## *A Tool-Based View of Theories of Evidence*

CHIEN-YANG HUANG

# A Tool-Based View of Theories of Evidence

Chien-Yang Huang

Thesis submitted for the degree of
Doctor of Philosophy


Department of Philosophy
Durham University
2019

# A Tool-Based View of Theories of Evidence

Chien-Yang Huang

## Abstract

Philosophical theories of evidence have been on offer, but they are mostly evaluated in terms of all-or-none desiderata — if they fail to meet one of the desiderata, they are not a satisfactory theory. In this thesis, I aim to accomplish three missions. Firstly, I construct a new way of evaluating theories of evidence, which I call a tool-based view. Secondly, I analyse the nature of what I will call the various relevance-mediating vehicles that each theory of evidence employs. Thirdly, I articulate the comparative core of evidential reasoning in the historical sciences, one which is overlooked in major theories of evidence.

On the first mission, I endorse a meta-thesis of pluralism on theories of evidence, namely a tool-based view. I regard a theory of evidence as a purpose-specific and setting-sensitive tool which has its own strengths, difficulties and limitations. Among the major theories of evidence I have reviewed, I focus on Achinstein's explanationist theory, Cartwright's argument theory and Reiss's inferentialist account, scrutinising and evaluating them against the purposes they set out and the scope of their applications.

On the second mission, I note that there is no such thing as intrinsically 'being evidence'. Rather, I hold that relevance-mediating vehicles configure data, materials or claims in such ways that some of them are labelled evidence. I identify the relevance-mediating vehicles that the theories of evidence employ.

On the final mission, I argue that the likelihoodist account is an appropriate tool for explaining the evidential reasoning in poorly specified settings where likelihoods can be only imprecisely compared. Such settings, I believe, are typical in the historical sciences. Using the reconstruction of proto-sounds in historical linguistics as a case study, I formalise the rationale behind it by means of the law of likelihood.

## Statement of Copyright

# Acknowledgements

This long, challenging and intellectually enlightening journey is one which I will always cherish. I owe a great debt of gratitude to many people who have been involved in the production of this thesis.

I have been fortunate indeed to have an outstanding supervising team assisting me. My primary supervisor, Nancy Cartwright, has been my philosophy heroine and was the main reason that I came to Durham. She has been a guiding light from the outset. I wish to thank Julian Reiss for acting as a secondary supervisor and constantly stimulating my thoughts. I am very grateful to Alison Wylie, who has been an additional supervisor; she inspired and encouraged me in various ways. Special thanks also go to Wendy Parker for providing sound advice.

I must thank my defence committee, Peter Vickers and Roman Frigg, for their invaluable comments. I also greatly appreciate the feedback from the Centre for Humanities Engaging Science and Society (CHESS), including William Peden, Rune Nyrup, Sarah Wieten, Erin Nash and Donal Khosrowi. My thanks are also due to Terri Edwards, Rebecca Ward, Tim Cannam, Alicia Ward, Michelle Joubert, Jiunn Wang and Tse-Chun Wang for making this thesis better. Special thanks go out to Nelson Goering for checking the methodological part of historical linguistics.

I would like to sincerely thank my aunt Pai-Ping Lin, Szu-Ting Chen, Chi-Chun Chiu, Robert Schoonmaker, Simon James, and especially Anuk Chang, for all their kind help and support.

Finally, I am indebted to my parents, Chiao-Wang Huang and Pai-Lan Lin, for their endless patience, support and love. I dedicate this thesis to them.

# Table of Contents

# 1
# Introduction

## 1.1 Relevance-Mediating Vehicles

Messy, chaotic and disorderly is the typical impression of a crime scene. Police officers and investigators working at a crime scene have to seek materials, gather eyewitnesses and collect anything that is thought relevant. The materials could be duct tape, bloodstains or fibres, some of which will be sent to a laboratory for further identification. The materials, witness or expert testimony, video tape or anything found relevant will be brought to court. Across different stages, evidence detection, identification and comparison proceed back and forth. If the materials, testimonies and images are viewed as evidence in this case, the reason is not that there are things inherently labelled 'evidence' that await discovery. Instead, it is because certain kinds of *relevance-mediating vehicles* mould those things into a web of evidence.

