Material analogical inferences—analogical inferences that are based on physical, rather than abstract structural similarities—are also subject to manipulated similarity. I will give two examples in which scientists gerrymander their representations of the source and target domain in order to create a degree of similarity or dissimilarity between the source and target that supports their intuitions about the plausibility of the analogical conjecture.

Under current accounts, the plausibility of an analogical conjecture is an effect and not a cause of scientists' choices about how to represent the source and target domain. Bartha says:

> When we reason by analogy that some feature of the source domain is present in the target domain, we must *first* determine which features of the source domain are relevant and how they relate to the analogical conclusion [in the source domain]. Then we investigate whether the crucial contributing features (or something similar) are present in the target domain. That is what explains the plausibility (or implausibility) of the argument [Bartha 2009, pg. 72].

This section will introduce two analogical inferences in biology and chemistry that violate the assumption that evaluations of *Source Domain Salience* and *Overlap* are epistemically prior to evaluations of an analogical conjecture's plausibility.

These examples also violate another assumption of current accounts. Under current accounts, if investigation of *Source Domain Salience* and *Overlap* reveals critical dissimilarities between the source and target domain, then scientists should conclude that the analogical conjecture is implausible (i.e. that the further feature of the source domain cannot be generalized to the target domain). In the examples provided in this section, scientists instead use critical dissimilarities as justification for altering the representations in the source and/or target domain in order to increase the degree of similarity between the source and target. I will call this *reverse analogical inference*, where scientists manipulate the source and/or target domain representations in order to increase the degree of similarity and make the analogy go through. Reverse analogical inferences are not recognized in the existing literature and violate current normative accounts.

The term 'reverse analogical inference' does not describe a new type of analogical inference, but rather describes the finding that when scientists make an analogical inference, there is

never a matter of fact about whether they should use *Source Domain Salience* and *Overlap* to evaluate the plausibility of the analogical inference, or use their intuitions about the plausibility of the analogical inference to evaluate their representations of the source and target domain and their standards for *Source Domain Salience* and *Overlap*. When critical differences are found between the source and target domain, scientists can either conclude that the analogical conjecture is implausible, or they can 'flip' the direction of the analogy and use the critical dissimilarity to justify revising their representations. This is what I call a *reverse analogical inference*.

The aim of the remainder of this section will be to show that (1) scientists sometimes interpret critical dissimilarities between the source and target as being justification for a reverse analogical inference rather than being justification of the implausibility of the analogical conjecture, and (2) making a reverse analogical inference rather than concluding that the analogical conjecture is implausible can sometimes be fruitful and of epistemic benefit to scientists.

Example 3.5.1. One example where reverse analogical inference occurred is in early microbiologists' development of the concept 'virus'. In the 1700's, the term 'virus', which came from the Latin word for poison, was a vague, broad term used to describe any entity that causes sickness. Prior to germ theory being developed in the nineteenth century, there was no concept of disease being caused by active agents. Instead, diseases were thought to be propagated by miasma, an imagined poisonous fume emanating from sewage, sick individuals, or decaying matter [D'Abramo 2020]. After Pasteur's work in developing germ theory in the late 1800s, scientists shifted from miasma theory to the belief that disease is caused by living microorganisms. From the mid-1800s to mid 1900s, all infectious diseases were thought to be caused by living microbes like bacteria or fungi. In a paper on rabies, Pasteur said,

> Anthrax in cattle (the malignant pustule in humans) is produced by a microbe,

> croup is produced by a microbe. The microbe of rabies have not yet been isolated but judging by the analogy, their existence must be admitted. In summary, all viruses are microbes. The conditions which govern their existence and their propagation are subject to the same general laws which regulate the birth and multiplication of animals and higher plants [Pasteur 1890 (translated from French via Google translate)].

Anthrax	Croup	Rabies
Infectious disease ***Q:*** Bacterium	Infectious disease ***Q:*** Bacterium	Infectious disease ***?***

Table 3.2: Analogical argument for 'all viruses are microbes'

Additionally, all disease-causing agents were thought to be governed by Koch's postulates [D'Abramo 2020]. Robert Koch was a microbiologist who expanded on Pasteur's work and developed a framework for establishing whether a disease was caused by a particular microbe. These postulates were highly fruitful in helping establish causal connections between diseases and their causative pathogens. However, the research from which Koch's postulates were derived was based only on bacteria, because viruses were too small to be visible on the microscopes of the time and thus were epistemically accessible. Thus, the postulates were based on characteristics of bacteria.

For example, one of Koch's postulates says that in order for there to be an established causal relationship between a disease and some pathogen, it must be the case that the pathogen can be separated from the host and propagated in a culture in a lab. This postulate was highly fruitful, and scientists assumed that all pathogens were alive and capable of self-propagation independent of a host. As a result, rabies, smallpox, tuberculosis, cholera, and typhoid were all assumed to be caused by small living organisms capable of self-propagation independently of a host.

During this period, the term 'virus' had fallen out of favor among scientists because of

its association with miasma theory; it was seen as an imprecise natural-language synonym for 'microbe'. The word 'virus' was associated with the idea that disease is caused by an inactive miasma or poison, and was not seen as a suitable explanatory construct now that it was accepted that disease was caused by active, living agents.

An anomaly to the dominant microbe concept arose in the late 1800s. A blight appeared that affected tobacco crops in Europe. Tobacco was an important part of the economy, so scientists were enlisted to find the cause of the disease, which was named 'tobacco mosaic' because it caused a mottled effect on the leaves and stems of tobacco plants caused by uneven chlorophyll production. It was believed by scientists that this was a case of a bacterium infecting the plants, particularly after biologist Adolph Mayer discovered that puréeing diseased plants and introducing the purée to healthy plants would cause the malady to spread to the healthy plants [Mayer et al. 1886/2018]. Mayer followed the guidelines of Koch's postulates, and tried to culture the microorganism in a Petri dish. However, he was not successful at producing more of the infectious substance through culturing, which suggested that the microorganism was not propagating in the Petri dish [Mayer et al. 1886/2018]. Another of Koch's postulates stated that microbes should be able to be removed from a substance by straining the substance through a fine-grained filter. Mayor reported that he was able to remove the virus from the juice by filtering it through a bacterium resistant filter [Mayer et al. 1886/2018]. However, in hindsight, it is unclear how he got this result, given that the filters that were used at the time were not small enough to capture the virus now known to be responsible for tobacco mosaic. Mayer concluded that the agent responsible for tobacco mosaic "is bacterial, but the infectious forms have not yet been isolated [Mayer et al. 1886/2018]." However, subsequent experiments were not able to replicate Mayer's finding of filterability.

Further experiments by other scientists found other discrepancies between tobacco mosaic and the concept of microbe dictated by Koch's postulates. For instance, it was also found

that tobacco mosaic was not visible under the microscopes of the time, not affected by temperature or humidity, and was not filterable.

Table 3.3 shows a tabular representation of the analogical inference:

Microbe (late 19th century)	Tobacco Mosaic
Cause contagious diseases	Appears to cause contagious disease
Filterable through paper or porcelain filter	Not filterable
Propagation is responsive to temperature and humidity	Unresponsive to temperature and humidity
Self-propagates in culture	Does not propagate in culture
Some are visible under microscope	Not visible under 19th century microscopes
Thought to include rabies and smallpox	
Is a living organism	
Q: Is a bacterium or fungi	***?***

Table 3.3: Tobacco mosaic analogical inference

This created a dilemma. The source domain of 'microbe' was the only available source domain from which to infer an explanation for what tobacco mosaic was, yet here were critical dissimilarities between the concept 'microbe' and tobacco mosaic. Three options were explored for resolving the puzzle. First, biologists considered that perhaps there was some way to make the negative analogy less concerning or dismiss some of the findings of dissimilarity. One suggestion was that perhaps tobacco mosaic *was* filterable, but in the preceding experiments, there might have been a gap in the filter that somehow allowed the microorganism through [Zaitlin et al. 1998]. Another posited explanation was that tobacco mosaic was a bacterium, but the substance making it through the filter was a toxin secreted by the bacterium rather than the microorganism itself. Microbiologist Dmitry Iwanowski proposed:

> According to the opinions prevalent today, it seems to me that the latter is to be explained most simply by the assumption of a toxin secreted by the bacteria present, which is dissolved in the filtered sap. Besides this there is another equally acceptable explanation possible, namely, that the bacteria of the tobacco plant

> penetrated through the pores of the Chamberland filter-candles, even though before every experiment I checked the filter used in the usual manner and convinced myself of the absence of fine leaks and openings [Zaitlin et al. 1998].

Through these hypotheses, scientists searched for an explanation that would preserve the similarity between tobacco mosaic and the current accepted definition of microbe, or at least render the dissimilarity non-critical. This endeavor was similar to the Botai example, in which archaeologists committed to the idea that Botai was a site of horse domestication attempted to generate an explanation that would explain the dissimilarities between Botai and known sites of horse domestication.

In the tobacco mosaic example, microbiologists struggled to come up with an auxiliary hypothesis that would lessen the severity of the observed dissimilarity between source and target domain.

Though it was initially considered that perhaps tobacco mosaic could be a toxin or an enzyme, these two options were unable to explain the similarities that the tobacco disease shared with microbes; tobacco mosaic was transmissible and appeared to propagate within the host. A toxin would become less efficacious over time as it spread to new plants because it is a finite substance. In contrast, the tobacco mosaic did not seem to be a finite substance; studies showed that it became no less efficacious as it spread between plants. This did not fit the expected characteristics of an enzyme or toxin.

Scientists were reluctant to conclude that the analogical conjecture was implausible and that tobacco mosaic was not a microbe. First, it *did* appear to be a contagious disease. Second, rejecting the plausibility of the analogical conjecture would place scientists in a position of doubt; tobacco mosaic would be entirely without a category classification. Even in the face of disconfirming evidence, scientists are often reluctant to reject a theory unless another plausible alternative exists [Kuhn 1962, pg. 77]. Kuhn says, that when anomalies

accumulate, scientists "may begin to lose faith and then to consider alternatives, [but] they do not renounce the paradigm that has led them into crisis...The decision to reject one paradigm is always simultaneously the decision to accept another [Kuhn 1965, pg. 77]."

Instead of settling for having no theory about tobacco mosaic, microbiologist Martinus Beijerinck proposed a revision of the source domain such that the concept of microbe would be generalizable to tobacco mosaic [Zaitlin et al. 1998]. He argued that tobacco mosaic was caused by a living infectious fluid, a *contagium vivum fluidum* [Bos 1999]. Under this hypothesis, tobacco mosaic was alive, but it was a smaller lifeform than the cellular bacteria that microbiologists had observed under microscopes. The idea that tobacco mosaic was a fluid, not a cell, would explain the finding that it was not filterable and not visible under a microscope. To explain the fact that tobacco mosaic was not able to self-propagate in a Petri dish, Beijerinck proposed that tobacco mosaic was alive and able to self-propagate, but could not do so in an artificial medium and required a host within which to propagate. He said, "Without being able to grow independently, it is drawn into the growth of the dividing cells and here increased to a great degree without losing in any way its own individuality in the process [Bos 1999]."

Finally, Beijerinck reintroduced the term 'virus'. Since the mid-1800s, 'virus' had been a broad, non-specific layman's term for any entity that causes disease. 'Microbe' was a more precise, scientific subset of the category 'virus'. However, Beijerinck proposed a change in definition in which 'virus' described a specific subset of the category 'microbe'. He proposed that there were two types of microbe, *contagium fixum* (cellular microbes, which obey Koch's postulates) and *contagium vivum fluidum* (fluid, non-cellular microbes that do not obey all of Koch's postulates). This amounted to a revision of the source domain, which had originally posited that all pathogens obey Koch's postulates (i.e. have the characteristics associated with bacteria).

Beijerinck's proposed revision of the source domain was fruitful for explaining anomalies in

other pathogens that had previously been identified as bacteria, like rabies and smallpox, which lacked evidence of some of the properties bacteria should have according to Koch's postulates.

Beijerinck's theory was not immediately accepted, but as dissimilarities between the existing microbe concept and new diseases increased, his theory began to gain traction. His proposed revision of the 'microbe' source domain is illustrated in Table 3.4.

Microbe (late 19th century)	Tobacco Mosaic
***Q:* is a bacterium**	
Cause contagious diseases	Is contagious
Filterable through paper or porcelain filter	Not filterable
Propagation is responsive to temperature and humidity	Unresponsive to temperature and humidity
Self-propagates in culture	Does not propagate in culture
Visible under microscope	Not visible under 19th century microscopes
Is a living organism	
Thought to include rabies and smallpox	
***Q:* Is a virus**	
Cause contagious diseases	
Not filterable through paper or porcelain filter	
Does not self-propagate in culture	
Not visible under 19th century microscope	

Table 3.4: Revision of the representation of the category 'microbe'

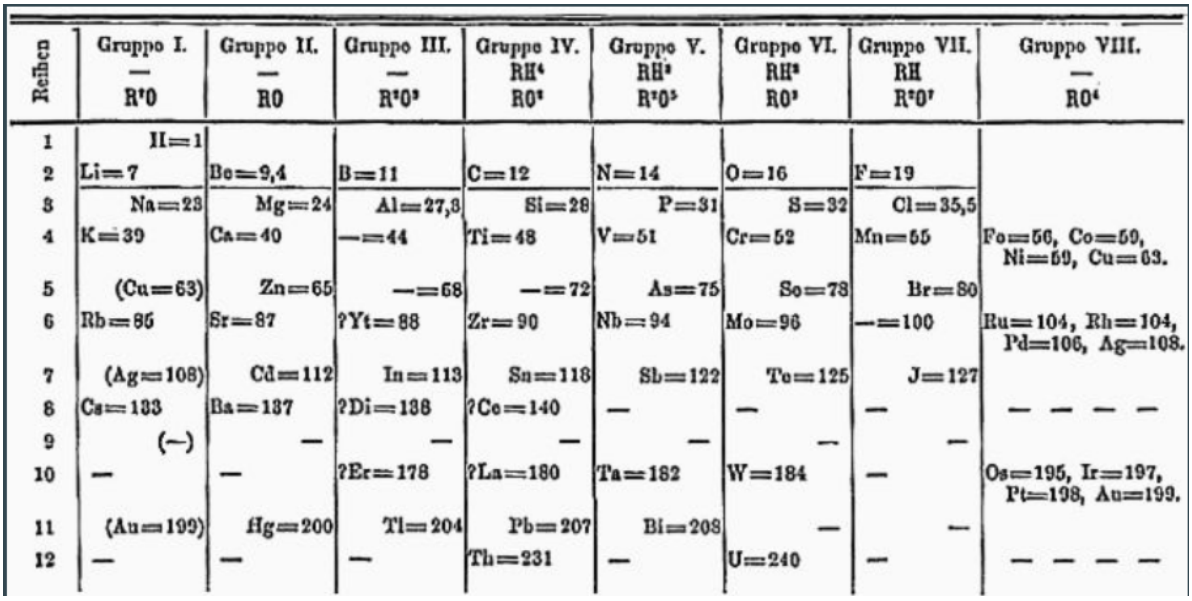
In this example, microbiologists did not merely accept the dissimilarities between tobacco mosaic and the paradigm-approved concept of 'microbe' and conclude that the analogical conjecture was implausible. Instead, they interpreted the dissimilarity as possible evidence for changing their source domain representation. This is a case of reverse analogical inference, in which scientists decide to manipulate the description of the source domain in order to increase the relevant similarity between the source and target domains and preserve the conjecture. Preserving the conjecture that tobacco mosaic is a microbe and altering the representation of the source domain was fruitful, as it led to a more precise and nuanced definition of 'microbe'.

Example 3.5.2. Another example of reverse analogical inference arises in Mendeleev's construction of the periodic table of the elements, which is discussed by Ross [2018]. Mendeleev discovered that if you organized the known elements sequentially by their atomic weight, then some similar physical properties would arise at regular intervals. Mendeleev called this relationship between atomic weight and physical characteristics 'the periodic law'. Mendeleev says, "Elements arranged according to the size of their atomic weights show clear periodic properties. All the comparisons which I have made...lead me to conclude that the size of the atomic weight determines the nature of the elements [Mendeleev 1869]."

Although chemists like Mendeleev and Wurtz did not yet have an explanation for why atomic weight and physical properties were linked, they believed that this regularity of periodicity was a law of nature, not an accidental correlation [Edwards 2020]. Mendeleev used the hypothesized periodic law to successfully predict the existence and properties of elements that had not yet been discovered by making analogical inferences from the properties of elements adjacent to blank spots in his periodic table [Scerri 2001].

However, there was a problem. Some of the elements appeared to violate the law of periodicity. One example of this was Tellurium. According to Mendeleev's periodic law, in order to belong to a row of chemically similar elements——oxygen, sulphur, and selenium——Tellurium's atomic weight should be between 123 and 126. Instead, it was 128, causing the periodic law to make an incorrect prediction about Tellurium's physical properties [Ross 2018].

Mendeleev believed that the periodic law was a fundamental law of nature, so it was concerning to him that periodicity could not be generalized to all of the elements. Rather than taking the irregularity of Tellurium to mean that the periodic law could not be generalized to Tellurium, Mendeleev instead concluded that the representation of Tellurium must be incorrect [Ross 2018]. He claimed that there must have been an error in scientists' measurement of Tellurium's atomic weight. Against experimental evidence, he changed Tellurium's



Reihen	Gruppo I. — R^1O	Gruppo II. — RO	Gruppo III. — R^2O^3	Gruppo IV. RH^4 RO^2	Gruppo V. RH^3 R^2O^5	Gruppo VI. RH^2 RO^3	Gruppo VII. RH R^2O^7	Gruppo VIII. — RO^4
1	H=1							
2	Li=7	Be=9,4	B=11	C=12	N=14	O=16	F=19	
3	Na=23	Mg=24	Al=27,3	Si=28	P=31	S=32	Cl=35,5	
4	K=39	Ca=40	—=44	Ti=48	V=51	Cr=52	Mn=55	Fe=56, Co=59, Ni=59, Cu=63.
5	(Cu=63)	Zn=65	—=68	—=72	As=75	Se=78	Br=80	
6	Rb=85	Sr=87	?Yt=88	Zr=90	Nb=94	Mo=96	—=100	Ru=104, Rh=104, Pd=106, Ag=108.
7	(Ag=108)	Cd=112	In=113	Sn=118	Sb=122	Te=125	J=127	
8	Cs=133	Ba=137	?Di=138	?Ce=140	—	—	—	— — — —
9	(—)	—	—	—	—	—	—	
10	—	—	?Er=178	?La=180	Ta=182	W=184	—	Os=195, Ir=197, Pt=198, Au=199.
11	(Au=199)	Hg=200	Tl=204	Pb=207	Bi=208	—	—	
12	—	—	—	Th=231	—	U=240	—	— — — —

Figure 3.2: Tellurium's revised atomic mass

described weight in his periodic table to make Tellurium obey periodicity [Ross 2018, Scerri 2007].

Mendeleev says: "The atomic weight of an element may sometimes be amended by a knowledge of those of the contiguous elements, thus the atomic weight of Tellurium must lie between 123 and 126 and cannot be 128 [Mendeleev 1869]." In summary, instead of concluding that the dissimilarities between the source and target domain made the analogical conjecture implausible, Mendeleev altered his representation of the target in order to protect the generalizability of the periodic law. Furthermore, he did this even though there was no empirical evidence for the revised representation of Tellurium.

Figure 3.2 shows Mendeleev's new periodic table after he revised the representation of Tellurium.

Tellurium's revised atomic mass dictated that it be moved up a row to share a row with oxygen, sulphur, and selenium. These elements share physical similarities with Tellurium, such as being poor thermal conductors [Scerri 2007].

This new atomic weight was incorrect, but the updated arrangement of Tellurium in the periodic table did correctly place it in the periodic table in terms of atomic number—the number of protons in the nucleus of an atom. The concept of atomic number had not yet been discovered, but atomic number captured the more fundamental cause of periodicity, electron configuration, which atomic mass was merely a proxy for [Ross 2018]. By altering his representation of Tellurium to preserve the analogy between Tellurium and the other elements, Mendeleev made the periodic table more precisely capture the periodic relationship between elements' atomic number and physical properties, even though scientists did not yet have the experimental instruments to find empirical evidence of atomic number.

This example is one in which a scientist has an intuition about the plausibility of an analogical conjecture, and uses this intuition to inform their beliefs about the degree of similarity between the source and target domain. This example is particularly interesting because the intuition is not merely used to interpret the existing empirical evidence and bridge the gap between evidence and interpretation of degree of similarity. Instead, Mendeleev used his epistemic commitment to the periodic law to generate auxiliary hypotheses about the reliability of the empirical evidence, and dismiss certain pieces of empirical evidence.

My goal in this section was to show that scientists often have intuitions or background assumptions about the plausibility of the analogical conjecture that operate independently of the explicit analogical inference. Sometimes scientists are more committed to the existence of a deep similarity between the source and target domain than they are to their existing representations of the source and target domain. Thus, when fundamental dissimilarities are discovered between a source and target domain, scientists sometimes are willing to sacrifice their representations in order to save the horizontal similarity relation and make the analogy go through.

A further moral from this section is that reverse analogical inferences are sometimes fruit-

ful. In the tobacco mosaic example, amending the source domain representation to preserve similarity between tobacco mosaic and other microbes had a positive payoff. The new representation was more precise and resolved some puzzles in microbiology, like the invisibility of rabies and smallpox under 19th century microscopes.

In the periodic table example, Mendeleev's manipulation of the representation of Tellurium resulted in the amended value for Tellurium's atomic mass being incorrect. Yet the manipulation of Tellurium's atomic mass played a positive role in the discovery of atomic number, which atomic mass was merely a proxy for [Ross 2018, Scerri 2007].

The prior two examples provided anecdotal accounts of reverse analogies in science. There is also evidence in the psychology literature that when people make analogical inferences, the process of comparing a source and target affects agents' representations of the source and target [Medin 1993, Hassin 2001].

Example 3.5.3. Consider the following example, which examines how the process of comparing a source and target domain affects agents' perception of the source and target domain's features [Medin 1993]. When agents were tasked with making an initial judgment on how similar the source and target domain are, this process of comparison and the agents' initial reported judgments affect their subsequent judgment of the metric lengths of shapes in the source and target domain [Hassin et al. 2001].

The Ebbinghaus illusion is a perceptual illusion in which the size difference between a central object and peripheral objects causes agents to misjudge the size of the central figure. Hassin et al. hypothesized that asking participants to report on the similarities between the central figures and peripheral figures in the Ebbinghaus illusion would cause participants to perceive the source and target in the Ebbinghaus illusion as being more similar relative to a control group. In other words, comparing the source and target would reduce the effect of the Ebbinghaus illusion. Likewise, Hassin et al. predicted that asking participants to report on

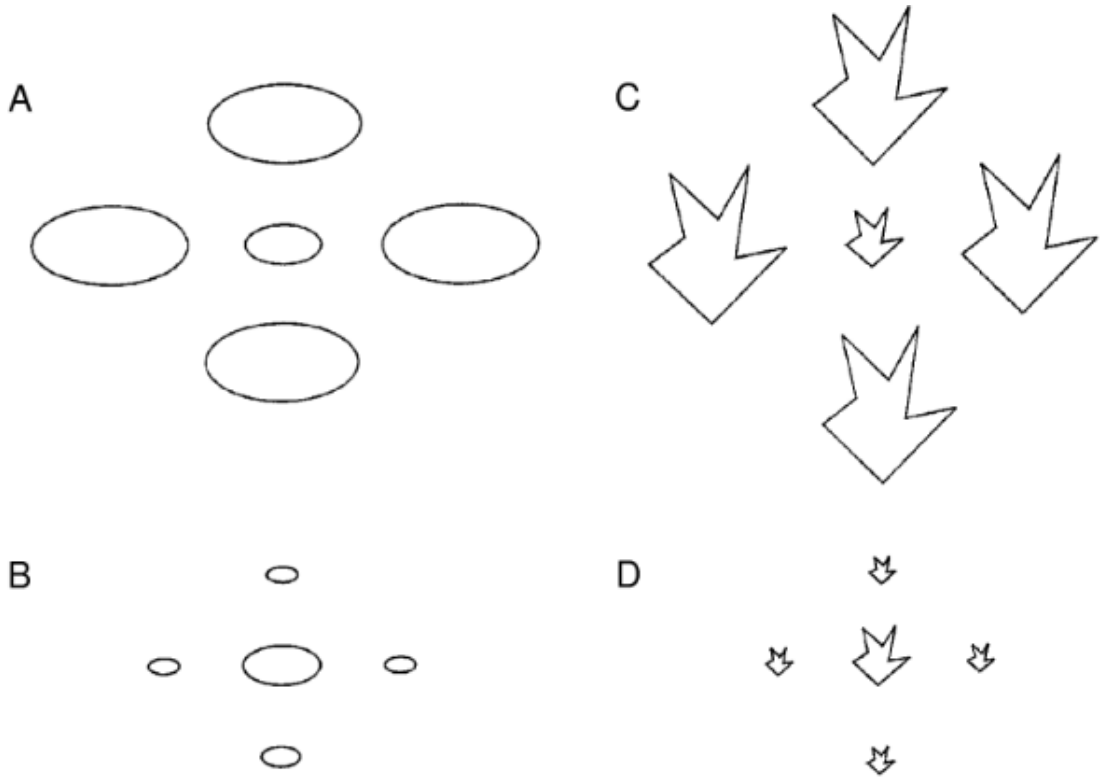


Figure 3.3: Ebbinghaus illusion

the differences between the central figures (X) and peripheral figures (Y) in the Ebbinghaus illusion would cause participants to perceive the source (X) and target (Z) in the Ebbinghaus illusion as being more different than a control group when asked to estimate metric length. In other words, it would increase the effect of the Ebbinghaus illusion.

This prediction was borne out by the evidence. Participants' judgments of similarity or difference systematically affected their estimations of the metric length of shapes relative to a control group who were not asked to make a similarity/difference judgment [Hassin et al. 2001]. Assessing similarity appears to prime agents to choose representations that are more similar to each other, and assessing difference appears to prime agents to choose representations that are more dissimilar. Hassin says, "Similarity judgments enhance perceptual similarity, whereas difference judgments enhance perceptual difference [Hassin 2001]." A similar result was found in [Boroditsky, 2007], which examined how the comparison of animals

on a dimension of either similarity or difference affected agents' subsequent representations of the animals' features. These findings suggest that analogical reasoning is not a linear process in which we settle on representation of the source and target and then judge the degree of similarity between those two representations. Rather, degree of similarity between source and target can be a factor in how people choose to represent the source and target. Medin says:

> It is natural to assume that, to constrain similarity comparison appropriately, the representation of each of the constituent terms must be rigid (i.e., context-insensitive). In contrast, our observations suggest that the effective representations of the constituents are determined in the context of the comparison, not prior to it [Medin 1993].