Relevance-mediating vehicles can be construed as *configurers*. They configure materials or claims in such ways that some of them are labelled evidence. Evidence is realised within arrangements rather than existing in the form of mysteriously metaphysical entities: this kind of realisation is ubiquitous in our daily lives and scientific practices. Inspired by Nancy Cartwright's (1999, ch. 2) suggestion for an appropriate way of understanding abstract scientific concepts, I also regard multiple kinds of abstract concepts are realised through physical entities or activities. Stir-frying egg rice is a case in point, illustrating concepts functioning as shorthand for a set of relatively concrete concepts. We pour some oil into a preheated wok, and when the oil gets very hot, we crack an egg into the wok, add rice, and start to stir the mixture until completely heated through. Then we season the dish with soy sauce and sugar and stir for a further minute before bringing the dish to the table. *Stir-frying* is not an additional existing activity which is attached to, as well as operating independently of, other

smaller activities, including heating, adding oil, cracking an egg, adding rice, stirring and seasoning. Nevertheless, the stir-frying makes sense of the entire cooking process: it places the action of adding oil into the wok in the cooking context instead of in the context of maintaining a wok.

Similarly — though in a relatively complicated manner, regarding evidential relations between evidence and hypothesis — materials or claims can be configured via relevance-mediating vehicles that bring them together and establish their relevance. For example, the police want to know whether the victim was murdered by his relative, and at the crime scene they found a piece of duct tape with some bloodstains and distinct isotopic characteristics. There is no 'evidence' label on this piece of duct tape; for it to be evidence, a relevance-mediating vehicle is required to distil evidence from a web of facts. The relevance-mediating vehicles, depending upon different configurations and perspectives, could be a sound argument, a probabilistic model, an explanatory relation or some other discipline-specific vehicle.

Viewing evidence as a convenient shorthand can be understood as an epistemic view of evidence. Such a way of characterising important philosophical concepts is not uncommon. For example, Michael Wilde and Jon Williamson (2016) propose an epistemic view of causality. On this view, causality is 'purely epistemic in the sense that our causal claims enable us to reason and interact with the world in certain ways; they are not claims about some causal relation that exists independently of us and our epistemic practices' (Wilde and Williamson, 2016, p. 36). They use the *trihoral* relation as an analogy to explicate the epistemic view of causality. Suppose there is an arrive-within-three-hour destination map in which places represented by nodes link to one another with a line. Every single line means a three-hour journey. There is no real geographical distance between nodes, but the information on the map is useful for people to plan breaks. The three-hour relation between nodes is obtained based on physical positions, the journey time and some other relevant travel conditions. The trihoral relation does not refer to any worldly place that we call 'trihorality'. Likewise, in Wilde and Williamson's view, causal claims are useful for explanation, prediction and

control; this explains our having the concept of causality without reference to causal relationships existing independently of us. Causality represents ‘a complex array of facts about the presence and absence of mechanisms, as well as the presence and absence of difference-making relationships and their magnitude’ (Wilde and Williamson, 2016, pp. 36–37).

A host of philosophical theories of evidence are concerned with the nature of evidence and evidential relationships between true claims, or between claims we already accept (E) and claims to be established (H). The nature of evidence can be understood in terms of *evidential relationships*. To stand in an evidential relationship, these claims must be connected by a relevance-mediating vehicle that links evidence (E) and hypothesis (H). I propose that we categorise the evidential relationships established by relevance-mediating vehicles into three different types:

I. OS (*objective support*): in what sense E guarantees H, or how plausible H is given E (e.g. the materials, testimonies and images labelled as evidence are good reasons for belief in the claim labelled as the hypothesis). This relationship is independent of one’s beliefs: E is or is not a good reason for H whether one takes it to be or not.

II. SA (*subjective acceptance*): when we are justified in believing H given E, and how confident we should be of H given E (e.g. given the same array of evidence, rational agents should all be justified in believing the same conclusion, or agents holding contrasting values should follow the same measures to evaluate evidence irrespective of reaching the same conclusion).

III. GD (*guidance*): how E guides us in further investigating H (e.g. a certain brand of duct tape indicates that someone is a suspect whose movements or bank account should be looked into).

Put plainly, objective support is concerned with the world itself, subjective acceptance with beliefs and justification, and guidance with a fair guess.

## 1.2 The Desideratum View versus the Tool-Based View

I begin with some prerequisites for a good theory of evidence from the desideratum perspective before introducing a tool-based view of theories of evidence. Achinstein (2014) identifies two desiderata that a theory of evidence must fulfil, namely the *good-reason-to-believe* desideratum (A1) and the *empirical* desideratum (A2). With respect to (A1), Achinstein maintains that E is evidence for H given background knowledge (B) only if E given B provides a good reason to believe H, where E is a good reason to believe H given B only if P(H/E&B)>1/2. The empirical desideratum further requires that whether E is evidence for H, given B, be a matter of empirical fact beyond E and B's being empirical facts. The good-reason-to-believe desideratum defines an objective standard by which H can be considered more certain than not-H given E as well as a subjective standard of one's justification for accepting H as true when knowing E. The empirical desideratum prohibits any *a priori* evidential relationships. For example, the hypothetico-deductive (H-D) account does not satisfy this desideratum, since for it, the relation between evidence and hypothesis is deductive entailment (given that E and B are true, E is evidence for H given B if and only if H in combination with B entails E), which is decidable *a priori*.

What is at stake is to articulate prerequisites that cover a whole landscape of evidential relationships and demarcate it into sensible divisions, even though I will later put forward a different view that no one theory of evidence should be expected to meet all requirements. Let us consider a more all-encompassing set of desiderata urged by Reiss (2015b, p. 37). He argues that a theory of evidence should be informative about how to gather evidence and when to be justified in believing a hypothesis on the basis of evidence in non-ideal scenarios as well as ideal ones. He puts forward four desiderata for a theory of evidence. It should:

(R1) be a theory of support;
(R2) be a theory of warrant;
(R3) apply to non-ideal scenarios;
(R4) be descriptively adequate.

For support, (R1) requires that a theory of evidence explain the role of evidence as a truth indicator where it is important that this does not require that evidence be a truth guarantor. For warrant, (R2) requires that evidence provide a reason, strong or weak, to believe in H to a certain degree or that demonstrates H. In short, an account of support is informative for gathering relevant facts to form evidential claims, whilst an account of warrant relates to assessing a statement in light of evidence and determining whether to believe it.

Perfect experiments, ideally conducted randomised controlled trials (RCTs) or anything that could be called flawless evidence are scarce. Typically, whatever studies practising scientists conduct are imperfect, for example, well-controlled experiments, well-conducted RCTs or observational studies. This reality, Reiss argues, should be taken into account for a theory of evidence. Grant that a theory of evidence takes a form like this: E is evidence for H if and only if certain conditions are satisfied. In this form, what satisfies the conditions listed is legitimately termed evidence. What matters here is that whether a theory of evidence fits scientific practice hinges upon whether it is possible to tune the content of the conditions to accommodate so-called imperfect evidence that was, according to these conditions, previously not taken as evidence. This is what (R3) requires. It stipulates that a theory of evidence should allow some principles that make sense of non-ideal scenarios, say, where for causal hypotheses, ideal controlled experiments (in an attempt to exclude known interfering factors) and ideal RCTs (in an attempt to balance known and unknown interfering factors) are rarely found or background knowledge is unreliable or insufficient. On a theory that meets (R3), it is recognised that evidential statements that are not conclusive evidence can be appraised in terms of degrees of evidential support, and statements that would not have been labelled as evidence can

otherwise become evidence under certain conditions. For instance, as Reiss (2015b, p. 57) notes, an H-D theorist can adjust background knowledge (B) to the extent that the hypothesis (H) and its new background knowledge (B′) deductively entail the claim (E), and 'describe ideal as well as non-ideal situations of evidence gathering'.

(R4) requires that even if the conditions listed in a theory of evidence have been tuned as (R3) demands, their relations and contents should actually be used by scientists. Take the H-D account as an example again. The adjusted component of background knowledge is not necessarily accepted by scientists, especially when it is manoeuvred *ad hoc* or its truth is impossible or implausible to ascertain. Furthermore, the deductive entailment (H&B entails E) is not a sufficient condition for evidence in actual use. Any trivial existential claim entailed by a hypothesis in question is not legitimate to be evidence for scientists (Achinstein, 2014, p. 387). For example, the claim 'calcium exists' is not evidence that calcium helps build strong bones.

Note that, compared to the three evidential relationships, objective support (OS), subjective acceptance (SA) and guidance (GD), (R1) is concerned with GD, and (R2), the justification function of evidence, with SA; the OS relationship is not explicitly allowed for in Reiss's desiderata. Neither (R3) nor (R4) fully or partly corresponds to OS, SA and GD, but the former are compatible with the latter. This brings to light the chief concern of Reiss's desiderata: a theory of evidence should work for us rather than articulating the objective evidential relationships that might not be useful for us.