These studies suggest that gerrymandering of the source and target domain in order to increase or decrease the degree of similarity between the source and target domain is a natural component of the psychological process of analogical reasoning. Furthermore, in some cases it can be epistemically fruitful for science, as shown in Example 3.5.2 and Example 3.5.1. This raises a question for normative accounts of analogical inference: When, if ever, are reverse analogical inferences epistemically justified? Current normative accounts suggest that reverse analogical inferences are universally unjustifiable. Reverse analogical inferences violate multiple assumptions of current accounts: they involve manipulations of the source domain to artificially increase source-target similarity, even in the absence of domain-internal reasons for the revised representations. They involve scientists failing to conclude that an analogical conjecture is implausible even when critical dissimilarities between the source and target are found—explicitly violating the recommendation of current accounts.

However, reverse analogical inferences have sometimes been fruitful for science. Because of this, if we want to advocate for a blanket rejection of reverse analogical inferences, we

should have principled reasons for doing so. The next section will investigate whether there are principled reasons for a normative account of analogical inference to advocate for a methodological rejection of reverse analogical inferences.

3.6 When are reverse analogical inferences justified?

When scientists encounter critical dissimilarities between a source and target domain, they must decide whether to conclude that (1) the analogical inference is implausible, or (2) that the dissimilarity provides justification for changing the source or target domain representation in order to increase the degree of similarity. In the last section, I showed that in practice, scientists sometimes do (2). When critical dissimilarities are discovered, scientists sometimes alter their representations of the source or target domain to preserve similarity and make the analogical inference go through. I call this *reverse analogical inference*.

When, if ever, are reverse analogical inferences justified?

The first option, which is claimed by current normative accounts, is that reverse analogical inferences are in-principle unjustifiable. Under current accounts, evaluation of *Source Domain Salience* and *Overlap* is a linear, two-step process, in which intuitions about the plausibility of the analogical conjecture should not be a factor in choice of representation for the source and target domain. If evaluating *Source Domain Salience* and *Overlap* yields critical dissimilarities between the source and target domain, then scientists should conclude that the analogical conjecture is implausible.

A second option is that reverse analogical inferences are justified or unjustified for precisely the same reasons that any analogical inference is justified or unjustified. If a particular reverse analogical inference is unjustified, it will be for local, domain-specific reasons. This would violate multiple assumptions of current normative accounts, as discussed in Sects.

3.4 and 3.5. It would also undermine the possibility of these accounts providing necessary and sufficient conditions for *prima facie* plausibility, which is an explicit goal of the relevant similarity framework accounts.

Do we want to commit ourselves to the first option and count reverse analogical inferences as paradigmatic examples of faulty reasoning? Given that scientists *do* make reverse analogical inferences, and given that reverse analogical inferences are sometimes fruitful, we should not commit ourselves to a methodological rejection of reverse analogies unless we have principled reasons for doing so.

The final section will provide two arguments against a methodological rejection of reverse analogical inferences. Then, I will argue that reverse analogical inferences should be judged just as all analogical inferences are: according to local, context-dependent norms for relevant similarity, by the epistemic values that are most relevant and virtuous by scientists' best lights, and according to scientists' context-dependent goals and epistemic commitments.

3.6.1 No matter-of-fact about whether existing schema or individual analogical conjectures are more plausible

One argument for there being no in-principle reason for rejecting the plausibility of reverse analogical inferences is that when scientists are in possession of a theory and come across a new piece of data, there is no domain-independent matter of fact about whether the theory should be used to interpret the data or the new data should be used to revise the theory. Sometimes it does not become obvious to scientists that their representation of the source domain is flawed until they attempt to generalize the source domain to new phenomena and discover critical differences.

With every analogical inference in which critical dissimilarities arise, scientists make a choice

between proceeding with the forward analogical inference or switching to a reverse analogical inference. This is an archetypal case of Goodman's notion of reflective equilibrium [Goodman 1955]; scientists weigh their commitment to existing representations against the potential epistemic payoffs of the analogical conjecture being true. Reverse analogical inferences facilitate revisions of entrenched representations when phenomena are observed that contradict the assumptions of existing representations. When scientists revise their representations of the source and/or target domain in order to achieve a greater degree of overlap between the source and target, the outcome can be a revision of the existing theoretical framework that extends its explanatory power and makes it able to handle anomalies.

A methodological rejection of reverse analogical inferences would entail a commitment to a foundationalist, top-down account of justification. This would be undesirable because there is no ultimate justification for our existing schema and inference rules. The foundationalist position is also incoherent with the history of science. Instead, we want a normative account of analogical inference that allows an iterative revision of scientists' intuitions about specific analogical inferences and their general schema and inference rules.

Under an account of analogical inference that permits a process of reflective equilibrium between representations and analogical conjectures, the potential epistemic payoff of an analogical conjecture would serve as indirect evidence for choosing a representation that licenses that conjecture. If an analogical conjecture being true would yield some positive epistemic payoff like explanatory power or predictive success, then this can constitute indirect evidence for choosing a representation that creates a higher degree of source-target similarity and makes the analogical conjecture plausible. To see this, suppose that there is slightly more empirical evidence for representation A than representation B, but representation B would create a higher degree of similarity between source and target domain. If there were a higher degree of similarity between source and target, then this would provide support for an analogical conjecture, Q. Q being true has some positive payoff like explanatory power,

uniting distinct domains, etc. As a result, the fact that B supports Q provides some indirect evidence for choosing representation B over representation A. It is not the only evidence scientists should consider when making their representation choice; they must also consider the independent merits of representations A and B. However, it is rational for intuitions about the analogical conjecture's plausibility to play an indirect role in representation choice.

What does it mean for evidence to play an indirect versus an indirect role? I propose something like a weaker reading of Bartha's *Internal Coherence* reading. Choosing a representation that maximizes the degree of similarity and makes the analogical conjecture plausible is not justified if making the analogy go through is a **direct** motivation for representation choice. "I chose representation A in order to create a higher degree of similarity and make the analogy go through" cannot be the sole consideration when choosing between A and B. However, it is justifiable for intuitions about the plausibility of the analogical conjecture to be an **indirect** motivation for representation choice, in weighing the evidence for different representations.[3]

Consider an example of the motivation to make the analogy go through playing an indirect role.

Example 3.6.1. Maxwell's electromagnetic field theory united magnetic and electric fields under one theory, electromagnetic field theory. Doing so involved some manipulations of current representations that seemed specious to scientists at the time. For instance, Maxwell's electromagnetic field theory relied on postulating an imaginary force, displacement current, that had no direct empirical support [Krishnasamy 2009]. Displacement current was necessary to explain the continuity of the magnetic field between gaps in electrical current, such as in insulation between capacitors. Maxwell proposed the displacement current as a different kind of current that was producing the magnetic effects in the absence of electrical current. The only evidence for there being such a current was indirect; its existence would explain

[3]This distinction between direct and indirect motivations is inspired by Heather Douglas' distinction between direct and indirect roles of social values in theory choice [Douglas 2016, pg. 10].

the presence of magnetic fields in cases where electric current is absent, and therefore make Maxwell's equations mathematically consistent [Krishnasamy 2009].

Maxwell's theory also implied the existence of electromagnetic waves, which had no empirical evidence until twenty years later, in 1887. However, Maxwell's field theory had a number of virtues, like explanatory power and uniting different physical laws. These epistemic virtues constituted indirect evidence for the existence of displacement current and electromagnetic waves even though there was not direct empirical evidence for these entities.

Another example is evidence for the existence of dark matter. The indirect evidence for dark matter is that it makes our current physical theories mathematically consistent, though this could alternatively be achieved by altering our representation of gravity, as modified Newtonian dynamics does [Baerdemaeker 2021].

Scientists judge the plausibility of an analogical inference against their overall web of belief. In that web of belief, scientists might have the belief that explanatory power is an epistemic virtue. They might also have the belief that if a particular analogical conjecture, Q, were true, Q would have explanatory power and would unite several distinct laws. Considered independently of an analogical argument for Q, perhaps a representation A has more empirical support than representation B. However, if representation B creates a stronger analogical argument for Q than A does, then this could constitute indirect evidence for choosing representation B depending on the *de facto* degrees of epistemic commitment scientists have to to A, B, and Q.

This weaker reading of *Internal Coherence* contradicts current normative accounts of analogical inference. Under current accounts, evaluation of an analogical inference's plausibility is a linear process consisting of two steps, evaluation of *Source Domain Salience* and evaluation of *Overlap*. Evaluation of *Source Domain Salience* and *Overlap* is supposed to be epistemically prior to evaluations of the plausibility of an analogical inference. Current accounts

say that if scientists complete their evaluations of *Source Domain Salience* and *Overlap* and discover a critical dissimilarity between the source and target domain, then they should conclude that the analogical inference is implausible. In contrast, *Internal Coherence (Weak)* suggests that it is justifiable for scientists' intuitions about the plausibility of an analogical conjecture to affect their evaluations of *Source Domain Salience* and *Overlap*. Thus, it is plausible to read Bartha as supporting *Internal Coherence (Strong)*.

However, this paper argues that *Internal Coherence (Strong)* is too strong and that *Internal Coherence (Weak)* is a better description of fruitful scientific and mathematical practice. Under this reading, if scientists discover critical dissimilarities between a source and target domain and choose to revise their source and target domain representations instead of concluding that the analogical conjecture is implausible, this move may be justified depending on their domain-specific epistemic commitments.

Internal Coherence (Weak) also yields several other epistemic virtues. First, *Internal Coherence (Weak)* avoids a theory entrenchment problem that *Internal Coherence (Strong)* is susceptible to. Gerrymandering a representation to preserve an analogical conjecture could be interpreted as a case of scientists exercising a confirmation bias. However, defending *Internal Coherence (Strong)* also makes our analogical inferences susceptible to confirmation bias in a different way. Under *Internal Coherence (Strong)*, analogical inference is a process in which an established conceptual framework in the source domain is used to shape theories about new domains, yet there is no route for these new domains to challenge or reshape the conceptual framework expressed in the source domain. Then, the process of analogical inference is a process of importing an existing conceptual framework and set of background assumptions into new domains. This can encourage a self-perpetuation of a set of entrenched assumptions, where scientists choose research questions, design studies, and interpret results through the lens of the dominant framework. Reverse analogical inferences help prevent this by allowing new data to shape the existing conceptual framework rather than only allowing

the existing conceptual framework to shape interpretations of new data.

A second epistemic virtue of reverse analogical inferences is that they can support investigating the potential fruitfulness of new conceptual frameworks. For instance, in Mendeleev's construction of the periodic table, reverse analogical inference helped preserve Mendeleev's concept of the periodic law until independent evidence for it could be found. Several philosophers suggest that scientists having a bit of 'stickiness' in their epistemic commitments to non-entrenched hypotheses may be an epistemic virtue. When a fringe hypothesis is developed, sometimes the hypothesis must reach a certain threshold of support in order for evidence for the hypothesis to be recognized as such. Longino [1990] argues that data itself does not have evidential significance—the data must be embedded in a framework that makes it significant as evidence [Longino 1990]. Kuhn claims that paradigms shape what counts as evidence [Kuhn 1962]. Feyerabend, similarly, suggests that a certain threshold of consensus is needed in order for a fringe theory to generate the evidence needed to obtain empirical support [Feyerabend 1975].

In summary, one argument against a methodological rejection of reverse analogical inferences is that reverse analogical inferences permit a fruitful process of reflective equilibrium, can help generate unconceieved alternative hypotheses, and help avoid dogmatic entrenchment.

3.6.2 Not a direct connection between analogical inference and plausibility of conjecture

A second reason to question a methodological rejection of reverse analogical inferences is that reverse analogical inferences may operate as a search method for finding strong analogical inferences in cases where there is a disconnect between intuitive analogical reasoning and explicit analogical inference.

Why does Bartha claim that scientists should find an analogical conjecture implausible if critical dissimilarities are found between the source and target domain? Under Bartha's account, the degree of relevant similarity between the source and target domain representations is the *reason* that scientists find the analogical conjecture plausible. Thus, finding that these representations are critically dissimilar invalidates the reasons that caused the analogical conjecture to be plausible to scientists.

It makes sense, then, that Bartha claims that it is irrational for scientists who discover critical dissimilarities to revise their representations to create similarity rather than concluding that the analogical inference is implausible. This gerrymandering seems like a fallacy of confirmation bias. However, this practice might be rational if it were the case that the reasons cited in the analogical inference diverged from the reasons that caused scientists to find the analogical conjecture plausible. If the factors that made the analogical conjecture plausible were not cited in the analogical inference, then the failure of the analogical inference need not spell the failure of the analogical conjecture.

Traditional accounts of reasoning assume that the relationship between a judgment and the reasoning that accompanies that judgment is one where the reasoning precedes and causes the judgment. In contrast to this assumption, the contemporary cognitive science literature claims that there are two different kinds of reasons involved in human cognition: one kind of reason that is causative in the psychological process of a conjecture being formed, and a different kind of *post hoc* reason that is invoked when reconstructing a reasoning process for the purpose of examination or justification [Sperber et al. 2012; Haidt 2001]. Cognitive scientists propose that human cognition involves two systems: System 1, which generates judgments quickly, intuitively, and subconsciously, and System 2, which is slow, conscious, and invokes explicit rules of inference.

Consider an example. When seeing someone's facial expression, the amygdala conducts an instantaneous evaluation of that person's emotion based on innate heuristics and condition-

alized pattern recognition [Fiske et al. 2007]. This is System 1 processing. However, after this initial assessment, one can also invoke System 2 processing: one might consider explicit rules like 'If the corners of the mouth are upturned, the person is happy or bemused'. People are particularly likely to invoke System 2 processing if they need to reconstruct their reasoning process for the purpose of justification, such as if someone questions the conclusion [Sperber 2011]. The deliberative process of constructing an analogical inference to justify an analogical conjecture is a System 2 process, but many cases of analogical reasoning appear to rely on tacit knowledge and intuition, which is the domain of System 1 [Holyoak 1995; Hofstadter 2013].

For instance, a philosopher of legal reasoning notes that:

> What makes analogical reasoning distinctive is that although people who draw analogies see similarities that are necessarily based on principles or theories, these principles or theories are often so embedded in their thought processes that they are not consciously perceived...there is no more reason to believe that the legal analogizer consciously retrieves [a priori] principles in drawing an analogy than there is reason to believe that the expert chess player consciously retrieves the principles of chess in making her next move, or that the expert musician consciously retrieves (or even understands) the principles of aural perception in hearing that a particular note is sharp or flat [Schauer 2017].

Much analogizing depends on knowing *how*, not knowing *that* and relies on associative learning and genetically and culturally inherited intuitions to fix respects for similarity [Gentner 1993]. One explanation for this is that explicit analogical inference is resource and memory-intensive [Halford 1992], and relying on less costly unconscious heuristics may be evolutionarily adaptive.

Early dual-process models proposed that System 2 serves the function of monitoring, re-

constructing, and rationally evaluating the steps performed by System 1 [Kahneman 2011]. However, more recent research suggests that this is a mischaracterization of System 2, for several reasons. First, people frequently lack epistemic access to the System 1 reasons for their judgments. An agent can give an account of what reasons someone could have to find an analogical conjecture plausible, but these cited reasons may diverge from the reasons that were actually causative in producing the judgment. Consider an example from Nisbett et al. [1977].

Example 3.6.2. A group of undergraduate psychology students were asked to memorize a list of word pairs, one of which was the pair 'ocean-moon'. Experimenters hypothesized that memorizing the pair 'ocean-moon' would cause an implicit association with the word 'tide', and would make participants more likely to name the brand 'Tide' when later asked to name a detergent brand. This did end up being the case; the participants whose word list included the 'ocean-moon' pair were twice as likely to name the detergent Tide as participants whose word list did not include 'ocean-moon'. Furthermore, experimenters found that participants lacked epistemic access to the reasons for their response. When asked why they responded 'Tide' instead of a different detergent brand, participants gave reasons that did not acknowledge any possible influence of the word pair memorization task, such as "I like the Tide box" [Nisbett et al. 1977].

This experiment, as well as numerous others [Maier 1931, Johansson et al. 2012] find that when reasoners are asked to report on their reasoning process, they provide a justification for their conclusion that often differs from the actual contents of their reasoning. Nisbett et al. suggest that we have epistemic access to the conclusions of our reasoning, but often lack epistemic access to the contents of the reasoning processes that lead to those conclusions.[4]

Nisbett and Wilson claim that agents are under an illusion in which, because they have

[4]More recent experiments by [Johnasson et al. 2012] and other cognitive scientists suggest that our beliefs as well as a our reasoning processes can be confabulated from environmental content rather than accessed introspectively.

introspective access to the *conclusions* of their reasoning processes, they mistakenly believe that they have epistemic access to the reasoning processes themselves [Nisbett et al 1977]. In reality, agents rely on *a priori*, folk psychology theories to interpret their own reasoning process just as they would if speculating about the reasoning of a third party [Nisbett et al. 1977]. Nisbett et al. say

> In the experiments reviewed above, then, subjects may have been making simple representativeness judgments when asked to introspect about their cognitive processes...The familiarity of a detergent and one's experience with it seem representative of reasons for its coming to mind in a free association task [Nisbett et al. 1977].

If this lack of introspective access occurs in analogical reasoning as well, then the reasons cited in an analogical inference can diverge from the reasons that were causative in the psychological process of analogical reasoning.

Several psychology studies support the claim that agents can lack epistemic access to the contents of their analogical reasoning. Schunn et al. [1996] demonstrate that agents can be primed to generalize a causal explanation from a source to a target domain without having conscious awareness that the source domain is affecting their reasoning. A more recent set of experiments conducted by Hristova et al. [2025] establish that (1) analogical reasoning can be unconsciously affected by external factors, without agents being aware of the effect of these factors and (2) agents use analogical reasoning to form conclusions without being aware of doing so. In such cases, agents are aware of holding a belief about a conjecture, without being aware of the contents, or even the existence of, the analogical reasoning process that led to that conjecture [Hristova 2025]. A 2007 study by Gentner et al. conducted several experiments with the goal of determining whether the lack of epistemic access to reasoning processes reported by Nisbett [1977] also occurs in cases of analogical

reasoning. They found that it does; in their experiments, participants' conclusions are affected by analogical transfer, yet when subsequently surveyed, participants reported no awareness of the analogical reasoning influencing their conclusion. In Example 4.2.3. in Chapter 4, I discuss a study by Gilovich [1981] that establishes agents' lack of epistemic access to the contents of their analogical reasoning.

There is an interesting phenomenon in the science literature that would be explained by the claim that constitutive ingredients of analogical reasoning can be unconscious and thus are missing from the explicit analogical inferences that reconstruct that analogical reasoning in the form of an argument. Many historical analogical inferences ended up being fruitful despite citing causal relationships and similarities that were subsequently discredited. This indicates that the explicitly cited relationships and similarities in these analogical inferences may not have been the sole, or even the primary, ingredient in the analogical reasoning. Instead, unconscious similarity judgments or background assumptions may have played a significant, yet never explicitly articulated role in motivating the analogical conjecture. Consider an example discussed by Bartha [2009].

Example 3.6.3. Salicylic acid, the main ingredient in aspirin, was discovered by an English reverend, Edward Stone, who spontaneously chewed on a piece of willow bark [Bartha 2009]. Stone noted similarities between willow bark and Peruvian bark, which was a known antipyretic. On the basis of these similarities, he inferred that English willow bark might also be an antipyretic.

In a letter to the president of the Royal Society justifying his conjecture, Stone wrote,

> I accidentally tasted [willow bark], and was surprised at its extraordinary bitterness; which immediately raised me a suspicion of its having the properties of the Peruvian bark. As this tree delights in a moist or wet soil, where agues chiefly abound, the general maxim, that many natural maladies carry their cures along

> with them, or that their remedies lie not far from their causes, was so very apposite to this particular case, that I could not help applying (the maxim)[Volmink 2008].

Under Bartha's account, an analogical inference must cite an accepted causal relationship in the source domain. Stone cited the 'doctrine of signatures', a widely-accepted belief of the time which posited that the causes of maladies tend to be physically connected to or spatially contiguous with their cures. Stone argued that damp locations with decaying matter tend to cause fever (in accordance with miasma theory), and that Peruvian willow bark grows in damp places with decaying plant matter. This property was thought to be causally related to Peruvian bark being efficacious in treating fever.

Noting that the English willow also thrives in damp, decaying environments, Stone proposed that willow might also serve as an antipyretic. The similarity in bitter taste that Stone noted may have also played a role in his analogical reasoning, but this would not qualify as a relevant similarity under normative accounts of analogical inference because there was no causal relationship linking bitterness to fever-reducing properties in the source domain [Bartha 2009]. Given that Stone's analogical conjecture was correct, it is possible that Stone's analogical conjecture was a result of him intuiting some salient similarity between Peruvian bark and English willow. However, if so, this similarity was not cited in the analogical inference that Stone used to justify his conclusion. Instead, when reconstructing his reasoning as an inference, he used an argument that would be persuasive to scientists of the time–the doctrine of signatures. The doctrine of signatures is more representative of 18th-century milieu accepted justificatory norms than are the unconscious, intuitive similarity judgments that may have actually underpinned Stone's conjecture.

Relying on representations that are widely accepted by one's scientific community makes for persuasive analogical inferences. However, the psychology research discussed above suggests

that the business of making persuasive analogical inferences may sometimes come apart from the business of reporting accurately on the analogical reasoning processes. Furthermore, the psychology research suggests that when a scientist's analogical reasoning and analogical inference do diverge, the scientist will not be aware of the divergence.

Further support for the claim that analogical reasoning and analogical inference may sometimes come apart comes from Sperber [2011, 2016], who suggests that explicit reconstructions of reasoning serve a different purpose than traditionally assumed. Sperber argues that System 2 reasoning serves the function of constructing arguments to persuade and justify oneself to others, and evaluating arguments offered by others. Under Sperber's account, System 2 reasoning is not truth-seeking—it does not monitor and refine the reasoning conducted by System 1. Instead, it merely comes up with arguments that support the conclusions generated by System 1. Sperber argues that many observed characteristics of System 2 reasoning, such as its propensity for confirmation bias, suggest that it has evolved to persuade and justify the conclusions reached by System 1, not to monitor and refine the reasoning process of System 1. This is supported by several psychology studies. Consider one example.

Example 3.6.4. Experimenters give participants the following puzzle:

'A bat and a ball cost $1.10 together. The bat costs $1 more than the ball. How much does the ball cost?'

Participants often quickly give the answer '10 cents', relying on quick System 1 processing and the heuristic that $1.10 minus $1 equals 10 cents [Frederick 2005]. If System 2 refines System 1 reasoning, then giving people more time to complete the puzzle should improve the quality of their answer. Instead, Frederick [2005] found that of the participants who *do* get the correct answer, the majority do so within a few seconds of being shown the problem, suggesting that they reach the correct answer through System 1 processes. Even more interestingly, giving participants more time to solve the problem did not significantly

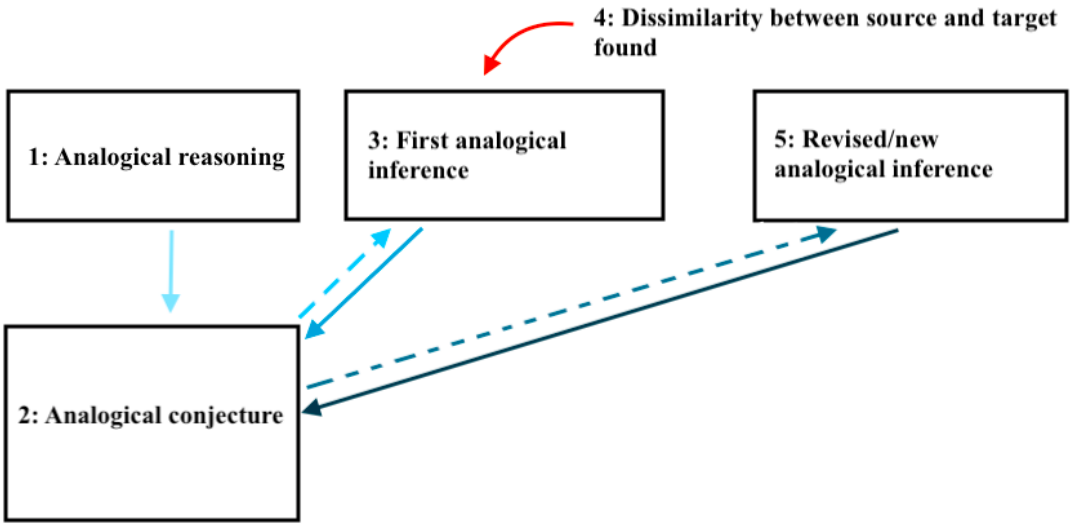


Figure 3.4: The indirect relationship between analogical reasoning and inference

improve the proportion of participants who get the correct answer, suggesting that System 2 may not have the function of monitoring and correcting errors made by System 1. These findings were supported by other studies [Sperber 2016; Bago et al. 2019].

Sperber suggests that rather than System 2 directly monitoring and correcting for the errors of System 1, System 2 instead creates publicly available arguments or justifications for the conclusions that are reached by System 1. Then, these reasons are evaluated by either other reasoners or reflectively by the agent themselves. This may be an effect of the inordinately social environment human reasoning evolved to operate in.

In Figure 3.4 below, I illustrate a picture of the relationship between analogical reasoning and analogical inference that is suggested by the above studies.

In the figure, the process of analogical reasoning directly affects the analogical conjecture. However, there is no direct connection between the analogical reasoning and analogical inference. Instead, agents epistemically access their analogical conjecture and then construct a *post hoc* analogical inference that explains and justifies that conjecture. If critical dissim-

ilarities are found between the source and target domain, then agents revise or construct a new analogical inference that would support the conjecture. The analogical inference is a *post hoc* theory about what the contents of the analogical reasoning process may have been that led to that conjecture.

In summary, two morals can be drawn from the above studies. First, it is plausible that when agents reconstruct their reasoning as an analogical inference to justify a conclusion, the inference may not distill out the components of the analogical reasoning that caused agents to find the analogical conjecture plausible. Rather, scientists confabulate a story about what factors *could* have caused them to find the analogical conjecture plausible. This confabulation will rely on whichever salient explanations are on hand, which will include the accepted laws, theories, and explanatory standards of the scientists' paradigm.

Analogical inferences may serve the purpose of justifying conjectures, rather than capturing the contents of the underlying analogical reasoning that led to the conjecture. The justification will appeal to whatever factors are salient to the agent's conscious awareness, and will miss the unconscious factors. What factors are salient to an agents' conscious awareness will depend on factors like what causal narratives are popular in their scientific community, epistemic and non-epistemic values, and what properties evolution has equipped them to be consciously aware of versus which are relegated to attention-saving unconscious heuristics.

If this portrayal of the relationship between analogical reasoning and analogical inference is correct, then this renders unsurprising the finding that when the reasons cited in an analogical inference are invalidated, scientists may not automatically reject the analogical conjecture. If there is an indirect connection between the analogical reasoning and the reasons cited in the analogical inference, then when the premises of the analogical inference are called into doubt, the analogical conjecture is not directly threatened–it is only threatened if no alternative argument for the conjecture can be found. If scientists lack epistemic access to the contents of their reasoning, then they do not know whether a threat to the analogical

inference is a threat to the contents of their actual reasoning process. Therefore, it may be rational for scientists who encounter dissimilarity between a source and target domain to explore whether a different representation could support the analogical conjecture rather than abandoning the conjecture forthwith.