On Reiss's account, the two desiderata that Achinstein identified for adequacy in a theory of evidence do not do justice to the full set of desiderata that a theory of evidence should satisfy. Where a good reason refers to sufficiently strong evidential warrant, Achinstein's good-reason-to-believe desideratum (A1) is similar to (R2) to the extent that both demand a certain degree of reason for H or to believe H. Nevertheless, (R2) can accommodate more degrees of reason regarding warrant strength, not restricted to $P(H/E\&B)>1/2$, as required in (A1), and (R2) does not

necessarily invoke probabilities, which is otherwise required in (A1). A theory of evidence that fulfils Achinstein's two desiderata (A1) and (A2) may nonetheless fall foul of the desideratum (R1) insofar as (A1) and (A2) do not consider the indicator role of evidence. For example, a piece of duct tape may indicate that someone is suspected of committing murder, or in a medical examination, the test for abnormality of liver function can signify liver cancer even if (A1) or (A2) are not met.

Suppose that a man is discovered to have an abnormal liver function and the clinician, given this evidence, estimates that the probability of the patient's contracting liver cancer is 1 / 1,000, since the occurrence of liver cancer in men is probabilistically low, say P(H) = 1 / 10,000 and there are more plausible causes (e.g. liver cirrhosis or other hepatic diseases) that equally explain the abnormal figures. In this case, the probability of H given E does not surpass half (~A1), yet the clinician, with knowledge of the result of the liver function test, can arrange a further medical imaging examination to try to detect the real problem. The measure of liver function given by the initial test serves an indicator role even if it is not decisive to the extent that (A1), the good-reason-to-believe criterion, demands.

Consider a card case in which claims do not meet Achinstein's empirical desideratum (A2) but still provide evidential information. There is a known statement that a card drawn at random from a standard deck is a king (K), which follows deductively from a hypothesis that 'a card drawn is the king of diamonds' (KD). With information K, support for hypothesis KD is higher than without the information, though the evidential relationship between K and KD is not a matter of empirical fact (~A2). Then again, even if P(H/E&B) < 1/2 and E is *a priori* evidence for H (~A1&~A2), some information can still indicate the truth of the conclusion of interest. Consider for example the card case. P(KD/K) < 1/2 and, as stated above, the relationship between K and KD is *a priori*, but K plays a role in indicating KD.

Still, an important question remains unanswered: namely, whether a theory of evidence must satisfy all of the desiderata. When Achinstein and Reiss set

out their desiderata, they intend them to be used in this way: invoking desiderata to judge whether a theory is legitimate is an all-or-none judgement. A theory is good if and only if it meets all of the desiderata; a theory that fails to meet one of them is otherwise considered unsatisfactory. Desiderata for a theory of evidence are by no means something invented or even fabricated by us merely with the intention of listing what we want. They need to and do take into account *real* evidential relationships. However, requiring a theory of evidence to satisfy *all* of the desiderata is something more than to simply cover evidential relationships. Clearly, Reiss's four desiderata not merely cover a considerable variety of real evidential relationships of the sort that I have outlined but also impose upon any theory of evidence a normative standard that should apply as an all-or-none requirement.

There is one other consideration that should not be ignored: desiderata for a theory of evidence, for the most part, do not include all dimensions of the evidential functions. Evidence can provide an explanation for natural or social phenomena or past events, can help with assessing the credibility of a causal hypothesis, and can also be useful for decision-making and prediction. In Achinstein's and Reiss's desideratum views, these types of evidential functions are either not involved or not characterised in detail.

The *desideratum* view essentially takes the form of 'E is evidence for H if and only if…', and some expression on the right-hand side such as: 'It is highly likely that there is an explanatory connection between E and H', 'There is an argumentative relation between E and H', or 'There is an inferential relation between E and H'. Advocates for this type of definition seek a universal account that covers all cases of evidence: invoking desiderata to judge whether a theory of evidence is legitimate is an all-or-none judgement. A theory is good if and only if it meets all of the desiderata; a theory that fails to meet one of them is considered unsatisfactory.

In contrast to this kind of all-or-none judgement, I endorse a meta-thesis of pluralism on this matter, which I call a *tool-based* view. The tool-based view

does not demand that E is evidence for H, only if certain conditions are met. This view takes the 'if' side, that is, if a theory of evidence can establish evidential relevance by a particular relevance-mediating vehicle and thus can account for certain aspects of evidence use in the scientific practices it intends to explain, then it is a good theory of evidence. From the tool-based view, a theory of evidence is a purpose-specific tool: a theory of evidence has its own strengths, difficulties and limitations. When considering strengths, it is appropriate to evaluate whether a theory of evidence operates positively against its own merits in relation to the purpose it is intended to serve. By difficulties, I mean to what extent an account does not accomplish the objectives it sets out for itself. For limitations, I seek to define the scope of its application by considering where else an account does not perform well.