In claiming that analogical inferences may be *post hoc* confabulations, I do not mean to imply that analogical inferences are not valuable tools that refine agents' reasoning. Bartha makes the point that:

> The point of articulating an analogical argument in precise form is not just to persuade others to accept a view that you have already accepted on independent grounds. The exercise helps you to check your own reasoning...making the reasoning explicit can aid in the creative process by bringing out weaknesses and suggesting remedies [Bartha 2009, pg. 5].

I am in agreement with Bartha. Reconstructing one's reasoning as an explicit argument plays a useful role in checking and refining one's beliefs. However, I suggest that the relationship between analogical reasoning and analogical inference is less direct than current accounts of analogical inference assume. Because of this, it is not de facto irrational for agents to weigh their epistemic commitment to their representations and analogical conjectures, and revise the reasons cited in the analogical inference if their initial reasons are invalidated.

3.7 A possible objection

One possible objection to my argument could be to capitalize on this distinction between *analogical reasoning*, the psychological process by which scientists find an analogical conjecture plausible, and *analogical inference*, a justification or argument for the plausibility of the

analogical conjecture. One could argue that reverse analogical inference is a component of analogical reasoning, but does not play a role in the justification of an analogical conjecture. Then, reverse analogical inference could be a factor in what causes scientists to find a conjecture, Q, plausible, but does not serve as a justification for the plausibility of Q.[5] Could it be argued that reverse analogical inferences are a part of analogical reasoning, but not analogical inference, and thus do not constitute a violation of current normative accounts?

In the current literature, analogical inference is generally treated as a distillation of the essential components of analogical reasoning. Bartha says, "An analogical argument is an explicit form of analogical reasoning, with premises and a conclusion [Bartha 2009]."

There are two ways in which an analogical inference could be an explicit form of analogical reasoning.

The first option is that analogical inferences capture the essential epistemic contents of analogical reasoning; an analogical inference is an explicit representation of the essential premises that cause a scientist, or community of scientists, to find it plausible that an analogical conjecture holds in the target domain.

The second option is that analogical inferences provide an explicit argument for why an analogical conjecture could plausibly hold in the target domain, but the reasons cited in the analogical inference can diverge from the reasons that actually caused scientists to find the analogical conjecture plausible. Analogical arguments give reasons to find a conjecture plausible, but not necessarily *the* reasons that caused scientists to find the conjecture plausible. If this were the case, then analogical reasoning cannot be reduced to the contents of the analogical inference that aims to justify that reasoning.

[5]To understand this distinction, consider someone who has the psychological disposition to always believe that red cars are always faster than gray cars. When asked to justify their belief that a red car will accelerate faster than a gray car, the person cites reasons like the engine size and the aerodynamics of the two cars. The belief that red cars are faster plays a role in their belief that the red car will be faster, but not in their argument justifying that belief.

Bartha explicitly states that he is not committed to either descriptive claim about the relationship between analogical reasoning and analogical inference. However, he says that he is committed to a normative claim: that analogical reasoning that cannot be formalized as an analogical inference is not justified. An analogical conjecture reached through analogical reasoning is not plausible unless the reasoning can be represented as an analogical argument that satisfies the normative criteria of the articulation model.

Bartha says,

> Why, if our ultimate goal is to understand analogical reasoning, should we focus our attention on analogical arguments? The basic reason is simple...a conclusion reached via analogical reasoning is justified only insofar as a reconstruction of the reasoning as an analogical argument can justify that conclusion...Justifiable analogical reasoning is capable of representation in argument form [Bartha 2009].

So, suppose that we want to claim that reverse analogical inferences do not violate current normative accounts because they are only components of the analogical reasoning that causes scientists to find an analogical conjecture plausible, not the analogical inference that justifies the conjecture's plausibility. Then we must commit ourselves to the claim that reverse analogical inferences are unjustifiable and should not influence scientists' assessments of plausibility.

This does not seem like the way we want to go about things. Reverse analogical inferences are essential machinery in the process of reflective equilibrium that scientists use to decide whether a conjecture is justified. Normative accounts should provide criteria for reverse analogical inference.

3.8 Conclusion

In this chapter, I showed that analogical inferences sometimes depart from the framework sanctioned by current normative accounts. When scientists discover critical dissimilarities between a source and target domain, they sometimes alter their representation in order to preserve the plausibility of the analogical conjecture. These reverse analogical inferences violate current normative accounts.

However, reverse analogical inferences can sometimes be fruitful. Given that scientists use reverse analogical inferences and that these inferences can be fruitful, I suggest that if normative accounts advocate for a methodological rejection of reverse analogical inferences, they should provide principled reasons for doing so. I then provide two arguments against a general, methodological rejection of reverse analogical inferences. Instead, I argue that reverse analogical inferences will be justified or unjustified for domain-specific reasons, just as standard analogical inferences are. Scientists use their *de facto* epistemic commitments to decide whether to make a reverse analogical inference or conclude that an analogical conjecture is implausible.

Chapter 4

Establishing the Epistemic Reliability of Similarity Judgments

"What makes Goodman's example a puzzle, however, is the dubious scientific standing of a general notion of similarity, or of kind. The dubiousness of this notion is itself a remarkable fact. For surely there is nothing more basic to thought and language than our sense of similarity; our sorting of things into kinds. We cannot easily imagine a more familiar or fundamental notion than this, or a notion more ubiquitous in its applications...On this score it is like the notions of logic: like identity, negation, alternation, and the rest. And yet, strangely, there is something logically repugnant about it. For we are baffled when we try to relate the general notion of similarity significantly to logical terms [Quine 1969, pg. 117]."

4.1 Introduction

Similarity judgments are an essential component of inductive and analogical reasoning. Hume's skeptical solution to the problem of induction is that our psychological nature compels us to form associations between similar events, and that this innate disposition, rather than reason, is our most reliable available guide to successful action. Quine emphasizes that similarity judgments are fundamental to inductive reasoning, and calls for naturalistic investigation of the cognitive processes underlying similarity.

Despite the important role that similarity judgments play in analogical inference, the notion of similarity has not received adequate examination in normative accounts of analogical inference. Instead, similarity has been treated as a foundational and unproblematic building block of analogical inference. This treatment spans both cognitive science accounts of analogy like Gentner and Holyoak's accounts, philosophical accounts like Bartha, Hesse, and Norton's, and formal accounts like Carnap's.

Normative accounts of analogical inference aim to justify the reliability of analogical inference as a form of argument and provide a toolkit for scientists to evaluate the analogical inferences they encounter and use analogical reasoning more fruitfully. Current normative accounts of analogical inference rely on similarity judgments as the basis of analogical inference, yet these accounts do not investigate the conditions under which scientists' similarity judgments are reliable.

A puzzle arises from the fact that similarity judgments play a central role in scientists' investigation of the nature of reality, yet psychology studies show that similarity judgments are constructed by our cognition and are often non-veridical [Hoffman 1998, Ghirlanda et al. 2003, Polk et al. 2002, Winawer et al. 2007]. There has not been integration of these empirical findings into philosophical accounts of similarity. Research also suggests that scientists often lack epistemic access to the factors that influence their similarity judgments [Maier 1930, Ghirlanda et al. 2003]. This highlights the importance of naturalistic investigation of the role of similarity in scientific reasoning rather than accounts that rely on the first-person authority of scientists.

This chapter will examine how similarity judgments are constructed when scientists generalize templates or models to new domains. First, I will consider findings from cognitive science, anthropology, and psychology that demonstrate the constructive and contingent nature of our similarity judgments. I also examine evolutionary narratives in anthropology to understand the types of survival problems that biases in similarity judgments may have evolved to solve, and how these contexts might differ from the demands of scientific enquiry. I discuss what this implies for the reliability of similarity judgments in scientific reasoning.

Our sense of similarity evolved to reason inductively in situations that we encountered in our evolutionary history, and thus are not always perfectly adapted for the context of scientific enquiry. Our similarity judgments rely on heuristics that produce unfruitful judgments in certain contexts. For instance, when presented with two stimuli, A and B, subjects

demonstrate asymmetric similarity judgments where A is judged to be more similar to B than B is to A, or vice versa, depending on the order or frequency at which A and B are shown to subjects [Polk et al. 2002]. Quine says, "The brute irrationality of our sense of similarity...offers little reason to expect that this sense is somehow in tune with the (truths of nature) [Quine 1969, pg. 13]."

I argue that normative accounts of analogical inference should provide epistemic norms that refine similarity-based reasoning in light of these findings. Normative accounts in epistemology exist not merely to explicate existing epistemic norms, but to bolster our reasoning in areas where we are susceptible to bias or irrationality. However, this normative account will diverge from previous normative accounts of similarity. Unlike prior accounts, my positive account does not purport to provide a once-and-for-all justification of similarity or characterization of what similarity judgments are reliable.

Instead, my account proposes a refinement and extension of Quine and Goodman's accounts. I examine the factors that *do* affect scientists' judgments of similarity, including psychological biases, cultural norms, individual learning, and scientific paradigms. Scientists must decide by their best lights what similarity norms to adopt in their local domains. My account uses historical cases and the natural science literature to posit epistemic values that can help inform scientists' deliberation.

4.2 What is similarity?

Similarity judgments are crucial for analogical and inductive reasoning. It is on the basis of our similarity judgments that we extrapolate knowledge from past experiences to new contexts and predict which actions will be successful. Hume says,

> "In reality, all arguments from experience are founded on the similarity which

> we discover among natural objects, and by which we are induced to expect effects similar to those which we have found to follow from such objects. And though none but a fool or madman will ever pretend to dispute the authority of experience, or to reject that great guide of human life, it may surely be allowed a philosopher to have so much curiosity at least as to examine the principle of human nature, which gives this mighty authority to experience, and makes us draw advantage from that similarity which nature has placed among different objects. From causes which appear similar we expect similar effects... When a new object, endowed with similar sensible qualities, is produced, we expect similar powers and forces, and look for a like effect [Hume EHU, IV. II].

Under Hume's account, all of our reasoning ultimately depends upon our *impressions*–the sensations and perceptions of our immediate experience. "Impressions [are associated] only by resemblance [Hume, Treatise 2.1.4.4]."

Under Hume's account, all *matter of fact* beliefs, his term for *a posteriori* beliefs, are built up from the associations between impressions. The three associations are resemblance, contiguity, and cause and effect. However, the latter two associations are dependent on resemblance under Hume's account. The association of cause and effect relies on resemblance, because in order for an association to be learned between A-type events and B-type events, the A-type events must resemble each other in some manner, and likewise the B-type events must resemble each other. Hume says that in order for a causal association to form, "The effect and cause must bear a similarity and resemblance to other effects and causes [Hume EHU 1.11]."

Contiguity is also dependent on resemblance. Every time I open the door of my favorite coffee shop, I experience a scene that is similar, but not identical, to prior scenes. Each time, chairs are arranged differently, the room has a slightly different orientation relative to the position

of the sun, the artwork on the walls may be different. Spatial and temporal contiguity is the association of similar (but not necessarily identical) A-type events with spatially or temporally contiguous to similar (but not necessarily identical) B-type events. Thus, for Hume, all inductive beliefs depend on the cognitive faculty that generates resemblance. In order for our inductive beliefs to be reliable, the cognitive faculty that generates resemblance must be reliable. Thus, the question of how the faculty of resemblance works and whether it is reliable is crucial to Hume's account.

Hume explicitly states this at multiple points in the Treatise and EHU. In the Treatise, he says, "The first [relation] is resemblance, and this is a relation without which no philosophical relation can exist since no objects will admit of comparison but which have some degree of resemblance [Hume, Treatise I.I.5]."

In the EHU, Hume says,

> If all the scenes of nature were continually shifted in such a manner that no two events bore any resemblance to each other, but every object was entirely new, without any similitude to whatever had been seen before, we should never, in that case, have attained the least idea of...a connexion among the objects [Hume EHU].

Though Hume claims that the faculty of resemblance underlies inductive reasoning, he does not examine the nature or reliability of this mental faculty. Instead, he treats the perception of resemblance as an innate, unconscious, primitive feature of the mind.

This can be seen in several places. In the Treatise, Hume emphasizes that resemblance is often instantaneous and automatic: "When any objects resemble each other, the resemblance will at first strike the eye, or rather the mind, and seldom requires a second examination [Hume, Treatise I.III.I.II]."

Hume claims that the faculty of resemblance is too simple and basic to admit any philosophical investigation. Instead, he treats resemblance as a primitive foundation of cognition that merely exists and cannot have its causes explicated by psychological investigation [Ross 1991].

Hume says that the association of resemblance:

> "is a kind of attraction, which...as to its causes, they are mostly unknown, and must be resolved into original qualities of human nature, which I pretend not to explain. Nothing is more requisite for a true philosopher, than to restrain the intemperate desire of searching into causes, and having established any doctrine upon a sufficient number of experiments, rest contented with that, when he sees a farther examination would lead him into obscure and uncertain speculations. In that case his enquiry would be much better employed in examining the effects than the causes of his principle [Hume Treatise].

So, Hume's answer to the question of what causes resemblance is merely that our mind naturally constructs the perception of resemblance according to an innate psychological process that we cannot epistemically access, yet cannot resist assenting to. Hume does not attempt to provide a description of this psychological process or a justification of its reliability for belief formation. Why does Hume think that the cognitive causes of resemblance cannot be discovered? Hume's assertion that resemblance is irreducible is surprising given his agenda to break down the machinery of the human mind into its simpler constituents, as Newton breaks down the interaction of bodies into constituent laws.

The answer to this is likely due to the reflective, introspective nature of Hume's account. He aims to discover the underlying laws that govern the mind by reflecting on his own conscious experience. When doing so, Hume finds himself unable to introspectively discover the causes of his perception of resemblance. When reflecting on his own conscious experience, the

association of resemblance appears to be contained in the impressions themselves. Thus, he concludes that the causes of resemblance, like the causes of his impressions themselves, originate "in the soul originally, from unknown causes [Hume Treatise, 1.1.2.1]" and cannot be discovered through introspection.

Hume's account thus aligns with contemporary psychology research in that he claims that many of our similarity judgments happen pre-consciously and involve processes that cannot be reliably accessed through introspective reflection. Psychology studies have found that we lack introspective access to the causes of many of our similarity judgments [Gilovich 1982, Maier 1930].

However, modern cognitive scientists depart from Hume in using the scientific method rather than introspection to analyze the causes of our similarity judgments. Empirical studies allow scientists to study how similarity is constructed by the mind, and what heuristics underlie this process. Understanding how similarity judgments are formed allows inferences about the reliability of our sense of similarity in different contexts and how we might improve this reliability.

4.2.1 Similarity and the problem of induction

Hume's problem of induction is that we have no rational justification for our belief that the future will resemble the past. When we experience an A-type event, and in the past, A-type events were accompanied by B-type events, we expect that this present A-type event will also be accompanied by a B-type event. There is no inductive or deductive argument that we can give, however, that will justify our belief.

This is the standard problem of induction. Yet there is a second dimension to inductive inference that also requires examination and justification, which has not received as much

attention. When we reason inductively, we hold the belief that our method for identifying A-type events is reliable.

Consider an analogy. When we get a new computer, we make the inductive conjecture that clicking on a 'home' icon on this computer will take us back to the home screen as it has done previously on the old computer. This requires two assumptions. First, an assumption of uniformity; that the underlying algorithm of this new computer is analogous to that of the old computer. I assume that this new computer has an algorithm that brings up the home screen when the 'home' icon is identified and clicked. The alternative is possible; perhaps this new computer involves an AI holographic desktop helper such that the user can only talk to the AI helper and never has the option of seeing the full layout of different folders and applications.

The second, separate assumption is that my similarity judgment is accurate; that *if* there is a home screen programmed into the computer's internal algorithm, that I am able to correctly identify the icon that is similar to the old 'home' one, even though it is likely a different color and shape from the 'home' icon on my old computer. My ability to make the correct similarity judgment is not a given. For instance, sometimes fake icons are embedded in ads. An ad might show the image of a 'home' icon or a '35 unread emails' icon in order to trick an unsavvy user into clicking on it. Even if the underlying algorithm is uniform between the old and new computer, it is not a given that I will correctly judge how to access it.

When reasoning inductively, I assume both that (1) the underlying algorithm of this computer is structurally analogous to my old one and (2) that I am able to judge the similarity of the icons to make a prediction about which icon click will access the part of the algorithm that I want. The judgment of similarity is crucial machinery in the inductive inference. But what justifies this similarity judgment?

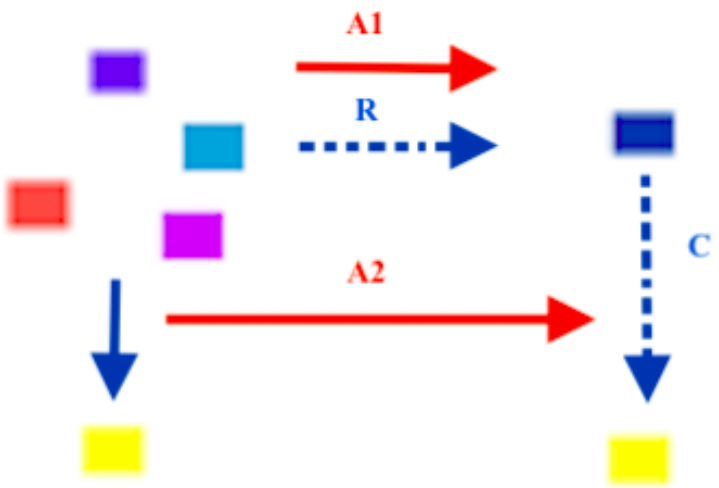


Figure 4.1: Two dimensions of induction

The below figure illustrates the two assumptions involved in inductive reasoning, which are labelled A1 and A2. A2 is the standard assumption of regularity that Hume discusses: when we encounter an event that is similar to events that led to B-type events in the past, we expect that future events that resemble those events will continue to yield B-type events. The other assumption, A1, is similarity. We assume that there is something about this present event that distinguishes it from other events and causes R, the feeling of resemblance. These two assumptions, similarity and regularity, generate our belief in C, the expectation that the new event will yield a B-type event. Both of these assumptions require justification.

4.2.2 Pragmatic responses to the problem of induction

Under Hume's account, resemblance is the foundation of habit and custom, and thus underlies inductive reasoning. Hume's skeptical solution to the problem of induction is that inductive reasoning is an innate component of our psychological nature, and we cannot help

but assent to it even though it is not rationally justified. Hume's account is primarily a descriptive account; with his account, we can describe how we come to hold a belief. There is a smidgen of normativity in Hume's skeptical solution, because we can provide an explanation for how we came to hold beliefs that were successful in the past. However, Hume's account certainly does not provide full-fledged normativity, because describing how we came to have a belief does not answer the question of whether we should hold that belief, unless more can be said about the reliability of that epistemic process.

Pragmatists have noted that the real motivation underlying our desire for rational justification is reliability–we want our inductive inferences to lead to successful action [Barrett 2024]. To a pragmatist, the problem of induction is really the problem of establishing and refining the reliability of the processes by which we form beliefs. The meaningful question to pragmaticists is not "is our inductive reasoning rationally justified?" but "How should knowledge of past events inform predictions about similar but unfamiliar events in the present?" Pragmatic responses to the problem of induction examine how custom and habit are formed and the reliability of this as a belief-forming mechanism. One area of focus has been on identifying cognitive shortcomings in our belief-updating processes and establishing epistemic norms for improving the reliability of our probability judgments in our inductive reasoning. For instance, Bayesian epistemology establishes a framework for rationally updating probabilistic beliefs. Consider an example.

Example 4.2.1. The following probability puzzle circulated the social media platform, X, after being posted by a University of Toronto mathematics professor [Litt 2024]:

> **You are given an urn containing 100 balls; n of them are red, and 100-n are green, where n is chosen uniformly at random in [0, 100]. You take a random ball out of the urn—it's red—and discard it. The next ball you pick (out of the 99 remaining) is:**
>
> **A. More likely to be red**

B. More likely to be green

C. Equally likely

D. Don't know/see results

22,563 votes were cast, and the final distribution of results was A: 22.6% , B: 37.1%, C: 20.9% and D: 19.5%.

The correct answer is A: 'More likely to be red' because the evidence obtained by drawing a red ball should update an agent's belief about the initial distribution of red to green balls in the urn. Suppose we want to use Bayes' theorem to calculate the conditional probability of the second ball being red.

Bayes theorem is:

$$\Pr(R_2|R_1) = \frac{\Pr(R_1|R_2)\Pr(R_2)}{\Pr(R_1)} \tag{4.1}$$

where R_2 is the probability of the second ball drawn being red and R_1 is the probability of the first ball drawn being red.

For the denominator, we need to have the prior probability of the first ball drawn being red. However, we cannot obtain this value for the particular urn that the ball was drawn from, because we do not known what the distribution of red and green balls was. Thus, we instead compute over all of the possible urn distributions. One can visualize that if we assume a set of urns with all possible initial distributions of red and green balls, and we draw a ball randomly from this set, the probability that the first ball drawn will be red is 1/2. Thus, $Pr(R_1)$ is 1/2.

For a given urn, the probability that if there are x red balls in the urn then both balls chosen

will be red is:[1]

$$\frac{1}{101} * \frac{x(x-1)}{100*99} \tag{4.2}$$

So, if we compute over all of the possible urns, we get the formula:

$$\sum_{x=0}^{100} \frac{x(x-1)}{101*100*99} \tag{4.3}$$

Combining the numerator and denominator of Bayes' theorem, we get:

$$2 * \sum_{x=0}^{100} \frac{x(x-1)}{101*100*99)} \tag{4.4}$$

This equals:

$$\frac{2}{101*100*99} * \sum_{x=0}^{100} x(x-1) \tag{4.5}$$

This reduces to:

$$\frac{2}{101*100*99} * \sum_{x=0}^{100} x^2 - \sum_{x=0}^{100} x \tag{4.6}$$

By the formula for the sum of squares and sum of first n integers, we get:

$$\frac{2}{101*100*99} * (\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2}) \tag{4.7}$$

[1]The fraction 1/101 represents the fact that there are 100 total balls in the urn, so there could be anywhere from 0 to 100 balls, which is 101 total possible values for x. $\frac{x(x-1)}{100*99}$ is the conditional probability that both drawn balls are red in the situation in which there are x red balls in the urn. For instance, if $x = 100$, then the probability of both drawn balls being red is 1.

This equals:

$$\frac{2}{101 * 100 * 99} * \frac{n(n+1)(n-1)}{3} \tag{4.8}$$

Plugging in 100 for n, we get the answer: $\Pr(R_2|R_1) = 2/3$ [Paterson 2024]. This is the conditional probability of the second ball drawn being red given the evidence of the first ball drawn being red. Someone who believes that the second ball is more likely to be green than red is vulnerable to having a Dutch book made against them. A Dutch book is a collection of bets that ensure a loss to someone whose probability beliefs are incoherent [Barrett 2024]. However, many people appear to rely on the simple heuristic that drawing and discarding a red ball reduces the number of red balls in the urn, and they base their reasoning on this heuristic without considering the information that the initial draw provides about the initial distribution of balls in the urn. This heuristic yields the incorrect belief that the next draw is more likely to be green than red.

There are numerous such cases where our inductive reasoning relies on simple rule-of-thumb heuristics that evidently worked well enough for survival in our evolutionary history, and saved us the energy and attention expenditure that more complex heuristics would require [Pinker 2005]. However, these fast and cheap heuristics, while adaptive in our evolutionary history, cause our inductive reasoning to lead us astray when solving puzzles that were not important for the survival of our evolutionary ancestors. Many of the inductive reasoning problems that we face in contemporary life are not ones that our evolved cognitive dispositions are adapted to solve. For instance, the importance of social relationships for our evolutionary ancestors' survival has resulted in us being more able to accurately solve logic puzzles when they are presented in a social context as opposed to a non-social context [Pinker 2005]. One example is the Wason selection task, which is a logic puzzle where participants are presented with a set of four cards and a conditional rule like "If a card has a vowel on

one side, then it has an even number on the other". Participants are asked which cards it is necessary to turn over to determine the validity of the rule. Over 75% of participants fail the task, but if the same logic puzzle is instead represented as the conditional "If a person is drinking beer, they must be over 21" and participants are shown a set of four cards representing people at a bar, with one side showing their drink and the other their age, then failure rates drop below 35% [Pinker 2005].

Additionally, it was more adaptive for our evolutionary ancestors' inductive inferences to be wrong in fitness-enhancing ways, than to be accurate. Suppose that while out gathering food, a swaying branch above you strikes you as being a mountain lion. However, it turns out that you are wrong and it is just a branch swaying in the wind. The disposition to be wrong in this case will be selected for more heavily than the disposition to be wrong in the opposite way: by seeing a mountain lion and mistaking it for a swaying branch. If our inductive reasoning were geared towards being accurate, then we would use a slower, more fine-grained heuristic for analyzing the shape in the tree, and would finally settle on the conclusion that it was a mountain lion after it was already mid-pounce. Our inductive reasoning has not evolved to give us accurate conclusions, but to give us conclusions that will optimize survival and reproduction. These cheap, adaptive heuristics sometimes generate deeply irrational conclusions when we reason inductively [Pinker 2005].

Pragmatic solutions to the problem of induction investigate ways to bolster the reliability of our inductive reasoning by providing epistemic norms that refine our reasoning. One such example is the use of the axioms of probability to prevent irrationality in our probabilistic reasoning.

Bayes' theorem is used in a variety of scientific contexts to refine scientists' inductive reasoning from particular instances to a general hypothesis, but using Bayes' theorem relies on assumptions about similarity. In Example 2.2, using Bayes' theorem requires that we assume that each draw from the urn is relevantly similar; that each draw is independent of the draw

that came before, that the draws are random; someone hasn't pre-arranged the distribution of the urn to have the red balls near the top ready to be picked first; that the number of balls in the urn stays constant and only reduces by one ball every time we draw and discard a ball.

These assumptions amount to an assumption that each draw is relevantly similar to those that came before such that it is suitable for updating the probability of the hypothesis. It is an implicit assumption of the probability problem expressed above that the urn is not rigged or manipulated. Assumptions about similarity are built into conditional probability questions. If a probability question asks what the probability of a six-sided dice landing on two is, the question is really asking "What is the probability of the six-sided dice landing on two, assuming that the dice is fair, not weighted?" The assumption of similarity may be fairly straightforward in the urn case, but when using Bayes' theorem to guide inductive inferences about natural phenomena, the assumption that individual draws on the urn are relevantly similar needs to be carefully justified. Consider the following example.

Example 4.2.2. Ecologists and conservationists use Bayes' theorem to estimate the abundance of a population in an ecosystem. This is crucial for making inferences about how endangered a species is and changing policy in response to population dips [Shertzer et al. 2022]. For instance, Bayesian analysis is used to estimate whether overfishing is occurring in a region and fluidly enact or repeal seasonal closures, annual catch limits, and restrictions on which fishing methods are allowed as fish populations ebb and wane.

To do this, ecologists often use a capture-recapture methodology, which involves settings traps, marking or tagging the animals that end up in a trap, and releasing them. Then, traps are reset and scientists record the proportion of animals from this second trapping that already have tags on them relative to the animals that are untagged [Shertzer et al. 2022]. This allows them to use Bayes' theorem to estimate the total population density.

In order to use Bayes' theorem, ecologists must postulate a conditional distribution over which the Bayesian analysis takes place. If ecologists think that the draws on the hypothetical urn of the traps are likely to be relevantly dissimilar to each other, such that one trapped animal has a different degree of relevance as evidence for the hypothesis, then this is not provided by the Bayesian analysis itself and has to be built in as a background assumption.

One consideration that has to be taken into account by ecologists using the capture-recapture method is that trapping an animal the first time can affect its likelihood of being trapped a second time. Animals can demonstrate trap-happy and trap-shy behavior, where being trapped once increases or decreases their odds of being trapped a second time. For instance, one study on vaccine efficacy in badger populations in Ireland was complicated by the fact that some badgers who experienced trapping learned to associate the trap with food and had an increased propensity to be trapped a second and third time [Byrne 2012]. If the organism being studied is trap-happy or trap-shy, then ecologists need to build this assumption into their model. However, this is not a simple task. In the badger study, some individuals in the badger population were trap-happy and others were trap-shy, and there was not a straightforward correlation to age, gender, or other quantifiable traits.

Using a capture-recapture technology also requires explicating an acceptable standard for category membership. One challenge that has affected past capture-recapture estimates is the existence of cryptic species. A cryptic species is one that is morphologically indistinguishable from a different species occupying the same ecosystem. If the population being studied is a cryptic species and scientists checking traps are likely to make inaccurate species identifications, then their reasoning must include an explicit condition for relevant similarity. Trapped animals must be identified by genetic testing, not mere perceptual similarity.

Finally, migration and behavioural patterns are another factor that affect the inductive strength of each urn draw in a natural set. The location of the traps, and the movement of a population of animals within their environment affect the propensity of a given individual

being trapped in a particular trap. A simple Bayesian analysis would merely assume that the population's behaviour and location is constant over time, but this assumption could be inaccurate. The population could be clustered near the sampling sites, or there could be a behavioural change in the population like a disease or a stage of the breeding cycle that changes the probability of a set of animals being caught [George 1990]. In these cases, ecologists develop a theory about the factors affecting the degree of similarity or difference between their samples, and then must build an extension onto the basic Bayesian model that takes into account the predicted representativeness of a given sample.

In summary, Bayesian analysis does refine our probability reasoning, but is dependent on background assumptions about relevant similarity. Whether explicitly stated or implicit, some assumption about similarity is embedded in every Bayesian analysis. Any inductive generalization relies both on how we use a sample to update our probabilities and an assumption about whether the sample is representative of the population or set that we are trying to reason about. Bayesian epistemology provides norms for the former, but the latter is an extension that must be built into the model. Bayesian epistemology provides epistemic norms for rationally updating our beliefs as new evidence comes in. Pragmatic solutions to the problem of induction must also include investigation of the reliability of our similarity judgments and epistemic norms for evaluating similarity.

4.2.3 Similarity is constructed and sometimes unreliable

A philosophical account of norms for similarity may seem unnecessary given that scientists empirically investigate and refine norms for similarity within particular scientific domains. These norms are local; a respect for similarity that is highly fruitful for one type of induction in one domain may be unproductive in another. Thus, a philosophical account of similarity may seem superfluous–what domain-general norms for similarity can be possible or useful?

However, there are a number of reasons that philosophical investigation of similarity is needed to help scientists decide in the trenches which similarity judgments are reliable. First, empirical investigation does not definitively determine one best respect for similarity. In Chap. 1, Example 2.1, I discussed scientific debates about the positioning of turtles in a phylogenetic tree. Phylogeneticists debate about whether morphological or genetic similarities are the more relevant similarity for determining turtles' ancestry [Sato 2023]. Even within these two approaches, there is disagreement about norms for relevant similarity. For instance, within the genetic analysis approach, some scientists advocate for prioritizing nucleotide sequences as a similarity criterion, whereas others advocate for using X-ray crystallography to obtain data about 3D protein structure. A benefit of the protein structure approach is that protein structures are thought to stay more constant over evolutionary time than nucleotide sequences [Koehl 2001]. On the other hand, there is considerable debate about what constitutes a similarity in protein structure. There is a plurality of norms for similarity of protein structure, and many of these yield different answers about whether two given proteins are structurally similar [Koehl 2001].

Even within a mature science like phylogenetics, there is often not scientific consensus about what constitutes a relevant similarity. In short, norms for similarity are not simply given by empirical investigation—scientists must choose them. In doing so, scientists consult the epistemic resources that they have available to them, which invariably include psychological, cultural, and paradigmatic norms, among others [Goldstone 1994, Kuhn 1962].

In particularly recalcitrant cases of disagreement about norms for conceptual similarity, scientists sometimes retreat to basic, pre-theoretic similarity judgments in order to determine which conceptual norm better satisfies their basic epistemic commitments [Stanford 2024]. I will give an example of this in sect. 3.3.

Scientists' pre-theoretic intuitions about similarity also hold a central role in analogical reasoning, where a source and target domain often belong to diverse fields such that there

are not established norms for generalizing between them. In these cases, scientists use creative or nonstandard respects for similarity [Holyoak 1994].

In summary, scientific similarity judgments are partially dependent on more innate, unscientific methods for fixing similarity, like psychological heuristics and cultural norms. As Quine says, "Science is not a substitute for common sense, but an extension of it [Quine 1957]." Thus, the existence of scientific norms for similarity does not eliminate the need to establish the reliability of the processes by which similarity judgments are fixed.

How are our basic, perceptual similarity judgments fixed? Under a traditional picture of cognition, our phenomenal experience provides a window through which we view approximations of similarities and differences existing in nature. Under this conventional picture, investigating the reliability of our similarity judgments means finding small blind spots or warped regions of the window. For instance, we lack the ability to perceive ultraviolet light and must use an optometer to measure similarities in ultraviolet light, and our similarity judgments can be muddled by optical illusions like the Ebbinghaus illusion. However, the psychology literature has found that this is an inaccurate depiction of the nature of our similarity judgments [Goldstone 1994, Hoffman 2019, Enquist et al. 2003]. Rather, even our most basic and fruitful similarity judgments are constructed by our cognition. The extreme sophistication and usefulness of our sense of similarity comes not from our cognitive faculties presenting us with a veridical representation of similarities that exist in the world, but rather from them constructing a phenomenal reality imbued with similarities and difference that maximize our likelihood of acting successfully.

Our visual system and other cognitive faculties continuously interpret raw data and construct a phenomenal reality imbued with similarities and differences [Hoffman 2019]. For instance, the incoming visual data that bounces off the retina is constantly changing from moment to moment. If a person moves their head a fraction of an inch, then every photosensitive cell on their retina is receiving different data than it was a moment before. Some of these

differences are due to their head moving, and some are due to changes in the coffee shop they are sitting in–a car has driven by outside, causing a flash of reflected light across half of their visual field, and a barista is approaching them with a slice of cake.

To protect them against this onslaught, the person's visual system presents them with a simplified story about similarity. It tells them that their coffee cup and the table are the same as they were a moment before, and presents these as stable even though the light bouncing off them careens across the person's visual field as their head tilts. It says that all of the objects in the coffee shop exist from one moment to the next even though they are momentarily a different color and brightness from the flash of light. By creating an illusion of similarity and stability, the visual system allows the person to attend to important changes in their visual field like the approach of their slice of cake. All of this construction of similarity happens below the level of conscious experience.

The predictive processing literature claims that our preconscious cognitive faculties make a guess about the content of a visual scene, and then filter out visual input that doesn't align with that visual scene in order to create a consistent conscious experience. Researchers in one psychology study "One might say that the higher visual areas ask of the primary visual cortex: 'This is what a disk in the foreground *would* look like–please check if this matches the input you're getting [Zhaoping 2021].'" If there are elements that do not match the presumed visual scene, then these contradictory visual inputs are not included in the visual experience that is presented by the visual system to the conscious experience of the agent. The disk looks similar to past disks that the agent has experienced because the brain *makes* it similar. If the incoming visual input diverges from what a disk is expected to look like based on past experience, then these anomalies are filtered out and never appear in the conscious experience of the viewer [Zhaoping 2021].

Our brain's endowment of the world with similarity and difference is a sophisticated adaptive mechanism. Cognitive scientists have used game-theoretic modeling to show that seeing the

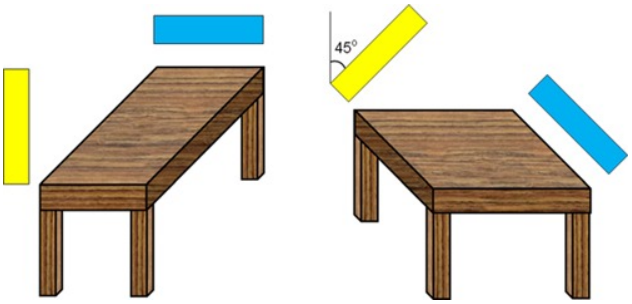


Figure 4.2: Shepard tables [1990]

world in a veridical manner rather than having similarity and difference preconsciously built into it is maladaptive [Hoffman 2015]. Agents who have veridical or partially veridical sense experiences survive and reproduce with less frequency than those who have a wholly constructive sense experience [Hoffman 2010].

Instead of presenting us with objective reality, our cognitive system uses heuristics, or rules-of-thumb, to create our similarity judgments. These bias us towards making certain judgments, often in line with what was adaptive in our evolutionary history. There are many example of these biases in the psychology literature. When presented with two stimuli, A and B, subjects demonstrate asymmetric similarity judgments where A is judged to be more similar to B than B is to A, or vice versa, depending on the order or frequency at which A and B are shown to subjects [Polk et al. 2002]. When judging the degree of similarity between a present situation and a previously experienced event, agents judge the similarity to be higher when the previous event was dangerous or inspired negative emotions [Kahneman et al. 2011]. As Hume notes, relevant similarity is built into our most basic impressions. Sometimes these biases, while adaptive in one context, cause unreliable judgments in other contexts.

The Shepard table illusion [1990], shown below, illustrates that even our simple perceptual similarity judgments can be unreliable. The two parallelograms forming the surfaces of the tables are equivalent in size and shape, but our brain cannot help but perceive them as being two different shapes. The context surrounding the quadrilaterals (the table legs)

alters the judgment of how similar the quadrilaterals are. A similar effect can be seen in the Muller-Lyer illusion [Judd 1899]. There are also uniformity illusions, in which the brain misrepresents shapes near the periphery of the visual field to make them more similar to shapes in the center of the visual field [Suarez-Pinilla 2018], or presents a changing visual scene as continuous and unchanging. Contemporary psychology studies illustrate numerous cases where similarity heuristics are unreliable in certain contexts [Goldstone 1994].

In short, similarity, even the seemingly straightforward perceptual similarities that we perceive between objects in day-to-day life, is a complex notion that requires philosophical scrutiny. This is especially important when considering the similarity judgments used in science, because the evolutionary selection pressures that shaped our innate similarity biases likely selected for different biases than the ones that would be ideal for scientific enquiry.

The remainder of this section will offer some preliminary discussion of how respects for similarity are fixed in practice.

In addition to being shaped by scientific theory (i.e. a dolphin is more perceptually similar to a shark that to a dog, but taxonomically more similar to a dog than a shark), our similarity judgments are also shaped by associative learning. A forager living in one environment might learn that morel mushrooms are edible, and that identical-looking false morels are poisonous. Agents in that environment will come to judge morels and false morels to be dissimilar, and those differences will be salient to them, even if perceptually it is challenging to distinguish between morels and false morels. Agents in that environment will experience the perceptual similarity between the morel and non-morel mushrooms, but this perceptual similarity will not cause a similar-category judgment. In a second environment, only morels exist, no false morels, so the perceptual similarity will cause a similar-category judgment.

Dispositions for similarity judgment biases can become genetically encoded. Over evolutionary time, if distinguishing between morels and non-morels is a reliably adaptive disposition,

then agents may evolve cognitive faculties for distinguishing between morels and non-morels, such that over time, their cognitive experience contains an increased number of dissimilarities between morels and non-morels. One example of this is in the evolution of predator visual processing in cases of prey mimicry. If a prey mimic is reliably an important food source, then a predator with a mutation that increases their ability to differentiate the mimic from genuine poisonous/venomous species will have increased fitness, resulting in changes in the perceptual capacities of that predator species over time. In turn, it will be adaptive for a prey species to have an appearance that exploits the predator's visual biases. A similar coevolutionary competition occurs in brood parasitism, where host species evolve mechanisms for differentiating between their own offspring and parasites, while the parasite species evolves increasing deceptions to make differentiation difficult for the host species [York 2021].

Learning can also affect basic perceptual similarity. In one study, it was found that people who formerly worked as judges at dog shows are able to differentiate between individual dogs belonging to the breed of dog they are experts in, by recognizing subtle similarities and differences that are not observed by non-experts [Diamond et al. 1986]. Psychology experiments have found that the source of experts' changed similarity judgments can be both (1) changes in attention allocation (they learn to attend to different features, and in more detail, than non-experts)[Harel et al. 2013], and (2) changes in the visual system's algorithms for preconsciously interpreting incoming visual data, even when experts and non-experts are attending to the same features [Goldstone 1994].

Studies in cognitive science suggest that judgments of perceptual similarity and judgments of similarity based on learned category membership are interdependent. Learning that two objects belong to the same category can alter test participants' visual perception of the objects [Goldstone 1994, Gauthier et al. 2002]. Additionally, through individual learning or imitation, individuals learn which perceptual features are salient for successful action and develop attentional biases for those features.

The predictive processing literature, which spans philosophy and cognitive science, posits that basic perceptions are hypotheses generated by the brain on the basis of past knowledge and incoming sensory data. Under this account, our sense of perceptual similarity is shaped by past experience. After riding a bike over a pothole, when we see another pothole we perceptually experience the new pothole as being similar to the old one, even though the properties that we perceive cannot be conveyed by vision alone. Our perception of the depth of the pothole is not fixed by our visual input, yet we experience depth. We can perceive the hardness of the asphalt and the coldness of the water in the pothole even though our visual input does not actually convey these properties; they are added by our cognition.

Hume notes that our perceptual experience of objects has more properties than the incoming sense data contains, "Tis a common observation, that the mind has a great propensity to spread itself on external objects, and to conjoin with them any internal impressions, which they occasion [Hume Treatise, 1.3.14.1]". Our minds create a perception of causality, similarity, and properties that are not immediately given by the sensory input we have immediate access to.

Just as learning shapes perception of similarity, innate perceptual biases place constraints on our category learning. For instance, a predator organism that has not evolved the visual apparatus that makes the differences between a poisonous snake and a harmless mimic species visually salient will avoid both species of snake and will not learn the slight differences that differentiate them.

One final important characteristic of similarity judgments is that scientists often lack epistemic access to the reasoning underlying their similarity judgments [Gilovich 1981, Maier 1931, Enquist 2003]. When reasoning analogically, scientists' analogical conjectures can be affected by superficial or irrelevant factors, and scientists are often unaware that those factors influence their similarity judgment [Gilovich 1981]. Consider an example.

Example 4.2.3. Gilovich [1981] surveyed a sample of political science undergraduate students to see how implicit, unconscious similarity judgments might influence their belief about the best method to handle a foreign policy crisis. The study found that (1) test participants were significantly influenced by superficial similarities between the target scenario and historical analogues and (2) they were unaware that these factors influenced their similarity judgment. The study presented three groups of participants with a hypothetical foreign policy crisis and asked them to recommend an appropriate US intervention. The hypothetical crisis was identical in all three conditions, except for a few superficial details. In one condition, superficial details of the scenario were designed to trigger an association between the hypothetical scenario and WWII. The president's home state is described as being New York, which is also the home state of FDR (who was the president during most of WWII). In that condition, victims of the foreign policy crisis were also described as fleeing the country in freight trains. In the second condition, the president is described as being born in Texas, which is the birthplace of Lyndon B. Johnson, the president during the Vietnam War. In this second condition, the victims of the crisis are described as fleeing the country in small boats. The third condition is neutral and specifies a presidential birthplace and means of victims fleeing the country that are not associated with either WWII or the Vietnam War. The study found that participants in the condition with superficial similarities to Vietnam recommended a more hands-off intervention strategy, while participants in the WWII condition recommended stronger interventions [Gilovich 1981]. After the study, participants were debriefed and asked whether the details about presidential birthplace and transportation method for victims fleeing the crisis may have had an effect on their recommended intervention. Most participants claimed that they noticed the parallels to historical analogues and that such details could hypothetically affect a person's recommendation, but reported that the details had not had any effect on their own policy recommendation [Gilovich 1981].

This highlights the fact that a normative account of induction should not rely on the first-

person authority of scientists, but rather should empirically identify contexts in which scientists' similarity judgments might go astray, and provide epistemic norms that constructively mediate possible sources of bias in such contexts.

In the next section, I will examine whether current accounts of analogical inference meet this need. Finding that they do not, I will then examine whether resources for answering this question can be found in the broader philosophy and cognitive science literature on inductive reasoning. Again finding that the existing literature is inadequate, I propose a positive account for how epistemic norms for similarity should be incorporated into normative accounts of analogical inference.

4.3 Current accounts of analogical inference lack norms for similarity

Guidelines for relevant similarity are particularly important for cases of inductive reasoning that are analogical. Generally, philosophers and cognitive scientists treat inductive inferences as analogical if they rely on similarities that are less entrenched and more abstract.

Analogical reasoning provides a mechanism to import successful strategies of action into unfamiliar environments. Consider the following example of analogical reasoning in crows [Smirnova 2015].

Example 4.3.1. Researchers trained crows with positive reinforcement to select matching pairs of objects. The crows were shown three upside-down cups, each with a picture on top. The middle cup always matched one of the cups on the right or left side of it, and food was always located under the matching cup on the right or left. The cup that did not match the middle cup did not have food under it. Crows learned that when they chose the cup that matched the middle cup, they would get a food reward. This is a case of straightforward

Figure 4.3: Training phase

Figure 4.4: Experimental phase

associative learning from perceptual similarity.

Next, experimenters set up a new scenario in which the cups had pictures that the crows had not been trained on. There were again three cups, but this time the center cup did not exactly match either of the cups on the left or right, but instead had a picture which was structurally analogous to a picture on the right or left cup. For example, the middle cup might show two crosses of different colors, and the cup on the left might show a triangle and square of different colors while the cup on the right shows two circles of different colors.

Experimenters wanted to see if crows would generalize the similarity-matching behaviour from the trained identity-matching case to the untrained relation-matching case. For instance, if the middle card displayed a circle and a cross, then the relation-matching choice would be to choose a card containing a square and a triangle rather than a card containing two squares. This inductive reasoning is more analogical in character; in order to choose the correct cup, the crows must extrapolate an abstract rule for similarity [Smirnova 2015].

Though only trained on the identity-matching cups during the training phase, the crows correctly chose the relationally matching cups in the experimental phase with high degree of accuracy (77.7%), despite never having been trained to associate the relationally similar pairs [Smirnova 2015].

When learning to select the matching pair during the trial run, the crows appear to have developed not merely an association between the matching pairs that they were trained on, but

an abstract rule for similarity. This rule was general enough that the crows easily extrapolated the rule to a new environment in which the salient similarity was not simple perceptual similarity, but abstract relational similarity. To do this, the crows used a sophisticated and abstract notion of similarity to generalize the learned behaviour from the unfamiliar to the new domain.

Being able to successfully extrapolate rules for similarity is a key component of analogical reasoning. Normative accounts of analogical inference aim to justify and refine the analogical inferences used in science [Bartha 2009]. Given that similarity judgments are the core of analogical inference, normative accounts should provide epistemic norms that justify and refine our similarity judgments.

Instead, both philosophers and cognitive scientists treat similarity as a primitive hinge of experience [Gentna 2020]. Some recent accounts, like Norton's material account, have tried to dispense with the notion of similarity to avoid the slippery issue of defining and justifying it. In this section, I will recount the sparse discourse on similarity that does exist and show that it is inadequate.

4.3.1 Carnap's framework in the *Aufbau*

In the *Aufbau*, the notion of similarity is crucial for Carnap's project. Carnap's account echoes Hume's in that similarity is the fundamental relation that constitutes and builds projectible properties. 'Recollection of similarity' is a relation that holds between remembered experiences and present experiences, and forms the foundation of Carnap's account. However, Carnap does not explicate the similarity judgments that underlie his account or justify their reliability. This was one of the main critiques that Goodman levels against Carnap's project.

It seems that Carnap assumes a foundation for similarity of the type proposed by Shepard's [1987] geometric model, where an objective set of respects for similarity underlie our similarity judgments, such that similarity judgments are justified by their alignment with the model. I will discuss the problem with this assumption of the standard geometric account in section 5.

4.3.2 Hesse's two-dimensional account

Mary Hesse's account introduces a distinction between two types of horizontal similarity relations, material similarities and formal similarities. A material similarity is a pre-theoretic similarity that can be observed independently of the existence of any established theory about the source and target domain. Material similarities are features or processes that "appear to be alike to fairly superficial observation [Hesse 1969, pg. 11]." For example, Hesse claims that similarities between sound waves and water ripples can be observed prior to any theory about them because they both share observable characteristics like oscillation, reflection, refraction, and frequency [Hesse 1969]. In contrast, a formal analogy occurs when there is some mapping between the formal models of the source and target domain. For instance, the mathematical laws describing heat and fluid flow share a similar mathematical structure [Bartha 2015].

Under Hesse's account, in order for a similarity to be reliable grounds for inductive inference, the similarity must be a pre-theoretic material similarity. Hesse claims that theory-laden structural similarities are not reliable grounds for induction because formal similarities can be artificially created through manipulations of the source and target domain like Goodman's grue/green puzzle or Bartha's examples of specious resemblance in formal analogies in mathematics [Bartha 2009, Nappo 2022]. Hesse thus stipulates the material condition: in order to be plausible, an analogical inference must have pre-theoretic, observable similarities

between the source and target domain. The material condition constitutes the entirety of Hesse's normative criteria for an analogical inference's horizontal similarity relations.

There are several factors that make the material condition inadequate for refining the similarity judgments used in analogical inference. The first problem is that excluding formal analogies does not exclude gerrymandering of the source domain. In the prior chapter, I describe two influential material analogies that involve gerrymandering of the source and target domain [Lauman-Lairson, forthcoming]. These are cases where Hesse's material condition does not eliminate cases of gerrymandered similarity. Having a criterion that prioritizes material analogies over formal analogies does prevent some implausible analogies from being treated as plausible, just by virtue of some formal analogies involving a manipulated source and target domain. However, the material condition also counts as illegitimate a number of fruitful formal analogical inferences that we should count as justified [Bartha 2009].

Another problem is that our observational similarity judgments are the very class of judgments that are in need of philosophical investigation and refinement. As Quine notes [1969], many perceptually salient, pre-theoretic similarities are irrelevant and unreliable for scientific induction. In contrast, it is often the case that the types of similarities that are fruitful for inductive reasoning in science only exist in the context of a theory.

Consider the example of iodine and fluorine, which occupy the same group in the periodic table but are perceptually dissimilar. At standard temperature, fluorine is a yellow gas which is highly toxic. In contrast, at standard temperatures iodine is a blue-black solid. However, they are similar in that they both have seven valence electrons. Their similar electron configuration causes them to have similar chemical behaviour. Though atomic number is an abstract, theory-laden similarity, it is a more fruitful source for inductive inferences about iodine and fluorine's chemical properties than pre-theoretic perceptual similarity.

A similar objection to Hesse's material condition is raised by Bartha [2009]. Bartha demon-

strates that analogies that are founded merely on formal, structural similarity are used fruitfully by scientists and argues that Hesse's criterion is too restrictive. The material condition labels as implausible a class of analogical inferences that are fruitfully used by scientists, like formal analogies between heat and electricity [Nappo 2021]. Abstract analogies of this sort, though fruitful in scientific reasoning, violate the material condition [Hesse 1966, pg. 63].

There is also another sense in which the similarity judgments used in science are theory-laden. Even in cases where do utilize pre-theoretic similarities as the grounds for inductive reasoning, we use scientific theories and instruments to refine and correct for errors in our pre-theoretic similarity judgments. Research in cognitive science shows that perceptual similarities are constructed by our cognition, and sometimes this construction can be unreliable [Hoffman 2015]. To correct for this, scientists subject their similarity judgments to empirical investigation and use this to interpret the pre-theoretic similarities they observe.

Consider the example of color. Hesse would say that Object A and Object B share a pre-theoretic, material similarity if they both are yellow. However, this similarity says more about the cognitive architecture of the observer than it does about the structure of objects A and B. In humans, the phenomenological perception of the color yellow can be achieved through several different input stimuli. First, it can be achieved by the visual input of light on the 'yellow' segment of the visible light spectrum, which a wavelength of 570nm to 590 nm. Alternatively, the same visual experience can be achieved by a mixture of equal amount of red and green light wavelength light–such as a combination of 520nm and 650nm light. This is because the visual perception of color involves an interpretive process rather than a simple veridical representation of light bouncing off the retina. The visual system has three types of cone cells, short-wavelength cones, medium-wavelength cones, and long-wavelength cones. To generate perception of red or green, the brain compares the amount of input of the long-wavelength (red) and medium-wavelength (green) cones. If the long-wavelength cones are more excited than the medium-wavelength cones, then the visual system perceives red.

If the reverse, then it perceives green. Either an equal proportion of long-wavelength and medium wavelength light, or light of an intermediate wavelength between long and medium, is interpreted as yellow light by the visual system.

As a result, there is not a direct correspondence between observational color similarity judgments and the actual spectral properties of objects. To make this even more tricky, not all color perception is vulnerable to construction to quite the same degree. For instance, blue is only perceived when blue light stimulates the short-wavelength cones and cannot be generated through any combination of stimulation of the green or red cones. So for blue, there is a more direct one-to-one correspondence between visual input and perceptual output. The similarity judgment that two objects share the pre-theoretic material similarity of being blue is a more reliable similarity judgment in tracking actual properties of the objects than the material similarity judgment that two objects are the same shade of yellow or purple.

In this case, in order to establish reliable similarity judgments about objects' reflectance properties, scientists must necessarily rely on a wave theory of light and scientific instruments like a spectrophotometer. Our formal scientific theories are crucial for differentiating between cases of observational similarity that do map more closely onto the world, like two objects appearing blue, versus observational similarities that map onto the world less closely, like the observation that two objects appear yellow. If we want to use similarity judgments to theorize about the spectral absorption properties of an object, relying on our pre-theoretic perception of color would be less reliable than relying on formal similarities like numerical measurements of wavelength derived from scientific instruments, a wave theory of light, and scientific theories about the efficacy of our measuring apparatus, etc. Formal theories refine our similarity judgments. Hesse's material condition is misguided and counterproductive because it overlooks the important role that theory plays in bolstering the reliability of pre-theoretic similarity judgments.

Another challenge is that some of the examples of similarities that Hesse claims are pre-

theoretic are actually cases where the similarity judgment is highly dependent on the existence of a theory. For instance, Hesse claims that similarities between pitch and color are pre-theoretic. Norton [2021] points out that

> "The analogy between pitch and color can be implemented only if we have numerical measures of pitch and color. Since these measures depend on a wave theory for both, are they still pre-theoretic? Since they are inferred from measurements, are they observables [Norton 2021]?"