According to the tool-based view, criteria for evaluating theories of evidence are comprised of, but not limited to, the evidential relationships (including OS, SA and GD) and the evidential functions (including explanation for natural or social phenomena or past events, warrant for causal relationships, and prediction). A theory of evidence can legitimately cover only certain evidential relationships or functions of the domain; it should not be faulted if it does not fulfil *all* of the desiderata that Achinstein or Reiss identifies. Theories of evidence may meet the same criterion, but this has no material consequence, as one theory may do the job better or meet more criteria than would another theory. By analogy, both a hammer and a Swiss army knife can drive nails, but the former carries out this function far more effectively and efficiently than the latter.

As an advocate for the tool-based view, I would not consider a theory of evidence inadequate unless it unequivocally fails to account for the evidence use it intends to explain. While I recognise proper theories of evidence when they are appropriately applied, is there a common core among the different concepts of evidence? Is there something linking them together and enabling them to serve as evidence? Although I would claim that evidence is the grounds upon which our hypotheses or theories are rationally confirmed or

refuted, this broad definition contributes little to our understanding of different uses of evidence in diverse domains. This is similar to a broad definition of the tool: a tool is identified by its function(s). While a screwdriver and a hammer both have their own function(s), this means of expression does not distinguish between their particular functions designed for performing particular tasks. The tool-based view aims to clearly articulate different uses of evidence in various settings.

It is worth noting that the tool-based view, a meta-thesis of pluralism, does not subscribe to relativism — any arbitrary theory of evidence should be acknowledged. One might worry that now that there is no fine-grained common core connecting different concepts of evidence, something meeting the conditions of an arbitrary theory of evidence is labelled evidence by virtue of that theory; if a theory of evidence can be formulated arbitrarily, anything in accordance with a correspondingly appropriate theory can be evidence for a hypothesis, be it scientifically desirable or not. For the tool-based view, one primary aim of a theory of evidence is to account for the aspects of evidence use in science it sets out to explain. The tool-based view does not prescribe what evidence should be; instead, by making explicit the conditions on the 'if' side, it analyses the rationales behind the evidence use judged by scientists in different aspects in diverse domains in science. If an arbitrary theory of evidence is used to account for actual evidence use in science, it is likely to face numerous counter-examples. However, the tool-based view does not rule out the following possibility: a theory of evidence specifically devised to account for the use of evidence in support of a scientifically undesirable theory (e.g. the flat-earth theory) may be taken as a good theory of evidence for certain communities, though not for the scientific community. However, although the tool-based view is open to this possibility, articulating the methodology of confirming scientifically undesirable theories is not typically the aim of philosophical theories of evidence.

What is at issue is discerning which meta-thesis concerning evidence of theories we should take: the desideratum view or the tool-based view. Here I do not argue that the tool-based view is correct and that the desideratum

view is mistaken. Instead, I wish to construct a new way of evaluating theories of evidence whereby we can measure them against the shared criteria and the purposes that they have set for themselves. More importantly, a good theory of evidence in the desideratum view advocates' mind is a 'super' theory of evidence likened to an all-encompassing and all-round 'super' Swiss army knife. But from the perspective of the tool-based view, we do not need a 'super' Swiss army knife — we need a toolkit. The super theory might not be accessible. In the subsequent chapters, I will argue that the theories of evidence under discussion have their own difficulties and limitations. The tool-based view can accommodate not only the evidential relationships and the functions that the desideratum approach has taken into account, but more relationships and functions in multiple dimensions, without regard to the all-or-none requirement.

In opposition to the desideratum view, the tool-based view is favourable in terms of breadth and depth. From the breadth point of view, a meta-thesis should cover, as far as we are concerned, all kinds of evidential relationships and functions, although a theory of evidence need not do so. For instance, Reiss's four desiderata are silent about objective support (OS). From the depth point of view, suppose $T_1$ is a theory of evidence that distinguishes between different degrees of evidential warrant (that is, how much we are warranted in believing H given E) in relation to probabilities, and suppose $T_2$ uses, in a rough way, probabilities to do this task, but additionally distinguishes between different degrees of evidential support, in a rough way as well. According to the desideratum view, it should be accepted that $T_2$ is superior to $T_1$, since $T_2$ meets more desiderata than $T_1$ does. However, $T_2$ may be so coarse-grained as to perform the first task less satisfactorily than $T_1$, which can otherwise be explained well in the meta-thesis of pluralism.