Hesse anticipates this possible objection [1966], and responds by claiming that for pitch and color, "the analogy is found in the language before any wave theory is thought of. It is independent of the particular theory of light we are considering and so can be used to develop this theory [Hesse 1966, pg. 31]." Hesse suggests that pre-theoretic similarities can be generated by characteristics of language, and thus observations of similarity are "relative to the assumptions of a given language community [Hesse 1966]."

While it is highly plausible that cultural norms and language *do* play a role in determining our similarity judgments, some justification is needed for the claim that cultural and linguistics norms are a reliable means of fixing scientific respects for similarity. Hesse does not provide justification for this claim. However, one possible answer could be that the scientific predicates that are reliably successful eventually become ingrained in natural language. But if so, why exclude theoretic similarities in the first place? Natural language often lags behind scientific theory and contains artifacts of earlier, defunct theories. For instance, a remnant of Newtonian mechanics is that gravity is often described as a force in natural language even though gravity is a curvature of spacetime, not a force, under general relativity.

Finally, because Hesse's account does not provide rules for differentiating theoretic from pre-theoretic similarities, her account seems to rely on scientists introspectively accessing the source of their similarity judgments to determine whether a similarity is pre-theoretic

or theory-laden. Psychology studies suggest that scientists often lack this epistemic access [Gilovich 1981].

Thus, Hesse's material condition does not constitute a successful strategy for establishing the reliability of the similarity judgments that analogical reasoning hinges on. Her account attempts to reduce the difficult question, "What makes a source and target similar?" to an easier question, "What properties of the source should be similar to properties of the target in order for the source and target to be similar?" Her answer to this latter question is that in order for the source and target to be genuinely similar, they should share pre-theoretic, material similarities. Her account then attempts to provide a once-and-for all justification of the reliability of material similarities. Even if taken as a starting point and not a fully-fledged account of similarity, the starting point that she proposes appears to be leading in the wrong direction.

4.3.3 Bartha's articulation model

Bartha's articulation model is a refinement and extension of Hesse's account.

Bartha's account purports to answer the question, "When is the source domain (S) relevantly similar to the target domain (T), such that a further feature from S can be generalized to T?" However, his answer reduces this original question to an easier, and slightly tangential question in the same way that Hesse's account does. His account says that S is relevantly similar to T when T contains features that are similar to the crucial features of the causal relationship in S. This answer makes a claim about *which* features of the source domain should be similar to features of the target domain in order for the source and target domains to be relevantly similar, but leaves unresolved the question of what constitutes this similarity.

The distinction between the original question and the question that Bartha's account answers

is summarized as follows:

Question A (The original question): What makes a source domain relevantly similar to a target domain?

Question B (The question that Bartha answers): Which features of a source domain need to be similar to features of a target domain in order for the source and target to be relevantly similar?

The remaining component of Question A that Bartha does not answer: What makes these features of the source domain relevantly similar to features of the target domain?

How does Bartha justify purporting to answer Question A by only answering Question B? Bartha justifies this by claiming that the notion of material similarity commonly used in science is unproblematic and self-evident. He says that in domains outside of science, such as in jurisprudence, the question of whether a similarity exists might be controversial and debated. For instance, when using analogical reasoning from a past court case to motivate a judgment on a negligence suit, lawyers will debate about whether the duty of care of the defendant in the present case is analogous to that of the guilty party in the former case. There is disagreement about what constitutes a relevant similarity.

Bartha claims that the same disagreement does not often occur in science. His rationale for this claim is that scientists generally have a shared agreement about how S and T are similar and dissimilar prior to any analogical inference being proposed. Thus, he leaves the question of what constitutes an acceptable similarity up to the discretion of scientists, saying that analogical "inference cites *accepted* similarities between two objects or systems [Bartha 2020]. Bartha says: "We seldom see analogical arguments that depend upon similarities that must themselves be established and justified analogically. Rather, analogical arguments in science depend upon basic similarities that are supposed to be unproblematic [Bartha 2009]."

Bartha does not provide further justification for this claim, but his answer may be motivated by Quine. Quine claims that our intuitive similarity judgments can be unreliable, but that the maturation of a science involves this primitive sense of similarity being replaced by more reliable scientific standards. For example, a primitive notion of water solubility is replaced by a precise articulation of the molecular structure that leads to a molecule being soluble. Quine says,

> One can redefine water-solubility by simply describing the structural conditions of that mechanism...That is, once we can legitimize a disposition term by defining the relevant similarity standard, we are apt to know the mechanism of the disposition, and so by-pass the similarity [Quine 1969, pg. 21].

A dogmatic reading of Quine could take this to mean that mature branches of science do not rely on similarity judgments, and instead have uncovered underlying mechanisms that can be uncontroversially imported into analogical inferences to determine similarity between a source and target. Under this picture, scientists do not need to justify their choice of a similarity predicate or choose between competing notions of similarity.

A more charitable reading of Quine's view would be to take Quine to be claiming that in mature domains of science, intuitive norms for similarity are replaced by more precise norms that are empirically refined and entrenched in the scientific theory. It does not follow that the similarity norms used in inductive reasoning are self-evident and uncontroversial. Even in a given domain, scientists must decide which norms are most applicable for a given induction and decide how to weigh those norms. Bartha's account does seem to assume a view of similarity in which relevant similarity in mature domains of science is fixed, uncontroversial, and reliable. This view has several challenges.

The first challenge is that there is often a plurality of norms for similarity within a domain. Even in purportedly straightforward cases like Quine's example of water solubility, scien-

tists still debate about what mechanism underlies a similarity judgment. For instance, the conditions for water solubility are a source of scientific disagreement.

Example 4.3.2. Water solubility is a key factor in bioavailability, so is of key interest in pharmacology [Llompart 2024]. But despite being a longtime subject of extensive research, the numerous, complex mechanisms underlying water solubility are still a source of scientific disagreement, and scientists lack a reliable method for predicting water solubility in new drugs. One chemist says,

> The truth is that solubility is a word that describes a large range of phenomena, from classic "yes/no" solubility of crystalline solids to phenomena such as spinodals (e.g. where a polymer is scarcely soluble in a solvent but the solvent is quite soluble in the polymer) and to the world of dispersions/colloids/nanoparticles where it is hard to decide what is "soluble" and what is "dispersed"...The world of solubility science is fragmented into little domains, each of which has its own language and priorities...[Chemists] adopt two extreme positions... The first way is to rely on intuition and experience, "how it's always been done". The second is to dig deep into profound theory in the hope that the answers will be found there [Abbott 2017].

Scientific standards for similarity have become more controversial as more precise, theoretic definitions of solubility are adopted. When observing two molecules, X and Y, some scientists agree that they share the property of solubility if solubility means a vague, pre-theoretic prediction, like "If X and Y were both placed in water they would dissolve." However, chemists' disagreement about an appropriate predicate increases as more precise mechanistic accounts with more embedded theoretical commitments are considered [Abbott 2017]. Abbott says,

> Someone who works on "dispersions" might never want to use Hansen Solubility Parameters, because the word dispersion instinctively creates the idea of

> something that is not soluble. Yet the evidence is overwhelming that HSP are hugely useful for solvent-based dispersions. Similarly, someone might be happy to use Fluctuation Theory of Solutions (KB) for solutions of proteins, starches or DNA, but would not choose to use it for dispersions of nanoparticles. Yet I cannot find any fundamental difference between solutions and dispersions, and the Gibbs Phase Rule approach of Shimizu demonstrates that there is no difference [Abbott 2017].

Rather than scientific predicates enjoying more secure epistemic footing than pre-theoretic predicates, scientifically precise expressions of similarity can sometimes be more contentious than their pre-theoretic counterparts. Scientists might agree that the proposition "X and Y are both soluble" is true when "is soluble" is a vague, intuitive predicate, but disagree when "is soluble" is replaced by any more scientifically precise notion about the mechanisms underlying the solubility.

A second, related, issue is that similarity judgments are an essential component of the process by which scientific predicates are formed and refined. Many scientific predicates are defined through ostension, where instances of a scientific predicate are linked by a network of similarity judgments, or family resemblances. Consider the example of the scientific predicate 'species'.

Darwin defines the term 'species' as follows: "I mean by species, those collections of individuals, which have commonly been so designated by naturalists [Darwin 1857]." Darwin seems to see the term 'species' more as a useful heuristic used by naturalists rather than a natural kind. In the *Origin of Species*, Darwin says, "No one definition has satisfied all naturalists; yet every naturalist knows vaguely what he means when he speaks of a species [Darwin 1859, pg. 38]."

Rather than defining conditions that separate species from varieties, Darwin defines species

by giving paradigm examples of species where the populations are reproductively isolated, have stable, distinct morphologies, and occupy different environments [Darwin 1859].

For instances that diverge from the paradigmatic cases, scientists have to decide how to weigh similarities and differences from the paradigm case to decide whether a population is a variety or a species. For instance, Darwin discusses the puzzle of primrose and cowslip, which were thought by many naturalists to be distinct species because they have distinct morphologies and usually grow in different environments (primroses in shaded areas and cowslip in fields), yet violated expectations by producing viable hybrids. Darwin says,

> Many of the cases of strongly-marked varieties or doubtful species well deserve consideration; for several interesting lines of argument, from geographical distribution, analogical variation, hybridism, etc., have been brought to bear on the attempt to determine their rank. I will here give only a single instance,—the well-known one of the primrose and cowslip, or Primula veris and elatior. These plants differ considerably in appearance; they have a different flavour and emit a different odour; they flower at slightly different periods; they grow in somewhat different stations; they ascend mountains to different heights; they have different geographical ranges; and lastly, according to very numerous experiments made during several years by that most careful observer Gartner, they can be crossed only with much difficulty. We could hardly wish for better evidence of the two forms being specifically distinct. On the other hand, they are united by many intermediate links, and it is very doubtful whether these links are hybrids; and there is, as it seems to me, an overwhelming amount of experimental evidence, showing that they descend from common parents, and consequently must be ranked as varieties [Darwin 1859, pg. 49].

When deciding whether a population is a species or a variety, scientists weigh similarities

and differences from the set of accepted exemplars. Similarity judgments go all the way to the core of the process, because similarity judgments are used to formulate the criteria for what makes a case relevantly similar or different to the exemplar case.

Scientists use existing norms for similarity to decide whether to include an instance as a case of 'variety' or 'species', and the inclusion or exclusion of that instance from the category shapes the scientifically sanctioned norms for similarity. For instance, in Darwin's time, shared morphology was weighed heavily in species identification. As a result, it was widely accepted that cowslip and primrose were distinct species. Over time, more populations with distinct morphologies and environmental niches were found to be capable of interbreeding, and the definition of 'species' came to depend more on reproductive isolation than morphology. In contemporary research, the concept of species has shifted to hinge more on phylogeny than reproductive isolation [Mace 2004].

Similarity judgments are thus a constitutive part of the process of constructing scientific kinds and norms for similarity. Thus, normative accounts of analogical inference should investigate and refine the reliability of the similarity judgment process.

Bartha's second justification for leaving criteria for relevant similarity out of his normative account of analogical inference is that he claims that respects for similarity are fixed externally and temporally prior to the analogical inference in which the similarities are cited. He says that similarities are established prior to the analogical argument and so philosophical investigation of similarity falls outside the scope of accounts of analogical inference [Bartha 2009]. Contrary to this claim, I showed in Chapter 3 that scientists' credence about the similarity of the source and target domain are often fixed in the context of their broader theoretical framework, which includes intuitions about the plausibility of the analogical conjecture. Psychology studies further support this claim, suggesting that similarity judgment and analogical inference are interdependent. Medin et al. [1993] say,

> It is natural to assume that, to constrain similarity comparison appropriately, the representation of each of the constituent terms must be rigid (i.e., context-insensitive). In contrast, our observations suggest that the effective representation of the constituents are determined in the context of the comparison, not prior to it [Medin et al. 1993.]

The process of analogically comparing a source and target domain is a constitutive part of the process by which agents fix respects for similarity. This further establishes the need for similarity judgments to be examined by normative accounts of analogical inference.

4.3.4 Gentner's structure-mapping account

In contrast to Bartha and Hesse's accounts, cognitive science accounts of analogical inference focus their analysis on the distinction between structural and attributional similarities. First, Gentner proposes a descriptive distinction, noting that some psychology studies show that shared relational structure appears to be more salient to agents than attributional similarities:

> "When asked to rate inferential soundness (described as "the degree to which an assertion that is true for one situation would hold in the other"), the same subjects rated matches sharing relational structure as both more sound and more similar than those sharing object attributes [Gentner et al. 1990]."

In her structure mapping account of analogical inference, Gentner expands this descriptive distinction into a normative distinction. She proposes the *systematicity principle*—that an analogical inference is more plausible if the similarities between the source and target domain are structural similarities, particularly higher-order structural similarities. For instance, in the analogy between mice and humans in testing the effects of a drug, the relevant similarity

for inductive inference would be physiological structure and biological pathways in mice and humans. Gentner's preference for structural similarity cuts crossways to Hesse's criterion for pre-theoretic similarities. Some structural similarities under Gentner's account would be material similarities in Hesse's sense because they are similarities in the observable causal and/or physical structure of the source and target domain. For instance, Hesse's example of pitch and color could be represented in terms of their structural relations in a way that would satisfy Gentner's systematicity requirement. However, there are also structural similarities under Gentner's account that would not be observable similarities under Hesse's account. For instance, in the analogy between the atom and the solar system, the important similarities under Gentner's account would be structural similarities like the planets orbiting the sun and electrons orbiting the nucleus, not attributes like the distance between the electrons and the nucleus versus distance between the sun and planets, or the similarity in shape of the sun and the nucleus.

Gentner's rationale for preferring structural similarities is that structural relations are more likely to have causal efficacy in causing the further feature of the source domain, whereas attributional features are often superficial and not linked to a deeper causal relationship. However, under Hesse's account the orbit of an electron around a nucleus would be a theoretical construct rather than a pre-theoretic observable phenomena. So Hesse and Gentner are in disagreement about some of their criteria for relevant similarity.

There are a few challenges with Gentner's systematicity principle, some of which I already discussed in Chap. 1. First, Gentner's distinction between relations and attributes is ill-defined; attributes can be expressed as relations and relations can be expressed as attributes. Second, feature-based similarities play an important role in analogical inferences, with attributional similarities playing a key role in analogies in archaeology, for instance [Bartha 2015]. Third, Gentner's account only allows syntactic, not semantic, information to be compared when judging the similarity of the source and target. A highly dissimilar source and

target domain can be represented with a syntactic representation that allows them to come out as similar under Gentner's account. Having only syntactic criteria for similarity readily permits uninformative, manipulated similarity predicates like 'grue' [Nappo 2019]. Finally, Gentner's account, like the accounts discussed above, sidesteps the question of what similarity *is*, giving criteria for which features (or in this case, relations) of the source domain need to be similar to target domain relations without defining what it actually means for them to be similar.

In a 1993 paper, Gentner more explicitly addresses the question of what similarity is and whether it has meaning as an explanatory construct. She responds to Goodman's influential 1972 critique of the notion of similarity. Goodman's paper questions the meaningfulness of the notion of similarity, noting that it only has meaning when the actual respect in which two entities are similar is explicated in terms of specific shared properties. For instance, if one hears that chicory is similar to coffee, the proposition only has meaning when the respect in which they are similar is fixed: i.e. are they similar in being grown in the same region, in the etymological origin of their names, in price, in flavor, or in caffeine content? The statement that they are similar provides no information without some context that fixes the respect in which they are similar. Thus, Goodman reproves philosophers' use of similarity as an explanatory construct, saying that similarity is "insidious, an imposter, a pretender, a quack [Goodman 1972]." Any entity is both similar and different to any other entity in a multitude of ways [Goodman 1973, Pierce, Putnam].

Gentner summarizes Goodman's concern: "For similarity to be a useful construct, one must be able to specify the ways or respects in which two things are similar [Gentner 1993]." In response to this, Gentner at al. [1993] aim to redeem the notion of similarity by showing that in practice, the meaning of similarity often is fixed by context. In practice, the context in which a similarity is expressed and shared assumptions about similarity fix the meaning of a statement of similarity.

Gentner argues that a process of relational alignment between the source and target domain is one way in which agents fix respects of similarity. For instance, when test participants are shown a picture in which one object of the scene has different attributes in the source and target, but similar relational structure to the rest of the scene, participants who are first asked to identify similarities between the scenes before being asked to map features of the source to target domain are more likely to map between relational features than participants who are not asked to identify similarities prior to be asked about mapping. In summary, thinking about similarity seems to predispose people to think about relations. Gentner uses this to argue that the respect in which two entities are similar is fixed in part relative to the structural connections between the two entities.

Gentner's 1993 paper fails to meet the normative goal I set out at the beginning of this chapter. Her account of respects for fixing similarity is descriptive, rather than normative. The [1993] account says that humans have tendencies to fix respects for similarity in particular ways, not whether we ought to fix similarity in those ways.

Holyoak and Thagard's connectionist accounts likewise describe psychological findings about how humans judge similarity in analogical inference, but do not provide justification for the reliability of similarity judgments or epistemic norms for refining them.

4.3.5 Norton's material account

Norton echoes the concern that I have expressed in this section: that current accounts of analogical inference hinge on the notion of similarity yet fail to adequately define the notion of similarity and justify its reliability. Of Hesse's account, Norton says, "Hesse struggles to explicate in general terms even the simple notion of similarity that constitutes the horizontal relations." Norton sees this as not being a symptom of neglect on Hesse's part, but rather a symptom of the impossibility of generating a universal notion of similarity. Instead, Norton

proposes that similarity is not a meaningful notion as similarity is context-dependent and that the expression of similarity is really an imprecise shorthand for the expression of a more specific empirical fact.

Norton aims to address this by formulating an account of analogical inference that dispenses with the notion of similarity. Norton says,

> "There are two notions in the material analysis. The first is the fact of an analogy, or just fact of analogy. This is a factual state of affairs that arises when two systems' properties are similar, with the exact mode of correspondence expressed as part of the fact. The fact is a local matter, differing from case to case. There is no universal, factual "principle of the uniformity of nature" that powers all inductive inference. Correspondingly, there is no universal, factual "principle of similarity" that powers analogical inference by asserting that things that share some properties must share others. The fact of an analogy will require no general, abstract theory of similarity; it will simply be some fact that embraces both systems [Norton 2021]."

Norton recognizes that the notion of similarity is ill-defined in the analogical inference literature, and he aims to sidestep the issue of defining and justifying similarity by replacing the vague notion of similarity with the 'fact of analogy', an inductive statement about how the two domains are similar which, if true, makes the analogical inference itself deductive.

The fact of analogy is an empirical fact about the source and target domain that, if true, makes the analogical conjecture necessarily hold in the target domain. Norton says that if their fact of analogy is true, then analogical "inferences proceed deductively" from the truth of the fact of analogy. "If the conjectured fact is unequivocal and held unconditionally, the analogical inference from one system to another may simply be deductive, with all the inductive risk associated with the acceptance of the fact of analogy [Norton 2021, pg.

133].” Norton says that his account aims to provide conditions for an analogical inference’s justification. “What makes some particular analogical inference a good one? My answer is that the chain of justification terminates materially, in a fact of analogy. The inference is only good in so far as that fact is a truth [Norton 2021].”

Norton claims that framing analogical inference in this way dispenses with the vague notion of similarity and leaves behind a more clear notion, sameness of properties, which similarity was merely a proxy for all along. Norton says,

“These material analogies reduce the similarity relation to sameness of properties. The earth and moon are analogous in their sphericity since they carry the same property, sphericity [Norton 2021].”

Giving another example, the liquid drop model of the atom, Norton says “We also see once again that the similarity of the source and target is a subsidiary matter. What matters to the analogy is what is expressed in the fact of analogy, that the liquid drop and nucleus share just the properties listed.” This attempt to reduce similarity to shared properties in order to pin it down as a philosophically meaningful notion is not new; Tverksy [1977] attempts the same strategy. This strategy is not a successful one, because basing similarity on shared properties requires determining what properties are inductively meaningful and how to weigh their relevance. As Quine puts it, “But properties in such a sense are no clearer than kinds. To start with such a notion of property, and define similarity on that basis, is no better than accepting similarity as undefined [Quine 1969].”

Let us examine more closely Norton’s attempt to reduce similarity to sameness of properties. Norton says that the ‘fact of analogy’ is (1) “is a factual state of affairs that arises when two systems’ properties are similar” and (2) “the exact mode of correspondence [is] expressed as part of that fact [Norton 2021]. Norton seems to require that in order for a fact of analogy to warrant an analogical conjecture, it must express an objective similarity between the source

and target domain. Not only this, but it must express *the* similarity that makes it necessary that the analogical conjecture holds in the target domain. An analogical inference is not warranted if the fact of analogy turns out to be an incorrect description of the underlying mechanisms causing the observed similarities between the source and target domain. Norton says "The fact of analogy must be a truth if the analogical inference is to be warranted. If it is not so, then the candidate analogical inference is an inductive fallacy [Norton 2021]." The fact of analogy also must be precisely expressed in order for the analogical conjecture to be plausible. Norton says, "If the fact of analogy is too vague to support the inference, then [the inference] is not warranted [Norton 2021]."

I have three concerns with this: (1) This condition for similarity is too restrictive, (2) this descriptive account of the *process* of analogical inference is inaccurate, and (3) it fails to meet one of Norton's main goals, which is to give a theory of analogical inference that gives proper attention to the local and context dependent nature of an analogical inference's plausibility.

First I will address the concern that this condition is too restrictive. Norton's approach, like Bartha's, appears to be partially motivated by Quine. Under Quine's account of similarity, similarity is a component of an epistemic process in which we start scientific enquiry with our biologically-furnished, innate sense of similarity. As a science develops, the innate sense of similarity is replaced by more complex, conceptual similarities that can be explicitly expressed in formal terms. In a more mature science, we are not reliant on a notion of similarity at all because we have a detailed scientific language to describe the exact correspondence in properties or structure that before was vaguely represented by the notion of similarity. Quine says that our sense of similarity progresses through a maturation "starting in its innate phase, developing over the years in the light of accumulated experience, passing then from the intuitive phase into theoretical similarity, and finally disappearing altogether [Quine, 1969, pg. 138]". For instance, the statement that Tellurium and Selenium are similar is replaced by the statement that they share certain properties such as photoconductivity

due to shared features of electron configuration.

Quine also notes that scientifically precise articulations do not totally do away with more intuitive, informal judgments of similarity:

> Between an innate concept of similarity or spacing of qualities and a scientifically sophisticated one, there are all gradations. Science, after all, differs from common sense only in degree of methodological sophistication...Something like our innate quality space continues to function alongside the more sophisticated regroupings that have been found by scientific experience to facilitate induction [Quine].

Norton, however, departs from Quine in that he denies that the vague notion of similarity ever plays a legitimate role in justifying analogical inferences. Instead, Norton requires that an inductive inference always rely on a clearly expressed fact of analogy and claims that the vague notion of similarity *never* factors into legitimate analogical inferences.

This delegitimizes a wide class of analogical inferences in which scientists recognize a regularity in nature but lack a detailed explanation of the underlying causal nature of those regularities. One example, discussed by Stegmann [2023] is pre-Darwinian biologists' recognition of homologies in organism features, like the phylogenetic homology between mammal forelimbs. Naturalists who preceded Darwin recognized that there was a deep similarity between, for instance, the arm of a human and the front leg of a cat. The actual 'fact of analogy' that causes that correspondence in structure, shared evolutionary history, was not recognized by naturalists because their research preceded Darwin. Instead, they relied on explanations from natural theology like intelligent design. Under Norton's account, this would not constitute an appropriate fact of analogy for analogical inference because it does not correctly describe the causal mechanism that creates the similarity in forelimb structure. Yet lacking the fact of analogy did not prevent scientists from correctly recognizing that a deep similarity exists and using that similarity to make correct predictions.

Stegmann proposes that the ability of pre-Darwinian naturalists to correctly categorize organisms in a taxonomy that tracks phylogentic homologies is due to an innate feature-mapping ability. In a psychology study, Stegmann shows that untrained non-experts make the same homology predictions that early naturalists did. Thus, Stegmann concludes that early naturalists judgments were guided by an innate, intuitive feature-matching ability that is an evolved component of human cognition [Stegmann 2023].

Norton's criteria for similarity disregards the efficacy of our innate sense of similarity. Many of our similarity judgments are based on innate, biologically evolved mechanisms that we lack epistemic access to. We might not have an exact theory of *how* two things are similar, but just feel that they *are*. Often the 'fact of analogy' is the subject of analogical inference rather than the grounds for it. As Quine notes, often the sense of similarity exists long before we have any detailed scientific theory that can explicate the nature of the similarity.

Other times, scientists' goal is to use a vague fact of analogy as a fallible tool rather than establishing the truth of the fact of analogy. Ankeny [2014] discusses scientists use of non-human animals as models to form conjectures about human behaviour: "Inferences as to whether the organism itself can reliably stand for human beings, and whether the experimental setting bears any relation to the social situations in which a putatively similar human behavior would be instantiated, remain open-ended." Scientists often use a similarity judgment as an epistemic tool without aiming to confirm or deny it. The observation that mice can be used as a model for humans is not an empirical fact that scientists aim to definitively confirm. Scientists merely use the observation that humans and mice share biological similarities as a fallible, but useful template to guide hypothesis generation. They use the vague notion of the similarity of mice and humans to create analogical conjectures to empirically test. Determining the precise nature of the fact of analogy would make the analogical conjectures themselves redundant.

Another counterexample to Norton's theory is Mendeleev's construction of the periodic table,

which I described in Chapter 3. Mendeleev believed that there was an underlying mechanism, periodicity, that caused the regularity he observed between atomic mass and chemical properties. He used the principle of periodicity as a 'fact of analogy' to form useful, true analogical conjectures about the chemical properties of unexamined elements, and to organize the elements into a periodic table. Despite its efficacy, his fact of analogy did not correctly capture the underlying mechanism that causes regularities between atomic mass and chemical properties. It is atomic number, not atomic mass, that causes the chemical properties of elements. The fact of analogy that Mendeleev used to justify his analogical inferences was incorrect, yet it functioned as a proxy for his deeper conviction that there was some deep causal relationship that determined chemical properties of the elements. In scientific practice, our language and theoretical commitments change as our scientific paradigms change, such that our fact of analogy is always changing. But our credence that deep similarities and regularities exist in nature between, say, artificial selection and natural selection endures even as our beliefs about the exact mechanism change. If Norton is correct, and analogical conjectures are only warranted if the fact of analogy is correct, then most of the analogical inferences made in the history of science would turn out to be unwarranted because scientists' understanding of the underlying causal mechanisms have changed since the time that the analogies were made.

One final counterexample to Norton's theory is the analogy that electricity is like a fluid, which was proposed after the discovery that electricity could be 'stored' in a Leyden jar, just as one would store a liquid. 'Electricity is like a fluid' was used as the grounds for analogical conjectures about electricity for several hundred years, though the exact correspondence between water and electricity was not known. This assertion of similarity would not meet Norton's conditions for a warranting fact of analogy because the assertion of similarity is imprecise and fails to describe a causal mechanism that, if true, would make the analogical conjecture deductively true. Yet this analogy was used fruitfully by scientists and drove research forward. Over a hundred years later, a more precise expression of the correspondence,

'electron flow is like water flow' was developed. Would this redeem the fluid/electricity analogy under Norton's account? It would not, because it still does not describe the precise nature of the similarities between fluid and electricity such that any analogical conjectures about electricity are deductive. Yet any more precise articulation of the fact of analogy will either be false, because it will predict some similarity between electricity and fluid that does not exist, or it will describe the exact nature of electricity that the analogical inference aimed to establish. Any articulation that is (1) adequately precise to make the analogical conjecture deductive, yet (2) does not make any false claims about the correspondences between electricity and fluid, will be exactly the description of the nature of electricity that the analogical inference is being used to investigate. The electricity/fluid analogy was fruitful for scientists as a heuristic tool for making fruitful analogical inferences, but these analogical inferences would not be warranted under Norton's account.

Norton's account also fails to be local and context dependent because it does not allow for the temporal and cultural relativity of scientific theory. Scientists' explanations for what empirical fact unites two domains changes over time as paradigms shift, but the credence that a similarity between two phenomena exists often endures through this process. The fact of analogy changes, but the similarity judgment endures. Suppose that scientists use an analogy between the spinning of a top and the spinning of a planet to form a conjecture about the motion of the planets. Under Newtonian mechanics, the 'fact of analogy' would be that objects fall to Earth and the planets orbit the sun because they are drawn by a gravitational force. Under general relativity, the 'fact of analogy' would be that objects travel in a straight line along curved spacetime. Do we want to reject a pre-GR analogy between the spin of a top and the spin of a planet because it relies on a 'fact of analogy' that is incorrect and describes gravity as a force rather than a curvature of spacetime? Causal explanations change over time, but the recognition of regularities in nature are often invariant across scientific communities even as the explanations for those regularities change.

Another concern is that Norton, in rejecting less precisely explicated forms of similarity, seems to disregard the role that analogy plays in the epistemic process of science. Often the fact of analogy is the end goal of analogical enquiry, whereas Norton, in trying to dispense with the notion of similarity, portrays it as one of the preconditions for analogy.

Another point, raised by Genta [2020] is that Norton's account appears unable to capture structural similarities, or at least a certain type of abstract structural similarity. Genta gives the example of an analogical inference from Holyoak et al. [1980], in which a doctor confronted with a malignant tumor recalls a story about a city being conquered by an army of small battalions approaching from multiple angles. The doctor uses this story to form a conjectured strategy of attacking the tumor from multiple angles with small amounts of radiation [Holyoak 1980]. Genta raises the concern that in the analogical inference, there is no material, empirically testable 'fact of analogy' that unites the source and target domain, because the similarity between source and target is an abstract structural similarity. Structural analogies of this type play an important role in scientific reasoning, yet appear to fall outside the scope of Norton's account [Genta 2020].

In summary, my claim in this section is that current accounts of analogical inference in philosophy and cognitive science neglect to adequately investigate the notion of similarity, and instead treat similarity as an unanalyzable hinge of experience. They fail to meet the desiderata that I laid out in the prior section.

In Chapter 2, I argued that the current two-dimensional accounts of analogical inference weigh relevant similarity too heavily when considering the factors that influence plausibility. I claimed that by treating relevant similarity as a necessary and sufficient condition for *prima facie* plausibility, they fail to acknowledge the other factors that affect plausibility like relevant *dissimilarity*, background assumptions local to the target domain, etc. This chapter offers a further challenge, arguing that current normative accounts fail to adequately explicate the notion of similarity that they so heavily rely on.

The neglect of similarity in current normative accounts of analogical inference is analogous to the treatment of combustion chamber pressure in early engines. In internal combustion engines, the power stroke is generated when the fuel combusts, increasing pressure in the engine and moving the piston. Energy output is a function of energy input, pressure inside the cylinder, and the design of the pistons (factors like shape, material, and friction affect piston efficiency). Early engines operated merely by using natural atmospheric pressure. When engineers worked on improving the efficiency of engines, they assumed that internal cylinder pressure was an unalterable fact of nature that could not be manipulated, and that increases in efficiency had to come from other areas, like the innovation of a double-acting engine where the piston is moved from both sides.

In the late nineteenth century, engineers realized that a previously unrecognized way to increase the efficiency of internal combustion engines was to artificially increase the amount of pressure inside the cylinder rather than relying on atmospheric pressure alone. They created the first engine with a compression stroke, which increased the pressure inside the combustion chamber. This was more efficient and resulted in more energy output than the prior strategy of relying on atmospheric pressure alone.

I propose that a similar assumption and neglect has been made in the analogical inference literature in regards to the notion of similarity. In current accounts, the plausibility of an analogical inference is a function of the strength of the source domain and the degree of similarity between source and target domain. As such, current accounts aim to give a once-and-for-all characterization and justification of relevant similarity. This is misguided. The solution to the need for a normative account of analogical inference is not to provide a universal, once-and-for-all characterization of relevant similarity, but to give scientists tools for empirically investigating norms for relevant similarity locally within their domain.

4.4 What epistemic norms for similarity *could* a normative account of analogical inference provide?

In the last section, I argued that current normative accounts of analogical inference do not offer justification or useful norms for the similarity judgments that underlie analogical inference. What justification and epistemic norms for similarity *could* they provide? The following section will review several attempts in philosophy and cognitive science to provide a justification of similarity-based reasoning. These accounts attempt to give a once-and-for-all justification for similarity. This is misguided. There is no ultimate justification for the psychological heuristics, cultural norms, and scientific paradigms that fix our similarity judgments. There is no definitive answer to the question of which similarity judgments are the relevant ones, as standards for relevant similarity will change as science evolves. I will use these accounts to show what a normative account of similarity ought not to strive for. In the subsequent section, I propose a naturalistic and pluralistic normative account of similarity. I examine how scientists do in fact fix respects for relevant similarity. Norms for similarity are formed in the trenches of domain-specific scientific enquiry. Yet what is common across domains is that norms for similarity are fixed by innate psychological characteristics, cultural values, and paradigm-specific norms. I argue that a successful normative account of similarity requires ongoing naturalistic investigation into the process by which scientists fix respects for similarity. This information should be used to construct a fallible and pluralistic toolkit of methodologies and epistemic values that improve the reliability of scientists' similarity judgments.

4.4.1 No ultimate justification or norms for similarity

The realist view

In the cognitive science literature, one attempt to justify similarity as a grounds for inductive inference is to claim that the similarity judgments we make should be approximations of objective similarities existing in the world. I will call this the *realist view*. This view is exemplified by geometric models of similarity like Shepard's that dominated the psychology literature until an influential criticism by Tversky [1977].

It is important to note that this view does not commit one to being a phenomenal realist. Instead, under this view, our qualia might be subjective, such that we do not experience the real, mind-independent nature of the world. However, even if our qualia are not representative of the world, the similarity relations between the contents of our experiences are objective and mind-independent.

Hume might be seen as a proponent of this view. Even though Hume advocates that ideas can be associated by relations of resemblance, he also suggests that resemblance is different from the other associative principles of contiguity and causation in that resemblance is more like a basic property contained in the impressions themselves. For Hume, our impressions are the most basic, irreducible, and immediate contents of our experience. Hume seems to suggest that resemblance can be equally basic, and in some instances is inseparable from the contents of our impressions. Hume gives the example of color: "Blue and green...are more resembling than blue and scarlet; tho' their perfect simplicity excludes all possibility of separation and distinction [Treatise 1.1.7]."

Blue, green, and scarlet are basic perceptions–they cannot be broken down into more primitive parts. The basic perception of blue, green, and scarlet is inseparable from the relations of resemblance holding between them–the similarity of blue and green seems to be built

into the impression of blue and green. We cannot percieve blue, green, and scarlet without perceiving the similarity of blue and green relative to the similarity of blue and scarlet. This makes resemblance different from contiguity and causation in terms of the conceivability, and thus metaphysical possibility, of them being otherwise. Hume says, "Whatever we conceive is possible, at least in a metaphysical sense [Hume Treatise]." We can conceive of a billiard ball striking another and coming to a standstill, or of opening the door of a favorite coffee shop and finding ourselves in the rainforest. However, we cannot conceive of blue resembling scarlet more than it resembles green.

Ross [1991] notes that Hume often speaks as though resemblance is a content of our impressions; Hume says that the relation of resemblance is "presented to the senses" and is "found among objects [Treatise]."

This does not imply that resemblance is ontologically real under Hume's account; it could be merely epistemically real. Under Hume's account, our mind-dependent impressions may or may not resemble the mind-independent world. So the resemblances between impressions, for Hume, are not necessarily relations that exist in the external world. Rather, resemblances are relations that hold between perceiver-dependent experiences, and the experiences may or may not resemble the world. However, for Hume, the *relations* of similarity and difference holding between a perceiver's experiences seem to be objective relations holding between the contents of the perceiver's subjective world.

Hume's goal was to reduce the workings of the mind to its constituent laws, as Newton does for physics. However, in his investigation, Hume did not include an investigation of the cognitive processes that generate resemblance, instead treating resemblance as an unanalyzable hinge of experience.

In an influential 1987 paper, cognitive scientist Roger Shepard aims to continue Hume's project by providing an explication of an objective law underlying the faculty of resem-

blance. Shepard aims to carry on Hume's project of examining "the question of whether psychological science has any hope of achieving a law that is comparable in generality...to Newton's universal law of gravitation [Shepard 1987]." Noting that similarity judgments form the foundation of all inductive inferences, Shepard says that "Psychology's first general law should, I suggest, be a law of generalization [Shepard 1987]."

Shepard describes the motivation for his geometrical account:

> I was now convinced that the problem of generalization was the most fundamental problem confronting learning theory. Because we never encounter exactly the same total situation twice, no theory of learning can be complete without a law governing how what is learned in one situation generalizes to another [Shepard 1987].

Shepard's geometric model represents similarity and dissimilarity as metrics in a coordinate space, where similar objects are closer in the coordinate space on more dimensions, and dissimilar objects are farther [Shepard 1974]. For example, the color, size, and shape of a set of balls could be represented in a three-dimensional coordinate space, with color occupying one dimension, surface area occupying another dimension, and mass occupying a third dimension. In the coordinate space, two balls, A and B, are similar to each other if they are located close to each other in the coordinate space, and they are located close to each other if they fall in a similar location on the spectrum for each of these three dimensions.

Definition 4.1. Universal Law of Generalization: The probability that an encountered stimulus has the potential for the consequence associated with a previous stimulus, despite their perceived difference, decreases exponentially with the distance between them in their representational space.

Under Shepard's model, an agent's likelihood of generalizing a behaviour from a familiar

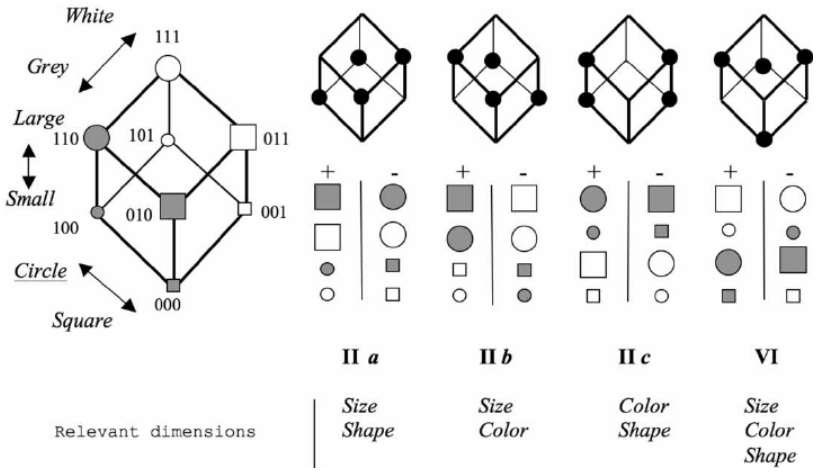


Figure 4.5: The geometric model

to an unfamiliar stimulus should be a function of the likelihood of them displaying that behaviour for the known stimulus and the distance between the known stimulus and the unknown stimulus in the coordinate space.

The goal of Shepard's project was essentially to do for similarity judgments what the axioms of probability do for reasoning about basic conditional probabilities. Shepard's *Universal Law of Generalization* is intended to be both normative and descriptive. It is descriptive because it describes an "optimum cognitive strategy in the world" that Shepard claims is universal across animal species and human cultures because natural selection selects for optimum cognitive strategies [Shepard 2004]. Shepard's account is also normative because Shepard claims that agents whose generalizations depart from the predicted generalizations of Shepard's model will be committing themselves to non-optimal behaviour.

In contrast to Hume's mostly descriptive account of resemblance, the geometric model offers norms for similarity. The model is posited as an explanation for the past success of our similarity judgments. It says that our similarity judgments are successful when they adhere to the axioms of the geometric model. So with the geometric model, we can not only say that our similarity judgments were successful in the past, but can also make a claim about why they were successful.

In contrast, Hume's account is descriptive and only offers the barest sliver of justification by appealing to the past success of our similarity judgments. His stance in the *Treatise* is that it is fruitless to look into the *causes* of resemblance, but that there is something to be gained from examining the *effects* of resemblance. I take this to be an appeal to the fact that our similarity judgments have guided successful action in the past.

In summary, the assumption of the realist view is that similarity can be objectively quantified, and that our similarity judgments are justified when they align with this metric. Of course, we do encounter situations where our brain and senses mislead us and present two things as being similar when they are in fact different, or different when they are in fact similar. The Shepard table is one such example. However, the realist assumes that aside from these occasional quirks of our cognition, our similarity judgments overall present an approximation of real similarities holding between the objects of our experience.

Challenges for the realist

There are several challenges for the realist view. The first problem is that there is no matter of fact about which properties are the relevant ones for determining the similarity between two objects. As has been stated by Putnam, Peirce, and Goodman, bare similarity without some filtering criterion for subjective salience is a vacuous notion. Charles Sanders Pierce says, "The forms of the words similarity and dissimilarity suggest that one is the negative of the other, which is absurd, since everything is both similar and dissimilar to everything else."

Goodman says:

> When, in general, are two things similar?...Since every two things have some property in common, this will make similarity a universal hence useless relation.

> That a given two things are similar will hardly be notable news if there are no two things that are not similar [Goodman 1972].

When we determine that two sets of features are similar, we are not picking out a special, privileged feature of the world; we are picking out a feature that is salient to scientists in a particular context and in aid of some particular goal. Judging similarity involves both selecting the relevant properties to be compared and assessing their relative relevance. Relevance will be relative to the scientist and particular domain in question. For example, a phylogenist may judge a snake and a mountain lion to be dissimilar while an ecologist studying the predation of gophers might consider them similar due to focusing on habitat and ecological role. Goodman concludes that similarity is too unconstrained and arbitrary to be a meaningful ground for inductive inference.

In practice, the choice of which properties we notice and use in our similarity judgments is constrained by psychology, cultural norms, individual learning, and scientific theory. However, while our similarity judgments are constrained by these factors, these factors merely constrain how we *do* fix respects for similarity, not how we *should*. There is no ultimate justification for using these factors to fix similarity judgments, so the fact that similarity *is* constrained does not help the geometric model achieve its normative goal. The geometric model also fails at its descriptive goal of describing how people do make similarity judgments. The above constraints on similarity create context-dependent, subjective respects for similarity, while the geometric model describes similarity as an objective, universal notion.

Consider an example of similarity being context-dependent. When children are given a block of a certain size, shape, and color and asked to identify what other objects are similar to it, they preferentially select objects that are a similar shape over ones that are a similar size or color [Smith 2000]. This appears to be an innate, evolved bias. In this case, the geometric model would make the wrong predictions as similarity of two objects is not a sum of metric

distance in the representation space; some properties are weighed more heavily than others.

There are also many examples of cultural variation in similarity judgments. Children in Western cultures perceive phenomena as being more similar if they share attribute-based similarities, whereas children in China prioritize complex relational similarities [Richland 2009].

Psychology studies suggest that language differences may create differences in how similarity and difference are perceived. The Russian language divides up the color spectrum differently and features a linguistic distinction between light blue and dark blue colors, which the English language lacks. A 2007 study found that when shown a range of shades of blue, Russian speakers differentiated between two shades of blue that English speakers perceived as identical [Winawer 2007]. Russian speakers were able to differentiate between two blue shades that fell on different sides of the Russian linguistic distinction between light blues and dark blues. It was hypothesized that the Russian linguistic distinction between light and dark blue made it possible for Russian test participants to discriminate between the two colors [Winawer 2007].

Some species, like Mantis shrimp, even demonstrate plasticity in their visual perception of similarity. Mantis shrimp have a mechanism by which they adapt the sensitivity of their color receptor cells to different wavelengths in response to their environment, such that their color perception varies across time [Street et al. 2022].

Finally, Tversky notes another way in which the geometric model fails to be descriptively accurate. Tversky notes that psychology studies on similarity judgment find effects such as asymmetry, non-transitivity, and non-identity that Shepard's model cannot account for [Tversky 1977].

For example, "the judged similarity of North Korea to Red China exceeds the judged similarity of Red China to North Korea. Likewise, an ellipse is more similar to a circle than a

circle is to an ellipse [Tversky 1977]." Asymmetric similarity judgments violate the geometric account, but can be fruitful.

Tversky also notes that the geometric model requires that properties be represented on a metric continuum. For the examples that Shepard's model focuses on, like color and mass, this is unproblematic. However, while some properties like color and mass can be represented as being binary and ordinal with minimal background assumptions, other properties are not binary or have no obvious ordinal ranking [Tversky 1977].

An alternative to the realist view, which I will call the *relativist view*, is to acknowledge that similarity judgments are dependent on the perspective of the observer.

The relativist view

There have been two influential accounts of similarity in response to the shortcomings of Shepard's account.

First, Tversky [1977], after noticing the shortcomings of Shepard's quantitative model of similarity, proposed a qualitative, feature-matching model of similarity. Under Tversky's model, two entities that are being evaluated for similarity each consist of a set of features, and the degree of similarity between the two entities is a function of their shared and mutually exclusive features.

A positive characteristic of Tversky's account is that it accommodates the asymmetric, non-transitive similarity judgments that can be seen in actual scientific practice. Because Tversky's model relies on overlap in features rather than a geometric metric, the model allows that an unfamiliar domain with few known features can be judged more similar to familiar domain with more salient features than the reverse. This accounts for asymmetric similarity judgments. Tverky's model also accommodates cases where similarity judgments

are based on features that are not binary or ordinal.

There are two issues with this, with regard to the aims I have described in this paper. First, Tversky's account aims to reduce similarity to shared properties in order to pin it down as a philosophically meaningful notion. However, the very problem this paper is trying to address is how to pin down similarity judgments, given that similarity appears to be the basis for properties. Tversky's account relies on the existence of properties in order to ground similarity. If comparing a cat with a mongoose, the similarity would be grounded in shared properties like 'has fur' and 'hunts prey'. However, using shared properties requires determining what properties should be compared and how to weigh the relevance of these properties. This is the very question that an account of similarity aims to address–what makes green objects more similar than grue objects? Quine makes this point: "But properties in such a sense are no clearer than kinds. To start with such a notion of property, and define similarity on that basis, is no better than accepting similarity as undefined [Quine 1969]."

Second, Tversky's account does not offer justification for our similarity judgments; it describes the similarity judgments that people tend to make, but does not say when we ought to rely on a similarity judgment for inductive reasoning. To make Tversky's account normative, some additional dimension would be needed to explain why an agent's similarity judgments aligning with the model's predictions would be rational.

The second type of relativist account is an extension of Shepard's geometric account. Accounts of this type come from Churchland [1992] and Gardenfors [2000]. These two accounts are an extension and refinement of the geometric approach proposed by Shepard in that entities are represented in the model as points in a multi-dimensional space, where the dimensions are properties of objects, and the similarity of two objects is roughly an inverse of their distance in the coordinate space. The main upshot of Churchland and Gardenfors' accounts is that they allow similarity judgment to be subjective; different agents may weigh different properties differently, resulting in different similarity judgments when viewing the

same objects. Though they rely on a geometrical structure to represent similarity, they do not follow Shepard in claiming that there is a *single* conceptual space that governs all similarity judgments. Instead, geometrical spaces are contextual; different contexts and properties have different conceptual spaces with different convex regions [Gardenfors 2000].

However, this relativism, while necessary, also makes the task of giving a normative account of similarity challenging. Gardenfors says, "Since a conceptual space may seem like a Kuhnian paradigm, aren't we thereby stuck with an unavoidable relativism? After all, anyone can pick her own conceptual space and in this way make her favorite properties come out natural in that space [Gardenfors, 2000]."

If the contextual geometric accounts purport to be normative, then they must have answers to two questions: (1) Is the process by which convex regions are fixed reliable? and (2) how should agents choose which conceptual space is appropriate for a particular context? Question (2) requires agents to make a similarity judgment in their choice of conceptual space.

A challenge for the relativist

Given that similarity judgments vary wildly between species, across cultures, and relative to the goals of the individual making the judgment, the relativist view faces the challenge of explaining why similarity judgments are a suitable foundation for scientific enquiry, which aims to uncover the objective structure of the world. In order for inductive reasoning to be reliable for scientific enquiry, it must be the case that our similarity judgments have some connection to the structure of the world. If an analogical conjecture in science hinges on the judgment that entities A and B are similar, then we must have reason to believe that the similarity of A and B is a fact of the world, not a fact of a particular cultural belief, or a genetically-evolved cognitive quirk that conflates A and B because it was reliably adaptive

to do so in humans' evolutionary history.

Under the relativist view, what makes our similarity judgments a reliable foundation for science?

One answer to the question is to claim that the similarity judgments used in inductive inference are shaped and refined by science, and this makes them reliable. A way that we can bolster the reliability of our similarity judgments, as Quine suggests, is to subject them to scientific enquiry and give precedence to the natural kind categories that agree with our scientific theories. However, this cannot be the only measure that we take, for two reasons. First, our similarity biases affect our scientific theory choice, and so our scientifically bolstered natural kind categories will also be products of our innate evolved disposition to see certain sorts of similarities. The very question that the analogical inference literature aims to address is how similarity forms a reliable ground for the analogical inferences that scientists use to inform theory choice.

It is circular for the justification for the similarity judgments scientists use in theory choice to be that the similarities that are relevant are those that feature as categories in a scientific theory. Because there is no ultimate justification for our similarity judgments, some degree of circularity is unavoidable. However, our innate sense of similarity judgments acts as a checkpoint on this process to prevent vicious circularity, and should not be disregarded.[2] Thus, when seeking a justification for the reliability of our similarity judgments, we need to have some guidelines for where and when our primitive, genetically-inherited similarity judgments are reliable.

[2]For instance, traditional modular theories of the brain predicted that the specific location of lesions in the brain was the best way to classify brain disorders. If two patients demonstrated similar cognitive symptoms, but MRI imaging showed that they had lesions in different regions, then the people with similar symptoms were treated as having distinct, unrelated brain disorders. A few scientists pointed out that this theory did not account for the high degree of similarity between the symptoms of patients with different lesion locations, and argued for prioritizing basic perceptual similarities over theory-sanctioned similarities [Geschwind 1965]. This ended up being fruitful and contributed to a distributed network model of the brain being developed, in which the same brain disorder can be caused by very different patterns and locations of brain lesions.

Second, our scientific theories change, and are culturally relative, and often our innate sense of similarity and dissimilarity is the common ground that we can agree on. For instance, philosophers who object to Kuhnian strong incommensurability claim that in cases of debate over paradigm choice, scientists often retreat to more pre-theoretic beliefs and observations in order to have a common language with which to form a starting point for debate [Stanford 2024]. Our simple, pre-theoretic similarity judgments can guide theory choice in these situations. When scientists disagree about which predicates correctly describe the sorts of features suitable for inductive inference, they can find common ground in their pre-theoretic similarity judgments.

So refining our natural kind categories by treating the ones that are products of scientific theory as a privileged class is not the only refinement work we need to do. We also need to establish the reliability of our basic perceptual similarity judgments.

One approach to establishing the reliability of pre-theoretic similarity has been to assert that our similarity judgments are evolutionarily adaptive, and that evolutionary adaptivity entails epistemic reliability. The argument is that if our similarity judgments did not approximate reality, then we would be at an evolutionary disadvantage and those of us with non-veridical similarity judgments would have a lower rate of survival and reproduction.

Philosophers like Kornblith and Gardenfors take this tack, following Quine.

Quine says,

> "There is some encouragement in Darwin. If people's innate spacing of qualities is a gene-linked trait, then the spacing that has made for the most successful inductions will have tended to predominate through natural selection. Creatures inveterately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing their kind [Quine 1969].

Philosophers and cognitive scientists have uncovered cases where our natural intuitions furnished by evolution have well-equipped scientists for scientific enquiry. One notable example comes from Stegmann et al., who shows that innate aptitudes for finding visual correspondences between different objects create similarity judgments that are indistinguishable from those made by pre-Darwinian biologists [Stegmann 2023].

However, some of our innate similarity dispositions are obstacles to successful scientific enquiry. As Quine notes,

> Color impresses man; raven black impresses Hempel; emerald green impresses Goodman. But color is cosmically secondary...If man were to live by basic science alone, natural selection would shift its support to the color-blind mutation. Living as he does by both bread and basic science, man is torn. Things about his innate similarity standards that are helpful in the one sphere can be a hindrance in the other [Quine 1969].

Our innate sense of similarity continually presents us with objects that are similar in color. These color similarities are psychologically salient, as they should be for any good foragers, but they are not salient for most scientific inquiry. A slightly more concerning issue is the one I raised earlier: that the colors we perceive do not have a one-to-one mapping to light wavelength. So not only is our sense of similarity of color evolved to fulfill a demand that is irrelevant to the demands of science, it is also the case that our judgment of similarity of color obscures the reality of the world by presenting similarities that are artifacts of our psychology rather than representative of any external structure.

However, a more modest proposal might be possible. One possibility is that even if our perceptions are not *representative* of reality, they are still *reliable* as a belief forming mechanism for scientific practice. This seems to be the strategy that Gardenfors wants to take to add a normative dimension to his account, saying,

> The appropriate question on the relation between the conceptual structure and the world is whether a conceptual structure is *viable* or not. Having a viable conceptual structure means that one is able to solve the essential problems when acting in the world. Via successful and less successful interactions with the world, the conceptual structure of an individual will adapt to the structure of reality. It must be emphasized, however, that this does not entail that the conceptual structure represents the world [Gardenfors 2000].

The challenge for this view is that research in cognitive science increasingly shows that evolutionary adaptivity and epistemic reliability come apart.

Evolution selects for traits that increase our ability to survive and reproduce. Research in psychology and cognitive science suggests that there are numerous points at which the demands that evolution places on our similarity judgments diverge from the demands of science.

First, our perceptual reality encodes cognitive shortcuts that guide successful action in environments resembling those encountered in our evolutionary history. This limits the transferability and generalizability of our similarity judgments to different domains. A rule is a fast, frugal way to approximate the behaviours that understanding would create. However, when an agent has access to a rule that governs successful responses to a phenomenon, but lacks access to the underlying *reasons* that make that rule successful, they can only act successfully as long as they are operating in an environment where that rule is applicable. Having simple rule-of-thumb heuristics instead of deeper understanding is adaptive because it rations attentional resources, When rules replace information, it creates scenarios like the five monkeys thought experiment where agents have a particular behaviour but don't have access to the original reasons for the behaviour.