The defence of the desideratum view could be that OS is pertinent to metaphysics and far from the actual use of evidence, and that the characterisation of the desiderata can be developed and given more detail to do justice to the complexity of evidential relationships. Advocates of the desideratum view might claim that desiderata, by nature, must be *fully*

satisfied. Although there are, to my knowledge, no positive reasons for the desideratum view as it stands, advocates of this view could further argue that a small leak may sink a great ship: non-fulfilment of just one desideratum may undermine a promising theory of evidence. Alternatively, they could argue that a theory achieving multiple important purposes all at once would be even better than one that meets fewer desiderata. For the purposes of this thesis, however, I set these issues aside.

For this thesis, the tool-based view is indispensable for critically examining theories of evidence. If the desideratum view were to be applied to analyse the theories of evidence under consideration, none of these theories would be deemed satisfactory or correct. For instance, according to their respective desiderata, Achinstein regards the Bayesian theory as problematic, and Reiss considers Achinstein's, Cartwright's and the Bayesian theories wrong. Accordingly, these theories should be dismissed unless they satisfy all of the desiderata, no matter how many strengths they have, how many holes they may patch and how many limitations they may overcome. I would not judge them wrong simply by revealing their problems while ignoring their abilities to capture certain facets of evidence. Until a 'super' theory of evidence is invented, it seems reasonable to let 'qualified' theories live side by side. If my analysis in the following chapters is correct in showing that the theories of evidence under consideration more or less have difficulties and limitations, we may have no good theories at all. This concern brings me to the tool-based view; a theory of evidence is analogous to a tool that performs particular functions and accounts for particular aspects of evidence use by the scientific community. In order to properly evaluate the strengths (where a theory can be applied), difficulties (where a theory is intended to be applicable but is done poorly) and limitations (where a theory cannot be applied) of theories of evidence, I adopt and endorse the tool-based view for the purposes of this thesis.

## 1.3 Outline

In this thesis, I aim to accomplish three missions. Firstly, I apply the tool-based framework to several contemporary theories of evidence, and analyse their strengths, difficulties and limitations. Secondly, I identify their relevance-mediating vehicles at work. Thirdly, I spell out evidential reasoning in historical linguistics by means of likelihoodism.

I proceed with the appraisal of each theory of evidence by checking whether it succeeds in the task it takes upon itself, rather than by invoking a full set of desiderata.

In Chapter 2, I discuss Peter Achinstein's explanationist theory of evidence as a springboard for further investigation into other theories of evidence, including the H-D account, the error-statistical account, the satisfaction view, the subjective Bayesian theory and the objective Bayesian theory. I follow Achinstein's (2014) analysis and focus on their difficulties by employing counter-examples. I argue that the explanationist theory encounters a difficulty that the merged objective Bayesian account of evidence (MOB) can avoid. In addition, I point out that Achinstein's explanationist theory employs explanatory connection as the relevance-mediating vehicle by which it captures OS (objective support) and SA (subjective acceptance), but with the omission of GD (guidance).

In Chapter 3, I examine Nancy Cartwright's argument theory of evidence, in which the mediating vehicle for evidential relevance is a sound argument. I argue that because of its demand for deductive entailment, the argument theory faces a number of difficulties that boil down to a failure to account for the transmission of epistemic warrant in specific types of cases. One of the major limitations is that looking for a valid argument and checking the truth of the premises fall outside this theory's job description. Nevertheless, it can remind us of inferential gaps and explain why what was previously accepted as evidence may no longer serve as evidence when some other premise

turns to be false or the argument becomes invalid. The argument theory emphasises metaphysical evidential relationships which amount to OS, but may be irrelevant to SA and GD.

In Chapter 4, I scrutinise Julian Reiss's inferentialist account, which harnesses loose implication and elimination as the relevance-mediating vehicles. I argue that its real difficulty is that it has to curtail the application scope that it originally attempts to reach in the biomedical and social sciences in accord with the quantitative evidence use, and that its limitations arise from its inability to make full use of information about probabilities. I also show that this account explicates why a variety of sources of evidence are deemed legitimate in the biomedical and social sciences. Although the inferentialist account delivers an account of SA and GD, it pays little attention to OS.

In Chapter 5, I argue that the likelihoodist account of evidence excels at spelling out evidential reasoning in poorly specified settings such as the historical sciences. By virtue of the comparison core built into the law of likelihood as the relevance-mediating vehicle, I formalise the reconstruction of proto-sounds in historical linguistics, illustrating that the likelihoodist account can make sense of the methodology, whilst the explanationist theory, the argument theory, the inferentialist account and the Bayesian approach all fail to do so. I also argue that the likelihoodist account is not well suited to composite hypotheses and well-specified and fully specified settings. This account is concerned with OS but is silent about SA and GD.