The problem is that veridical representations are context-invariant in that the representa-

tions still provide useful information if the environment changes, but heuristics are context-dependent. If evolution has selected for useful heuristics over veridicality in our sense of similarity, then these heuristics will work in the contexts that arose in our evolutionary history but not in novel contexts. Studies suggest that the demands of science diverge in important ways from the survival problems that humans encountered in their evolutionary history [Kahneman 1974; Cosmides et al. 1992].

Second, having inaccurate perceptions can be more adaptive than having accurate perceptions. Being wrong in the 'right' way can be selected for more than being right. Hoffman et al. [2010] use game theoretic modeling to provide support for the claim that having non-veridical perceptions is adaptive. In his model, one set of agents have fully veridical perceptions, one set have partially veridical perceptions, and one set have fully constructed, non-veridical perceptions. The game theoretic model simulated evolution of five hundred generations of the organisms, and found that the population with non-veridical perception had the best adaptive fitness.

From these results, Hoffman concludes that our perceptual experience is populated by icons rather than roughly veridical representations of our environment. Hoffman uses the analogy of icons on a computer. Icons on a computer help guide successful action by obscuring the underlying software and hardware from us, and instead presenting us with an interface that makes it easy for us to make the correct choices. If one icon shaped like a folder says 'Important files' and another icon shaped like a wastepaper bin says 'trash', we are guided away from the unsuccessful action of accidentally deleting our important files.

Under Hoffman's theory, our perceptual system performs the same function, acting as an interface that mediates and guides successful action in our environment. Hoffman's theory claims that our cognitive faculties have evolved to misrepresent the world, because perceiving the real structure of the world would be maladaptive. This suggests that our sense of similarity could be inaccurate by design.

Philosophical, psychological, and anthropological investigation is needed to see what rules our cognition uses to construct similarity judgments, and whether these rules are epistemically reliable for scientific enquiry. The demands of science likely diverge in places from the selection pressures of our evolutionary ancestors' environments. If so, then the locations of these divergences need to be discovered, and epistemic norms postulated that bolster the rationality of our similarity judgments in these areas.

Another relativist account: Goodman's entrenchment

Hume acknowledges that the cognitive process that generates resemblance is not free from error. He notes, "Of the three relations above-mention'd that of resemblance is the most fertile source of error; and indeed there are few mistakes in reasoning, which do not borrow largely from that origin [Hume Treatise 1.2.5.21]." In the *EHU*, Hume says "Nothing so like as eggs; yet no one, on account of this appearing similarity, expects the same taste and relish in all of them [EHU I.I.V]."

When making similarity claims, reliability and relevance can come apart. When we make a similarity claim, we generally intend to say something about the world, not just about our psychology [Quine 1969]. Someone might claim that their similarity judgment is reliable if there is reason to believe that the claimed similarity is more than a psychological artifact. But this comes apart from the question of whether the similarity tracks a relevant structure in the world such that it is relevant grounds for a particular induction [Goodman 1955].

For Hume, the answer to the question of how we form associations between experiences in order to reason inductively about the future, is that experiencing associations between A-type events and B-type events causes us to develop a habit such that when we experience A-type events we naturally expect B-type events.

Goodman [1955] proposes a refinement of Hume's account. He argues that a successful

account of resemblance must say what *sorts* of similarities are the basis for our inductive inferences. We do not form habits on the basis of *all* of the associations we experience, but only on the basis of certain types of similarities. Goodman says,

> The real inadequacy of Hume's account lay not in his descriptive approach but in the imprecision of his description. Regularities in experience, according to him, give rise to habits of expectation; and thus it is predictions conforming to past regularities that are normal or valid. But Hume overlooks the fact that some regularities do and some do not establish such habits; that predictions based on some regularities are valid while predictions based on other regularities are not [Goodman 1955, pg. 82].

Cases of manipulated similarity, like Goodman's riddle of induction, have brought philosophical attention to the question of what makes a similarity relevant for inductive inference. In Goodman's puzzle, the predicate 'grue' applies to an object if and only if it is green and is observed before an arbitrary but particular time t, or if it is not observed before time t and is blue. For an observer before time t, the same empirical evidence is equally supportive of the proposition that all emeralds are grue as of the proposition that all emeralds are green. Yet our intuition says that the former proposition will make incorrect predictions about the future.

Successful induction appears not to be merely founded on similarity judgments, but on particular kinds of similarity judgments. How do we identify and justify these reliable and relevant forms of similarity?

For Goodman, the answer is entrenchment. A projectible predicate is one which has been reliably used in successful past inductions [Goodman 1955]. There are reasons to think that the biography of a predicate does track reliability to an extent. However, there are a few reasons that a richer and more naturalistic account than Goodman's is needed. First, the

factors that cause a predicate to become entrenched include cultural norms, psychological bias, individual learning, and scientific paradigms. A predicate can become entrenched because it tracks truth, or it can become entrenched because our psychology, cultural, or paradigm makes the predicate salient.

Consider Kuhn's example of Ptolemy's geocentric model. Aristotle's philosophy entrenched the notions of perfection, uniform motion, and circularity as being projectible predicates for induction about celestial objects. For a time, the entrenchment of these predicates insulated them against revision. When new evidence necessitated revision of scientists' theoretical framework, the entrenched predicates were protected from falsification because scientists revised other elements of the theoretical framework [Kuhn 1962]. For instance, Ptolemy's geocentric model of the solar system was developed as a response to empirical observations like retrograde motion and variations in the observed speed and brightness of planets. These observations threatened the Aristotelian notions of uniform motion and perfect circularity. Rather than demoting perfection, uniform motion, and circularity to the status of accidental rather than lawlike generalization, Ptolemy developed the concept of epicycles, which preserved the predicates that Aristotle's philosophy deemed salient.

In short, Quinean holism poses a challenge to Goodman's entrenchment solution because the failure of a predicate to be projected is partially a matter of scientific choice. A predicate fails to be projected when scientists *choose* to revise that predicate rather than other predicates in their web of belief. There are numerous studies in psychology that find that particular predicates are psychologically salient independently of whether they have a tendency to lead to correct predictions, such that they are preferentially projected more than other predicates without necessarily being more projectible.

For example, a [2000] study found that when children are given a wooden block and asked to identify other objects that are similar to it, they preferentially select objects that are a similar shape over ones that are a similar size or color [Smith 2000]. This appears to be an innate,

evolved bias. Shape may have been particularly salient for some evolutionary puzzle to the extent that humans evolved a bias to project shape predicates over size or color predicates. As a result, if shape is preferentially *projected* more than size it does not necessarily follow that shape is more *projectible* than size. The social epistemology literature contains various examples of unprojectible predicates being recalcitrantly projected because of sexist and racist social values that made those predicates salient to scientists [Longino 1990, Douglas 2014]. I discussed this in the context of analogical reasoning in [Lauman-Lairson 2025].

Goodman's notion of entrenchment also fails to offer guidelines for similarity judgments in instances of analogical reasoning. Analogical reasoning often depends on making similarity judgments in new domains where there is not established history of a predicate's use. I give an example in Sect. 3.1. In these cases, similarity judgments are central; scientists must judge how to import a predicate that applies to a familiar category of kinds to a category of kinds in which that predicate has no established history. For instance, in astrobiology, the presence of liquid water is an entrenched predicate in life-supporting planets when we consider the set of all planets that are known to support life, which consists of Earth. But we cannot appeal to entrenchment when deciding whether this predicate is projectible to the domain of all possible life-supporting planets. Instead, scientists must make a similarity judgment.

So, Goodman's critique of Hume is that not all regularities lead to inductive inferences. My further point is that scientists do not take all entrenched predicates to be projectible by virtue of their entrenchment. Rather, scientists are aware of the possibility that their inductive reasoning may be impacted by cultural, psychological, or paradigmatic biases. When scientists have reason to believe that such a bias could be active in the realm in which they are working, they take this into account when evaluating the validity of their inductions.

Thus, a satisfactory account of similarity must describe the processes that determine how scientists choose which predicates to project, and a successful normative account would give

scientists tools to improve the reliability of those processes (such as with considerations from social epistemology).

My critique is not in opposition to Goodman's account but rather is an extension of it. Goodman thinks that Hume is correct that there can be no ultimate justification for our inductive inferences [Goodman 1955, pg. 64]. He proposes that instead our inferences are locally justified by a process of reflective equilibrium. Goodman says,

> An [inference] rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either [Goodman 1955, pg. 64].

In order to bring our normative accounts of similarity into alignment with scientists' in-practice judgments of inductive validity, we must recognize that part of scientific practice involves identifying areas in which scientific practice is at risk of being unreliable. Consider the example of color. The Aristotelian notion was that colors are a property of objects that are transmitted to the eye [2023]. Newton challenged this notion by claiming that our psychology disposes us to think of color as a property of objects. He argues that color is not a property of objects, but a psychological sensation that is generated through interaction with light. Newton says,

> If at any time I speak of Light and Rays as coloured or endued with Colours, I would be understood to speak not philosophically and properly, but grossly, and accordingly to such Conceptions as vulgar People in seeing all these Experiments would be apt to frame. For the Rays to speak properly are not coloured. In them there is nothing else than a certain Power and Disposition to stir up a Sensation of this or that Colour [Newton 1704, pg. 125].

A component of scientists' web of belief are beliefs about ways in which predicates could become entrenched without being projectible. When weighing their commitment to an inductive conjecture, a component of scientists' evaluation is their belief about whether the process that resulted in a predicate becoming entrenched was a reliable one. If, for instance, scientists believe that a predicate became entrenched because we have an evolutionary bias to preferentially project that predicate, then scientists will consider whether that bias is one which is likely to be fruitful in the context in question. If the evolutionary problem that the bias evolved to solve differs from the epistemic demands of the scientific context in question, then scientists will weigh the possibility of unconscious bias into their degree of commitment to the analogical conjecture. Because this is a criterion in scientists' similarity judgments, norms for this evaluation must be included in normative accounts of analogical inference in order for them to be descriptively adequate.

4.4.2 What epistemic norms *are* possible?

The moral of the past section is that there is no objective set of norms for similarity. Yet, we should not be satisfied with a purely descriptive account of similarity. When we say that this piece of silver is similar to that piece of copper, and infer that the piece of silver will likewise conduct electricity, we are making a claim about the external world, not just about our psychology. Thus, even though there is no ultimate justification or epistemic norms for similarity, it is not satisfactory to abandon the normative project. Quine says,

> A man's judgments of similarity do and should depend on his theory, on his beliefs; but similarity itself, what the man's judgments purport to be judgments of, purports to be an objective relation in the world. It belongs in the subject matter not of our theory of theorizing about the world, but of our theory of the world itself [Quine 1969].

What is the first step of the normative project? First, we must understand what norms for similarity scientists are in fact committed to, and how these are formed. The similarity norms that scientists use are local norms generated in the context of domain-specific empirical investigation. As a scientific domain matures, psychological, primitive standards for similarity are replaced by theory-mediated standards that are local to the domain [Quine 1969]. Physicists' norms for similarity involve notions of symmetry and invariance across transformations. Phylogeneticts debate about whether similarities in DNA or morphology are more relevant for a given analogy.

A successful normative account of analogical inference will look different from the existing accounts described in the past sections. A successful normative account will not provide a once-and-for all justification for our similarity judgments, nor should it postulate a characterization of the kinds of similarity that are relevant. Standards for relevant similarity are domain-specific and are constructed in the trenches of scientific inquiry occurring in specific fields. Furthermore, standards for relevant similarity will change as science progresses. Instead, a successful normative account will look at how norms for similarity are fixed in different scientific domains and provide tools for refining this process. In practice, scientists' ways of fixing similarity are shaped by social values, history, and evolved cognitive biases, among other factors. A successful normative account of analogical inference will posit epistemic values that constructively refine this process. The epistemic values that I will postulate fall into two main categories: epistemic values that serve the aims of social epistemology, and epistemic values that help mitigate cases where evolved cognitive biases affect our similarity judgments.

Adding criteria for similarity to normative accounts of analogical inference would enrich the current epistemic toolkit that we have for evaluating analogical inferences. One of the pragmatic outputs of a normative account of analogical inference is to help us predict the plausibility of analogical inferences. Another is to help track the reasons that analogical

inferences succeed and fail. Adding criteria for similarity is necessary for both of these goals.

Values from social epistemology

Judgments of similarity are often constrained by social values. For instance, a variety of social factors led to the male body traditionally being used as a baseline for medical studies. This was historically seen as an obvious and unproblematic choice. Men have fewer hormonal fluctuations than women, so studies of men did not have to control for menstrual cycles. Women were also excluded from clinical trials due to various other social factors like (1) there being more men than women at universities where test participants are often recruited, (2) paternalistic beliefs about women being objects that need to be protected at home and excluded from the political and economic sphere, and (3) concerns about clinical trials having negative effects on womens' fertility and on unborn fetuses. Conducting clinical trials using only male participants has historically been seen as an obvious and unproblematic choice in the context of the background assumption that the male body is the standard, while female bodies are deviations from the standard.

The consequence of this similarity judgment is that inductions about male nutritional needs, disease symptoms, and drug efficacy are taken to be generalizable to women. This has resulted in women being 50% more likely to have a misdiagnosed heart attack than men [Hamid 2024], underdiagnosis of anemia in women [Levi 2019], and women experiencing dosage mistakes, side effects, and reduced efficacy in drugs whose pharmaceutical trials focused on males.

The concept of race was socially constructed by selecting and emphasizing arbitrary similarities and differences between humans' superficial physical features. Biased investigation of these similarity judgments led to these creation of biased similarity categories which were then used to justify scientific racism in the 19th and 20th centuries. There are numerous

other examples in archaeology, anthropology, medicine, and biology in which gender bias, racial stereotypes, and other social values have shaped the norms for relevant similarity in a given domain, which have in turn shaped what inductive inferences are possible.

To constructively mediate the role of social values, a successful normative account of similarity should include values from social epistemology. These could be values like requiring that similarity judgments be explicitly expressed and public, so that they can be subject to debate and scrutiny [Longino 2008]. In chapt. 1, I showed that when scientists disagree on the plausibility of an analogical inference, their disagreement often lies in disagreement about what constitutes a relevant similarity. However, the scientific communities in question are often unaware of the background assumptions that are motivating their norms for relevant similarity. Weisber says,

> In many cases theorists won't necessarily articulate the grounds for the judgments of similarity – the judgments are just made. Nevertheless, when it matters, such as in cases of disagreement, theorists should be able to work out the grounds for their similarity judgment [Weisberg 2013].

Values to constructively manage psychological biases

Another consideration in the toolkit should be to incorporate findings from psychology and anthropology research on the nature of our similarity biases. Then, scientists should establish norms for similarity in those domains that constructively mediate those innate biases. For example, psychological studies on perception in radiology diagnosis has found that radiologists regularly demonstrate (1) inattention blindness and (2) fail to recognize features that diverge from their expectations of what they will find in their analysis of a particular radiographic image [Drew 2013]. One study found that radiologists fail to notice a large image of a gorilla in the middle of the radiographic image while analyzing the image for lung

nodules [Drew 2013]. To mediate this bias, some teams now have multiple radiologists analyze an image, which seems to significantly increase the probability of an anomalous feature on the image being detected [Drew 2013].

To illustrate how naturalistic investigation can refine the reliability of our similarity judgments, I will use two examples. One describes a more abstract similarity bias that operates across different domains. The second is a perceptual similarity bias that has been of concern in the domain of cancer screening (different from the inattention blindness case discussed above). I will discuss the abstract example first.

Example 4.4.1. Cultural learning

Bayes' theorem provides a restriction on the updating of beliefs that prevents some cases of irrational belief-updating behaviour when encountering new evidence. In situations where Bayes' theorem can be applied, updating one's beliefs in a way that is inconsistent with Bayes theorem is irrational, in that one would be accepting a Dutch book bet [Barrett 2024]. In a variety of studies, it has been found that in both simple and complex environments, non-human animals are adept at updating their beliefs in a manner that is consistent with Bayes' theorem [Valone 1992, 2006]. Bayes' theorem predicts different results for rational behaviour in variable versus non-variable environments. For instance, in a foraging environment with berry patches of variable quality—some very rich patches but many poor quality berry patches–behaving in a way that rationally accords with Bayes' theorem will be different than behaving rationally in a low-variety environment where all of the patches are of a similar size and quality.

The concept is similar to the urn example I discussed in sect. 1. In a variable environment, when a forager encounters food in a patch, Bayes' theorem predicts that their estimate of the quality of the patch should increase because finding food should increase their credence that they are in a rich patch. Therefore, even though finding a berry and eating it has

decreased the total number of berries in the patch, the fact that the patch has one less berry is outweighed by the fact that the likelihood of this being a rich patch has increased. In contrast, in a low-variance environment populated by poor patches, finding a berry in a patch and eating it should not increase the forager's credence that they have hit upon a rich patch, because rich patches are very unlikely. They also have the information that they have already removed and eaten one of the berries in this patch. Therefore, their estimate of the quality of this current patch should decline after finding and eating a berry from this patch. Non-human animals demonstrate behaviour that is consistent with Bayesian rationality in both high-variance and low-variance environments [Valone 2006]. For instance, studies have found that animal behaviour in both low-variance and high-variance environments is consistent with the predictions of Bayes' theorem in parakeet foraging patterns [Ghiraldeau 1993], crustacean mating behaviour [Hunte et al. 1985], and female blue razorfish mate choice [Luttbeg 1999].

Like these non-human animals, humans display reasoning that approximates the predictions of Bayes' theorem in certain scenarios, such as when interpreting visual input [Harrison 2023] and when reasoning about low-variance environments [Valone 2006]. However, humans are unique from many other non-human animals in that our reasoning does not update according to Bayesian norms in some situations, particularly when environmental stimuli is highly variable [Valone 2006, Kahneman et al. 1974, Pinker 2010]. Valone [2006] says, "These deviations...result from evolved tendencies by which humans systematically fail to learn meaningfully from high-variance environmental feedback [Valone 2006]."

This finding is surprising given that human brains are inordinately large and this adaptation has had extremely costly effects like increased altriciality. We would expect that if encephalization (evolving large brains) is so evolutionarily costly, then the fact that we evolved large brains means that large brains confer some tremendous adaptive increase in ability to reason in a way that will confer maximum payoffs. Bayes' theorem is supposed to offer a constraint on reasoning, such that a reasoner who violates Bayes' theorem commits themselves to a

Dutch book, a series of bets that guarantee a loss [Barrett 2024]. Why is our reasoning about variable environments less Bayesian than that of other animals?

Anthropologists and cognitive scientists theorize that the answer is social learning. Large brains allow us to learn, teach, and remember domain-specific knowledge that gets passed down through generations and refined over multiple generations [Valone 2006]. Cognitive scientists claim that the evolutionary pressures that shaped our cognition selected for the ability to acquire culturally evolved knowledge [Hrdy 2014, Boyd et al. 1985].

Culturally evolved knowledge is often more valuable than information learned directly from an agent's environment, for several reasons. First, it encodes longer spans of patterns of payoff results than one agent can learn in their lifetime. Second, it can encode knowledge that an agent would be in-principle unable to incorporate into their knowledge structure through direct interaction with the environment, such as that that mushroom over there is deadly poisonous. This information would never be learned if social learning were unavailable, because an agent who tested the mushroom would die, and thus not be able to update their beliefs after eating the mushroom. Other information that a culturally evolved strategy could contain that could never be acquired through individual learning are (1) infrequent data points, like the information that cicadas emerge every seventeen years, and (2) non-obvious strategies that have a low probability of being discovered. For example, some science historians argue that if penicillin had not been discovered by Fleming by accident, it could be that it would still be undiscovered and that we would currently rely on sulphonamides or other antibacterials instead [Alharbi et al. 2014]. With social learning, agents can have knowledge as part of their repertoire that they would never have chanced upon in their lifetime through individual experimentation in their environment.

Finally, social learning improves efficiency, particularly in variable environments. In variable environments, the number of strategies and payoffs that an agent has to keep track of are much larger than in non-variable environments. Encoding strategy and payoff results in a

high-variability environment carries high fitness costs [Valone 2006]. Accurately encoding the information in the environment takes time and other cognitive resources away from other other important activities, like watching for predators. Thus, in a variable environment, if you are an organism who has the cognitive capacity for social learning, it is more reliably adaptive to follow a template that has been culturally inherited rather than to attempt to develop a template from scratch from the limited data set of your own interactions with the environment.

As a result, studies in cognitive science suggest that humans have evolved an innate tendency to *under-attend* to environmental payoffs and *over-attend* to domain specific knowledge that they have acquired through social learning. When we have acquired a credence through social learning, we have a bias to not update our credence with incoming environmental data that conflicts with the culturally inherited credence [Valone 2006].

This evolved disposition amounts to a similarity judgment bias. Suppose that an agent has acquired knowledge from social learning about a particular type of berry and how to forage it. From imitating a more experienced agent, they learn that the berries should only be picked when they are light red and not when they are purple. From imitating the role model, the agent learns that light red berries belong to an A-type set that are correlated with the B-type set 'good to eat' in the context of foraging selection.

After learning this rule, the agent is not perfect in their attempts to judge similarity. By accident, they eat two or three purple berries. They don't notice any ill-effects from eating the berries. If they were to attend to this environmental feedback and update their credence in accordance with Bayes' theorem, this would amount to updating their similarity judgment category. Purple berries that are perceptually similar to the ones they ate with no ill-effects would begin to be associated by a similarity relation with the light red berries (on a category level, not a perceptual level).

Is is rational for the agent to allow environmental payoffs to affect their similarity judgments in accordance with the standard formulation of Bayes' theorem? If the agent has a strong credence about the role model's reliability, then it is likely more adaptive for them to practice high-fidelity imitation of the role model's similarity judgments rather than innovating their own similarity categories. Consider a few options that could have caused the foraging of only light-red berries to be selected for and culturally inherited. One option is that the berries, when overripe and purple, can harbor dangerous parasites. However, it takes several weeks for the parasites to have an effect on the organism who ingested them. Therefore, it is unlikely that an agent, through individual learning, would be able to make the connection between the purple berries and the category 'not good to eat'. They would merely know that they are sick for whatever reason but they would not be able to connect the negative payoff to having eaten the berries.

Another option is that only four to seven percent of berries have parasites. However, the parasites are very severe if an organism does eat berries with the parasites. In this instance, deviating from the domain-specific socially learned similarity judgment and using environmental payoff to update their beliefs will lead to them eating more purple berries, further updating their belief that purple berries are similar to those berries which are associated with the 'good to eat' outcome. This will lead to more purple berry eating behaviour and further similarity category updates until they eventually eat one of the berries belonging to the small percentage that does have parasites. Innovation will lead to getting the parasites while rote imitation results in not getting parasites.

Perhaps that on very dry years, the rate of parasites drops low enough that it is adaptive to eat purple berries. This payoff information could be encoded and passed down through cultural evolution, but would be unlikely to be discovered by the individual agent.

Finally, the agent has limited cognitive resources. Any attention it allocates to innovating a new berry gathering strategy instead of faithfully following the culturally inherited strategy

is attention that is taken away from other cognitive tasks. Some of the agent's cognitive tasks involve less variable environments where individual innovation is more adaptive than in highly variable environments. By ceding to the authority of the role model in the similarity judgments for this category, the agent can allocate their attention to innovating strategies in less variable environments.

In this instance, where the role model does have a successful method, then betting against the role model would be a maladaptive strategy for the agent. It is better for them to under-attend to the environmental payoffs and prioritize the role model-sanctioned similarity relations rather than responding to the environmental feedback and altering their similarity judgments to align with the environmental associations they experience.

Cognitive scientists use Bayesian models to show that it is likely that humans have responded to our unique capacity for social learning by evolving an innate disposition to under-attend to environmental payoffs in high-variation environments.

However, while this disposition evolved because it was adaptive for our evolutionary ancestors, it likely is the source of similarity biases in our scientific reasoning that are not epistemically fruitful. Consider the example of continental drift.

Example 4.4.2. In the 1500s, a prominent view among geologists was catastrophism, the theory that Earth's history has included many major, catastrophic events like floods and earthquakes and that the Earth's geologic features are products of these huge events. Proponents of this view noted that Europe, America, and Africa all had a similar shape along the coasts that faced towards each other, and proposed that this was further evidence for catastrophism. Ortelius, a Dutch cartographer, said that the continents must have been "torn away from [each other] by earthquakes and floods...[Ortlelius 1596]." Couched in the theory of catastrophism, the similarity between the shapes of neighboring coasts seemed telling and apparent. This view remained prominent for several hundred years, and geologists agreed

that there was a notable similarity in the shapes of the coasts that must be indicative of some deeper similarity. However, in the early 1800s, an influential proponent of uniformitarianism, James Dwight, criticized the theory that the continents had been torn apart. He likely saw the theory as inextricably tied to catastrophism. The idea that the continents had been torn apart began to fade in the early 1800s as catastrophism shifted to uniformitarianism. However, in the fringes, catastrophic theorists still argued that the continents had once been conjoined, and proposed less biblical explanations for the rift, like volcanic activity. In the early 1900s, Wegener followed this tradition and proposed the theory of continental drift, though he did not offer a definitive explanation for the drift, and merely postulated a few suggestions, such as that the continents had been pulled apart by the centrifugal force of the planet's rotation. In evidence for his theory, he pointed out the similarities in shape of the neighboring coasts, and the fossil records and geological properties that these coasts shared. His claim was that an adequate theory should explain the existence of these similarities. His theory was largely rejected, probably because the unpopularity of catastrophism left the observed similarities unwedded to any convincing overarching narrative. Chamberlin said that Wegener's theory of continental drift ""takes considerable liberties with our globe...If we are to believe Wegener's hypothesis we must forget everything which has been learned in the last 70 years and start all over again [Bullard 1975]."

Scientists judged the similarity of the coastlines, fossil record, living flora and fauna, and geological features as being irrelevant and not evidence for updating the theory. Wegener's evidence did not support the dominant theory, uniformitarianism, and because he was unable to provide a convincing mechanism for how the continents might have drifted, there was not a framework to fit the postulated similarities into.

Rejecting an established, culturally inherited theory that makes correct predictions in favor of updating our beliefs to accord with new environmental evidence goes against our innate disposition to prioritize culturally inherited rules over innovation. Yet this example shows

that adherence to this bias can cause scientists to make errors in their similarity judgments. Geologists dismissed the similarities of the coastlines, geological features, species, and fossils as being irrelevant and accused Wegener of gerrymandering the similarities. On one hand, this kind of epistemic vigilance is an epistemic virtue. Another bias that we have is the tendency to overgeneralize from similarities. One example is the homeopathic belief 'the principle of similitude', the belief that substances which cause symptoms in healthy individuals must also cure those symptoms in sick individuals. Requiring a satisfactory account of an underlying mechanism, as critics of Wegener did, is a good way to guard against overgeneralizing from an observed similarity. On the other hand, in the epistemic process of science, understanding of mechanisms and a dominant theory that explains a theory that often follow quite a ways behind the identification of a relevant similarity. Some of the most innovative discoverers (Einstein, Mendeleev, Darwin) are those who cling to the observation of a similarity long enough for a mechanism to be discovered. Regardless, being aware that we have this bias can help us develop a more principled way to carry out this epistemic vigilance.

Example 4.4.3. Serial Dependence

A more simple perceptual similarity bias is of concern in medicine. Researchers have found that when radiologists evaluate mammograms for markers of breast cancer, they are affected by a visual similarity bias that causes them to under-diagnose cancerous tissue. Cancer causes clusters of calcification spots in the breast tissue, and evaluation involves identifying these clusters, which appear as white dots on the x-ray images. However, only clusters of calcification are markers of cancer; more diffuse distributions of calcification spots across the breast are normal and not associated with cancer.

When evaluating images, radiologists have to differentiate between clustered and diffuse calcification patterns. However, a visual bias causes radiologists to report many clustered spots as diffuse. Studies have shown that this visual bias is equally present in trained

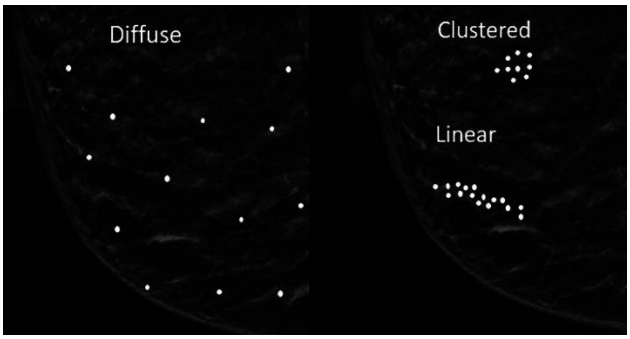


Figure 4.6: Clustered versus diffuse calcification [Witt 2022]

radiologists as it is in the general public. In a psychology study, it was found that this bias leads to false negatives ranging from one to fifteen percent [Witt 2022].

As a result, scientists are working on designing AI systems that replace radiologists in analyzing mammograms.

Understanding the nature of the bias has potential to help scientists anticipate and adjust for this bias in other fields. There are two possible evolutionary explanations that would explain the nature and cause of the bias. One possibility is that the bias is a case of serial dependence.

Serial dependence is a similarity bias in which an agent observing a stimulus perceives it as being more similar to a previously observed stimulus than it actually is. It as though the perceptual system takes an average of the past observed stimuli and uses this to represent the current stimuli. Visual experience is not a product of sensory input alone, but also memory of past visual experiences. One example where this can be seen is in agents' visual perception of an ambiguous image with black and white abstract shapes. After agents are told that it is an image of a dog, most are able to see a dalmatian in the ambiguous shapes. However, after gaining this perspective, they are unable to go back to seeing the image as a collection of abstract shapes; they intractably see the dog. This effect can last for a long time, where even when shown the image years later they see the dog instead of the abstract shape they initially perceived [Cicchini 2024].

One possibility, which has not been considered in the literature, is that serial dependence could explain the clustering bias in mammogram evaluation. The majority of mammogram images do not have clustered calcification, so radiologists conducting evaluation see non-clustered tissue over and over again all day. This could trigger the serial dependence bias and make radiologists more likely to perceive clustered tissue as non-clustered. The effect of the serial dependence bias is almost like an average generated over our past observations that our visual system uses to construct our current visual experience. If the majority of the input for that average function is diffuse calcification, then this would increase the likelihood of new perceptions being perceived as also diffuse.

In an evolutionary history in which an agent's environment was generally stable, it is adaptive to have the brain develop shortcuts to ensure that similar behaviours are elicited by similar stimuli. Then past learning will be utilized for reacting to new situations. Thus, when constructing perceptual experiences, the brain makes current stimuli more similar to past stimuli than they actually are. Game theoretic models have shown that this is an adaptive trait and does lead to higher fitness than veridical representations in stable environments [Burr et al. 2018]. Having this bias occur at a preconscious level and rely on simple, rule-of-thumb heuristics rather than deliberate comparison of past and present stimuli is adaptive because it reduces response times and preserves agents' attentional resources. Selection pressures favor simple similarity heuristics that make fast, approximately true predictions in stable environments, not costly heuristics with higher accuracy and reliability in unstable environments, which were less common in our evolutionary history.

In summary, a satisfactory naturalistic normative account of similarity must investigate the psychological heuristics and biases that fix our similarity judgments, and investigate these innate heuristics. Part of understanding these innate heuristics involves consulting evolutionary narratives to understand the types of evolutionary contexts our similarity judgments evolved to operate in.

Figure 4.7: One-shot learning

4.5 Conclusion

Scientific norms for similarity are determined by the de facto epistemic commitments of scientists within a particular domain. These norms are revised over time through domain-specific empirical investigation. Thus, no once-and-for-all characterization or justification of relevant similarity is possible. What a normative account of similarity can contribute to science is to enrich the epistemic toolkit that scientists use to investigate norms of similarity within their domains. The process by which scientists fix norms for similarity is affected by cultural norms, innate psychological dispositions [Quine 1969], local background knowledge, and scientists' goals, among other factors. The goal of a normative account should be to attend to research in natural science that investigates how these variables affect humans' similarity judgments. A normative account should formulate a pluralistic set of epistemic values that enrich the epistemic toolkit that scientists have at their disposal when judging similarity. These epistemic values must be updated as the research on how humans make similarity judgments advances.

Chapter 5

Conclusion

5.1 Conclusion

This dissertation has proposed a framework for a middle ground between formal normative accounts of analogical inference, like those proposed by Bartha, Hesse, and Gentner, and material accounts like Norton's. While the relevant similarity framework accounts fail to deliver necessary and sufficient conditions for analogical plausibility, these accounts do capture patterns that tend to characterize fruitful analogical inference. Norton is right to claim that evaluation of analogical inferences is local and empirical, but giving up the normative project would risk collapsing into a form of naive empiricism.

To avoid this, I advocate for a pluralistic normative account that acknowledges that norms are locally determined but does not take this as justification for giving up the normative project. I argue that a successful normative account of analogical inference, rather than providing necessary and sufficient conditions, should provide a pluralistic set of epistemic values that tend to characterize fruitful scientific enquiry.

I have proposed several such epistemic values. I showed that under current normative accounts, two conditions, *Source Domain Salience* and *Overlap*, are thought to be jointly necessary and sufficient for *prima facie* plausibility. I show that in practice, scientists also consider a third criterion which I call *Target Domain Salience*. In Chapter 2, I showed that in many cases in which scientists disagree about the plausibility of an analogical conjecture, scientists agree that the analogical inference satisfies *Source Domain Salience* and *Overlap*. In these cases, the source of their disagreement often lies in *Target Domain Salience*, which often consists of implicit background assumptions.

These implicit background assumptions fall outside the scope of *Source Domain Salience* and *Overlap* and so fail to be included in the explicit articulation of an analogical inference under the relevant similarity framework accounts. This is a significant problem for current accounts of analogical inference because the plausibility of an analogical inference often hinges on these

implicit assumptions. *Target Domain Salience* can include assumptions about causality, relevant similarity, and the stability of the relationship in the source domain across time and different contexts [Currie 2016]. However, scientists rarely articulate these *Target Domain Salience* assumptions when presenting analogical inferences, which can obscure the actual grounds for the inference and make it difficult to assess its plausibility. Including *Target Domain Salience* as a criterion in normative accounts makes these background assumptions explicit and allows them to be subjected to critical evaluation and considerations from social epistemology.

In Chapter 3, I challenged an assumption made by current relevant similarity framework accounts, the methodological rejection of representational gerrymandering. Current normative accounts assume that gerrymandering of the source domain to increase the degree of similarity between a source and target domain is an unscientific and unjustifiable use of analogical inference. If scientists can arbitrarily select representations that will make an analogical conjecture go through, then this threatens the integrity of analogical inference as a method of induction. If normative accounts permit this, then an analogical conjecture can appear plausible not because there is a genuine relevant similarity between the source and target, but because the domains being compared have been selectively represented in order to engineer artificial similarities between the source and target.

In order to prevent undesirable manipulation from occurring, Bartha and Nappo propose an umbrella rejection of reverse analogical inferences as a valid form of analogical inference.

I show that this umbrella rejection of reverse analogical inferences is counterproductive. The fact that scientists can always choose between making a forward analogical inference or a reverse analogical inference facilitates a process of reflective equilibrium between representation choice and analogical conjecture. This is an important mechanism to prevent problematic cases of entrenchment. The analogical inference literature has long considered the problem of dogmatic entrenchment in analogical inference. Bartha notes that when background as-

sumptions and norms for relevant similarity become entrenched, scientists can become biased towards making analogical inferences that confirm entrenched predicates. This can prevent scientists from making novel comparisons and uncovering plausible unconceived alternatives [Bartha 2009, pg. 12]. This is particularly concerning because the predicates and background assumptions used in analogical inferences can self-perpetuate. Stepan says,

> Because a metaphor or analogy does not directly present a pre-existing nature but instead helps "construct" that nature, the metaphor generates data that conform to it, and accommodates data that are in apparent contradiction to it, so that nature is seen via the metaphor and the metaphor becomes part of the logic of science itself [Stepan 1996].

When a set of background assumptions are entrenched in a scientific community, analogical inferences run the risk of being replication machines for entrenched predicates. This is problematic because analogical inferences are an important tool for generating new hypotheses and discovering unconceived alternatives, and the entrenchment problem threatens their ability to do this.

In Chapter 3, I argue that the ability of scientists to switch inference direction from forward analogical inference to reverse analogical inference in response to their local epistemic commitments is an epistemic virtue. It facilitates a process of reflective equilibrium in which representations, norms for relevant similarity, and particular analogical conjectures, are mutually adjusted. Rather than an entrenched representation being accepted uncritically, it is weighed against scientists' other epistemic commitments in the context of the analogy. Analogical inferences act as a mechanism for maintaining the coherence of scientists' belief system while also responding to new evidence. Then, even a deeply entrenched representation may be revised if doing so would yield an analogical conjecture that scientists are epistemically committed to or have strong intuitions about. Reverse analogical inferences

are circular, but they can be virtuously circular, and are a tool for avoiding bias and the vicious circularity of dogmatic entrenchment.

Another epistemic virtue of reverse analogical inferences is that they can help scientists uncover strong reasons for finding an analogical conjecture plausible that might be initially epistemically inaccessible to them. The psychology literature shows abundant evidence that scientists often lack epistemic access to the factors that caused them to find an analogical conjecture plausible [Dunbar 2000]. Thus, while current accounts claim that analogical inferences are explicit representations of scientists' analogical reasoning processes, the reality is more likely to be that analogical inferences are *post hoc* constructions in which scientists cite paradigm-sanctioned reasons for why a scientist would find a given analogical inference plausible. The literature has numerous cases in which plausible analogical conjectures were founded on spurious analogical inferences that cited historically contingent reasons for accepting the analogical conjecture. In these cases, a scientists' analogical reasoning might be sound, even if their *post hoc* construction of that reasoning employs norms for similarity that are not accepted by contemporary scientists. I describe one possible case of this in Example 3.6.3. In these cases, reverse analogical inference can be seen as a search mechanism—a way to search for the strongest *post hoc* construction. Because there is a disconnect between analogical inference and analogical reasoning, when critical differences are found that invalidate an analogical inference, scientists do not know whether their analogical reasoning was implausible, or just their analogical inference was implausible. Thus, it is coherent for them to use reverse analogical inference to see if another plausible representation supports their analogical conjecture.

Finally, in Chapter 4, I argued that refining analogical reasoning requires sensitivity to the natural science literature on how humans form similarity judgments. I argued that a normative account of analogical inference should include investigation of the innate biases and subconscious inductive inferences that affect scientists' similarity judgments. For instance,

cultural norms and innate psychological mechanisms like the *serial dependence* and *cultural learning* biases shape scientists' similarity judgments. Furthermore, these biases often operate below the threshold of conscious awareness, such that scientists' similarity judgments can be affected by contextual factors without scientists having conscious awareness that these factors are affecting their similarity judgments.

In summary, although no universal logic for analogical inference is possible, it is possible for a normative account to provide useful epistemic norms that are grounded in considerations from psychology and social epistemology. My dissertation proposes an account that is a middle ground between the formal, relevant similarity framework accounts and material accounts like Norton's. The norms I postulate are not necessary and sufficient conditions for plausibility, but rather are values that scientists must weigh according to their local epistemic commitments. Having these epistemic values in their toolkit provides scientists tools for debating about plausibility, examining their background assumptions, incorporating considerations from social epistemology, and using the natural science literature to extrapolate epistemic norms for evaluating relevant similarity.

Bibliography

[1] Z. Adam et al. Estimating the capacity for production of formamide by radioactive minerals on the prebiotic earth. 2018.

[2] C. Barranco. The first live attenuated vaccines. *Nat Milestones*, 284:S7, 2020.

[3] P. Bartha. *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford University Press, 2009.

[4] P. Bartha. Analogy and Analogical Reasoning. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2013.

[5] P. Bartha. Analogy in the natural sciences: meeting Hesse's challenge. 2015.

[6] P. Bartha. Norton's material theory of analogy. 2020.

[7] K. Bora et al. Cd-hpf: New habitability score via data analytic modeling. *Astronomy and Computing*, 17:129–143, 2016.

[8] A. Chawda, E. Guinan, and S. Engle. A Better Earth? The Age and X-UV Irradiance of the Superhabitable Earth-Like Exoplanet, KOI-4878.01. *Bulletin of the AAS*, 56(2), feb 7 2024. https://baas.aas.org/pub/2024n2i178p03.

[9] N. Crawford et al. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. 2012.

[10] A. Currie. Ethnographic analogy, the comparative method, and archaeological special pleading. *Studies in History and Philosophy of Science Part A*, 55:84–94, 2016. Special Section - Historiography and the Philosophy of the Sciences Special Section - Scientific Knowledge of the Deep Past.

[11] A. Currie. Ethnographic analogy, the comparative method, and archaeological special pleading. *Studies in History and Philosophy of Science Part A*, 55:84–94, 2016. Special Section - Historiography and the Philosophy of the Sciences Special Section - Scientific Knowledge of the Deep Past.

[12] F. D'Abramo and S. Neumeyer. A historical and political epistemology of microbes. *Centaurus; international magazine of the history of science and medicine*, 62, 2020.

[13] C. Darwin and G. Beer. *On the Origin of Species*. Oxford University Press, New York, revised edition, 2008. Original Date: 1859.

[14] S. De Baerdemaeker. Method-driven experiments and the search for dark matter. *Philosophy of Science*, 88(1):124–144, 2021.

[15] J. R. Denbow. The toutswe tradition: a study in socio-economic change. *Settlement in Botswana*, 73:86, 1982.

[16] A. S. Dias. *Analogy in Archaeological Theory*, pages 297–301. Springer International Publishing, Cham, 2014.

[17] H. E. Douglas. Values in science. In P. Humphreys, editor, *The Oxford Handbook of Philosophy of Science*, pages 609–630. 2014.

[18] A. Einstein. Letter from Albert Einstein to Moritz Schlick. 1915/2005.

[19] A. Einstein. *The evolution of physics: the growth of ideas from early concepts to relativity and quanta*. Simon and Schuster, 1938.

[20] A. Einstein, H. Gutfreund, and J. Renn. *Relativity: The Special and the General Theory*. Princeton University Press, 100th anniversary edition edition, 2015/1917.

[21] Einstein, A; et al. *The Principle of Relativity*. Dover Publications, New York, reprint edition, 1952. A collection of original memoirs on the special and general theory of relativity.

[22] S. Evers and R. Benson. A new phylogenetic hypothesis of turtles with implications for the timing and number of evolutionary transitions to marine lifestyles in the group. 2019.

[23] A. Fages, A. Seguin-Orlando, M. Germonpré, and L. Orlando. Horse males became over-represented in archaeological assemblages during the bronze age. *Journal of Archaeological Science: Reports*, 31:102364, 2020.

[24] K. J. Fewster. The potential of analogy in post-processual archaeologies: A case study from basimane ward, serowe, botswana. 2006.

[25] P. Feyerabend. *Against Method*. New Left Books, London, 1993.

[26] S. T. Fiske, A. J. Cuddy, and P. Glick. Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, 11(2):77–83, 2007.

[27] M. Fox-Powell et al. Ionic strength is a barrier to the habitability of mars. 2016.

[28] M. Fujita et al. Turtle phylogeny: insights from a novel nuclear intron. 2004.

[29] B. Genta. How to think about analogical inferences: A reply to Norton. *Studies in History and Philosophy of Science Part A*, 82:17–24, 2020.

[30] D. Gentner. Structure-mapping: A theoretical framework for analogy. 1983.

[31] D. Gentner and Day. Nonintentional analogical inference in text comprehension. *Memory Cognition*, 35.

[32] S. Ghirlanda et al. A century of generalization. *Animal Behaviour*, 66, 2003.

[33] T. Gilovich. Seeing the past in the present: The effect of associations to familiar events on judgments and decisions. *Journal of Personality and Social Psychology*, 40, 1981.

[34] N. Goodman. *Fact, fiction, forecast*. Harvard University Press.

[35] J. Grabenstein et al. Immunization to protect the US armed forces: Heritage, current practice, and prospects. 2006.

[36] M. H and S. D. Why do humans reason? arguments for an argumentative theory. *Behav Brain Sci*, 2011.

[37] G. S. Halford and G. Andrews. *Reasoning and Problem Solving*, chapter 13. John Wiley Sons, Ltd, 2007.

[38] H. Hanson. Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology*, 58(5):321–334, 1959.

[39] R. Hassin. Making features similar: Comparison processes affect perception. *Psychonomic Bulletin Review*, 8, 2001.

[40] M. Hesse. Models and analogies in science. *Philosophy and Rhetoric*, 3(3), 1966.

[41] I. Hodder. *The domestication of Europe: structure and contingency in neolithic societies*. Wiley, United Kingdom, 1990.

[42] D. R. Hofstadter. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. New York, 2013.

[43] K. J. Holyoak. et al. Bayesian generic priors for causal learning. *Psychological Review*, 115(4), 2008.

[44] K. J. Holyoak and P. Thagard. *Mental leaps: Analogy in creative thought*. The MIT Press., 1995.

[45] S. B. Hrdy. *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Harvard University Press, 2009.

[46] P. Hristova, P. Bg, B. Kokinov, and B. Bg. Anxiety can influence analogy-making both positively and negatively depending on the complexity of the mapping task. 08 2013.

[47] T. Huffman. Broederstroom and the central cattle pattern. *South African Journal of Science*, 89:220–226, 1993.

[48] T. Huffman. The central cattle pattern and interpreting the past. *Southern African Humanities*, 13(1), 2001.

[49] T. Huffman et al. Caddoan archaeology on the high plains: A conceptual nexus of bison, lodges, maize, and rock art. *American Antiquity*, 79(4):655–678, 2014.

[50] T. N. Huffman. Archaeology and ethnohistory of the African Iron Age. *Annual Review of Anthropology*, 11:133–150, 1982.

[51] T. N. Huffman. Broederstroom and the origins of cattle-keeping in southern Africa. *African Studies*, 1990.

[52] D. Hume. *A Treatise of Human Nature*. Oxford University Press, 1739.

[53] D. Hume. *Enquiries concerning Human Understanding and concerning the Principles of Morals*. Clarendon Press, Oxford, 3rd edition, 1975.

[54] W. Isaacson. *Einstein: His Life and Universe*. Simon & Schuster, United States, 2017.

[55] V. J. The willow as a hottentot (khoikhoi) remedy for rheumatic fever. *Journal of the Royal Society of Medicine*, 101, 2008.

[56] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[57] I. Kant. Universal natural history and theory of the heavens or essay on the constitution and the mechanical origin of the whole universe according to Newtonian principles (1755). In E. Watkins, editor, *Kant: Natural Science*, pages 182–308. Cambridge University Press, Cambridge, 2012.

[58] B. L. Beijerinck's work on tobacco mosaic virus: historical context and legacy. *Philosophical transactions of the Royal Society of London*, 354, 1999.

[59] P. J. Lane. The use and abuse of ethnography in the study of the southern African Iron Age. *Azania*, 5, 1994.

[60] P. J. Lane. Reconstructing Tswana townscapes: toward a critical historical archaeology. In *African Historical Archaeologies*. Routledge, London, 2004.

[61] P. J. Lane. Barbarous tribes and unrewarding gyrations? the changing role of ethnographic imagination in African archaeology. In A. B. Stahl, editor, *African Archaeology: A Critical Introduction*, pages 24–54. Blackwell, Oxford, 2005.

[62] M. Lee. Reptile relationships turn turtle. *Nature*, 389:245, 1997.

[63] T. Lyson et al. Micrornas support a turtle + lizard clade. *Biology Letters*, 2012.

[64] N. Maier. Reasoning in humans: The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, 1931.

[65] A. Mayer, D. Ivanowski, M. W. Beijerinck, and E. Baur. *Early Papers on Tobacco Mosaic and Infectious Variegation*, pages 9–62. 1886/2018.

[66] J. Mazon et al. Aircraft clouds: From chemtrail pseudoscience to the science of contrails. *Metode Science Studies Journal*, 8:181–187, 2018.

[67] D. Medin, R. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 1993.

[68] D. Mendeleev. On the relationship of the properties of the elements to their atomic weights. *Zeitschrift für Chemie*, 12:405–406, 1869.

[69] D. Mendeleev. The relation between the properties and atomic weights of the elements. *Journal of the Russian Chemical Society*, 1:60–77, 1869.

[70] F. Nappo and N. Cangiotti. Reasoning by analogy in mathematical practice. *Philosophia Mathematica*, 31(2):176–215, 2023.

[71] I. Newton. *Newton's Principia: the mathematical principles of natural philosophy*. Daniel Adee, New York, 1846.

[72] R. E. Nisbett and T. D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 1977.

[73] J. D. Norton. Thought experiments in Einstein's work. In *Thought Experiments In Science and Philosophy*. Rowman and Littlefield, Savage, MD, 1991.

[74] J. D. Norton. *The Material Theory of Induction*. University of Calgary Press, Canada, 2021.

[75] S. Olsen. Documenting domestication: New genetic and archaeological paradigms, eds zeder ma, bradley dg, emshwiller e, smith bd, 2006.

[76] S. Olsen. Early horse domestication on the eurasian steppe. *Documenting Domestication: New Genetic and Archaeological Paradigms*, pages 245–269, 06 2006.

[77] S. Olsen. Ancient genomes revisit the ancestry of domestic and przewalski's horses. *Science*, 360(6384):111–114, 2018.

[78] S. L. Olsen, S. Grant, A. M. Choyke, and L. Bartosiewicz. *Horses and humans: the evolution of human-equine relationships*. Archaeopress Oxford, UK, 2006.

[79] E. P. P. and E. R. G. The periodic law of the chemical elements: 'the new system of atomic weights which renders evident the analogies which exist between bodies'. *Royal Society Publishing*, 2020.

[80] J. P and Hall. Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. 2012.

[81] L. Pasteur. De l'attbnuation du virus du chokra des poules. *Comptes rendus de l'Academie des sciences*, 91:673–680, 1880.

[82] L. Pasteur. Sur les maladies virulentes, et en particulier sur la maladie appelee vulgairement cholera des poules. *Comptes Rendus de l' Acad. Sci.*, 90:249–248, 1880.

[83] G. Polya. *Induction and Analogy in Mathematics: Vol. 1 of Mathematics and Plausible Reasoning*. Princeton University Press, 1954.

[84] N. Price et al. Viking warrior women? reassessing Birka chamber grave. *Antiquity*, 93(367):181–198, 2019.

[85] W. V. Quine. *Natural Kinds*, pages 5–23. Springer Netherlands, Dordrecht, 1969.

[86] L. M. Reder and C. D. Schunn. Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder, editor, *Implicit Memory and Metacognition*. Lawrence Erlbaum, 1996.

[87] A. Reid et al. Tswana architecture and responses to colonialism. *World Archaeology*, 28(3):370–392, 1997.

[88] L. N. Ross. Causal explanation and the periodic table. *Synthese*, 198(1):79–103, 2018.

[89] R. Saladino et al. Formamide and the origin of life. *Physics of Life Reviews*, 9(1):84–104, 2012.

[90] C. Sarkissian. Evolutionary genomics and conservation of the endangered przewalski's horse. *Curr Biol.*, 2015.

[91] H. Sato et al. Turtle skull development unveils a molecular basis for amniote cranial diversity. *Science Advances*, 2023.

[92] E. R. Scerri. *The Periodic Table, its Story and its Significance*. Oxford University Press, New York, Oxford, 2007.

[93] E. R. Scerri and J. Worrall. Prediction and the periodic table. *Studies in History and Philosophy of Science Part A*, 32(3):407–452, 2001.

[94] F. Schauer and B. A. Spellman. 2017. *University of Chicago Law Review*, 84.

[95] M. Schlick. Positivism and realism. *Synthese*, 7:478–505, 1948.

[96] D. Schulze-Makuch et al. In search for a planet better than Earth: Top contenders for a superhabitable world. *Astrobiology*, 20(12):1394–1404, 2020.

[97] K. T. Selvan. A revisiting of scientific and philosophical perspectives on maxwell's displacement current. *IEEE Antennas and Propagation Magazine*, 51(3):36–46, 2009.

[98] H. Shaffer et al. Tests of turtle phylogeny: molecular, morphological, and paleontological approaches. *Systematic Biology*, 1997.

[99] M. Shanks et al. Processual, postprocessual and interpretive archaeologies. In *Museums in the Material World*. 2007.

[100] A. Smirnova, Z. Zorina, and T. Obozova. Crows spontaneously exhibit analogical reasoning. *Current biology*, 2015.

[101] K. Smith. Wanted, an anthrax vaccine: Dead or alive? *Medical Immunology*, 4(1):5, 2005.

[102] K. Smith. Louis Pasteur, the father of immunology? *Frontiers in Immunology*, 3:68, 2012.

[103] N. Spence. Tobacco mosaic virus: One hundred years of contributions to virology, eds k. b. g. scholthof, j. g. shaw m. zaitlin, vii+256 pp. st paul, minnesota: The american phytopathological society. us 53.00($hardback$).$isbn$0890542368. $Journal of Agricultural Science - JAGRSCI, 136 : 253 - -255, 032001$.

[104] K. Stanford. University of California, Irvine, 2024. Naturalized Epistemology [Powerpoint].

[105] U. E. Stegmann and F. Schmidt. Homology judgements of pre-evolutionary naturalists explained by general human shape matching abilities. *Scientific reports*, 13, 2023.

[106] A. Streets and H. England. Colour vision in stomatopod crustaceans: more questions than answers., 2022.

[107] W. Sullivan. Extraterrestrial life as the great analogy, two centuries ago and in modern astrobiology. In *Astrobiology, History, and Society*, pages 73–. 2013.

[108] Taylor. Rethinking the evidence for early horse domestication at botai. *Sci Rep*, 11, 2021.

[109] J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785, 2007.

[110] M. Zaitlin. *The Discovery of the Causal Agent of the Tobacco Mosaic Disease*, pages 105–110.