# 2

# Necessitating Explanatory Connection: Achinstein's Explanationist Theory of Evidence

Why have scientists paid little or no attention to philosophical theories of evidence? In *The Book of Evidence*, Peter Achinstein (2001) posits that it is because theories of evidence are of little or no scientific importance and practical value. A challenge from a dean of science, Achinstein characterises, expresses the concerns. The dean's utilitarian mind demands that a theory of evidence should resolve scientists' disputes about evidence, such as whether the data suffice to say that a hypothesis is true. His analytical mind desires a theory of evidence that brings to light the rationale behind how practising scientists reason evidentially.

Taking on the dean's challenge, Achinstein (2001, 2010a, 2014) attempts to proffer a satisfactory theory of evidence that can make sense of objective support (OS) and, less importantly, subjective acceptance (SA), both of which are evidential relationships which I have articulated in Chapter 1. The approach he offers calls for an explanatory connection between a hypothesis and claims to confirm the hypothesis. For this reason, I will call it the *explanationist theory of evidence*, which is as follows:

> E is evidence for H given B if and only if
> (a) P(Exp(H,E)/E&B) > 1/2;
> (b) E and B are true;
> (c) E does not entail H,[1] where H is a hypothesis, B is background knowledge and Exp(H,E) means that there is an explanatory

[1] Achinstein (2010a, p. 5) rejects E as evidence for H by E's entailing H, for example, the fact that she is a female student does not evidence the hypothesis that she is a student, since 'it is too good to be evidence'. In opposition to his view, I argue in Chapter 3 that if E entails H, E can be evidence for H if and only if E succeeds in transmitting an evidential warrant to H. To illustrate this point, let us consider a simple example. Suppose there are 366 people in the room (E), we can guarantee that there must be a pair having the same birthday (H) on the basis that there are 365 days in a non-leap year. Intuitively, E is evidence for H, though E entails H.

> connection between H and E. That is, H correctly explains the truth of E, or E correctly explains the truth of H, or some hypothesis correctly explains both the truth of E and the truth of H.

Different explanatory directions are exemplified as follows: (1) H explains E: smoking as a cause of lung cancer (H) explains the correlation between smoking and lung cancer (E); (2) E explains H: Tom's having no heartbeat (E) explains his death (H); (3) another hypothesis (H′) explains both E and H: 100 tosses of a coin are all heads (E) and the hypothesis (H) is that the 101st toss will land heads; the fact that the coin is extremely biased (H′) explains both E and H. Achinstein's definition of evidence is a threshold concept, shown in the requirement that the probability should be higher than half, which he names *potential evidence*. I will return to the threshold concept and potential evidence in due course.

I begin by explicating the concept of explanatory connection and the constituents of a correct explanation. I will discuss Achinstein's four concepts of evidence, with potential evidence providing a foundation for the other three, and I will analyse Achinstein's good-reason-to-believe assumption and his empirical assumption. Following that, I will use Darrell Rowbottom's (2013) criticisms to clarify Achinstein's empirical assumption. Armed with the two assumptions that a theory of evidence should have, following Achinstein's analysis, I will present the unsatisfactory aspects of five theories of evidence. Finally, from the perspective of the tool-based view I endorse, according to which a theory of evidence is equipped with some relevance-mediating vehicle(s) and is purpose-specific as to capturing particular different kinds of evidential relationships in distinct settings, I shall argue that the explanationist theory encounters difficulties relative to its purpose and scope.

## 2.1 Explanatory Connection

Several terms of Achinstein's definition of evidence should be unpacked. I begin with 'correctly explain'. Achinstein (2010a, pp. 24–25) demands that a

correct explanatory connection should comply with objectivity, context insensitivity and non-circularity.[2] An example of a theory of explanation that satisfies these conditions is Carl Hempel's D-N (deductive-nomological) model of explanation. To elucidate these conditions, I briefly introduce the D-N model.

Hempel (1965) believes that the standard form of all scientific explanations is the D-N model, which takes the form of a deductive argument:

$$\begin{array}{l} C_1, C_2, \ldots, C_n \\ L_1, L_2, \ldots, L_m \\ \hline P \end{array}$$

The D-N model is composed of two core components: *explanandum* and *explanans* (Hempel and Oppenheim, 1948, p. 136). An explanandum is a sentence 'describing the phenomenon to be explained (not that phenomenon itself)' and an explanans is 'the class of those sentences which are adduced to account for the phenomenon' (Hempel, 1965, p. 247). $C_1, C_2, \ldots, C_n$ are initial conditions describing relevant empirical facts and $L_1, L_2, \ldots, L_m$ are general laws of nature. The sentences in the two groups together are the premises called the explanans. Conclusion P, a statement reporting the phenomenon of interest, is the explanandum. Any phenomenon to be explained should be entailed by the explanans, consisting of a law, or laws, of nature in addition to some initial conditions involving verifiable empirical content. Moreover, the explanans should be true. The inverse holds as well: if a phenomenon is a deductive consequence of an explanans consisting of true claims that contain at least one law, then the explanans constitutes an admissible scientific explanation for the phenomenon. The requirements of deductive entailment and of the explanans containing a law of nature reveal

[2] Achinstein (2010a, p. xi) distinguishes between 'correct explanation' and 'good explanation', asserting that correctness and goodness are both crucial in assessing explanations. From his viewpoint, correctness is determined independently of epistemic situations, whereas goodness varies depending on different epistemic situations. However, since the basic concept of evidence used among scientists is, as he claims, required to be objective, the concept of explanation that underwrites the objective concept of evidence, if there is one, should also be objective.

the origin of the name of the deductive-nomological model, also called the ‘covering law’ model. A case of a D-N explanation can be seen in Kepler’s laws, which explain the motion of Mars, insofar as Kepler’s laws entail the statements about Mars’ motion.

Now let us turn to the three features that Achinstein (2001, p. 160) requires of a correct explanation used to define (potential) evidence:

1. *Objectivity*: Whether an explanation is correct is free from one’s subjective beliefs. The explanatory connection in a D-N explanation is an illustration. It is objective insofar as no one’s knowledge or beliefs affect whether the deductive entailment in the D-N explanation holds or the explanans and the explanandum are true. Whether the motion of Mars is a deductive consequence of Kepler’s laws and whether Kepler’s laws and the statements describing how Mars moves in the solar system are true do not depend on what we think or believe.

2. *Context insensitivity*: A correct explanation is context-free rather than being appropriate for one context but inappropriate for another. The notion of correct explanation is analogous in certain respects to a D-N explanation, like both are not determined by the different standards of various contexts and crosses over all of them. Once Kepler’s laws plus certain initial conditions correctly explain the motion of Mars in the proper D-N manner, it is a correct explanation thereafter, whatever the more profound understanding of the universe Newton’s laws may provide.

3. *Non-circularity*: The required notion of correct explanation should not draw on the notion of evidence, which avoids circularity when defining evidence. This is analogous to the way that Hempel defines ‘explanation’ by the D-N model, which requires only deductive entailment as well as the explanans that is true and consists of a law, without appeal to the notion of explanation itself. Recall that Achinstein uses ‘correct explanation’ to define the notion of evidence,

> so if, upon spelling out what counts as a correct explanation, he stows away the notion of evidence beforehand, it would be circular. It might be thought that the notion of evidence enters the D-N account as the premises in a D-N explanation are supposed to be true, and that if Achinstein adopts the D-N model to define the notion of correction explanation, how else is truth to be ascertained other than by evidence? This is not the case, however, since Hempel requires only that the premises of a D-N explanatory argument be true, not that we (empirically) know them to be.[3]

As Achinstein (2010a) pointed out, despite explanations in Hempel's D-N model being objective, non-contextual and non-circular so that it appears to serve the purpose of providing a suitable definition of correct explanation, this model of explanation has its own problems that make it too unsatisfactory to be integrated into the definition of evidence that the explanationist theory requires.[4] For the D-N model, all considerations are *a priori* except the truth of explanans: considerations of they are whether the explanans deduces the explanandum and whether sentences contained in the explanans are lawlike. Neither is determined empirically: the former is determined by logic, and in Hempel's view, the latter 'depends only on [a sentence's] syntactical form and the semantical interpretation of its terms (Achinstein, 1983, p. 165). For instance, take the premise (a) that Mr. X drank a moderate amount of water every day of his life is true and the lawlike premise (b) that anyone who drinks water every day dies before the age of 150 is also true. It follows from both premises that (c) Mr. X died before the

[3] Hempel (1965, p. 248) stipulates that the premises must be true rather than highly confirmed, the latter of which of course means that they are supported by evidence. The reason for this stipulation is to avoid awkward consequences, where 'a certain phenomenon was explained at an earlier stage of science, by means of an explanans which was well supported by the evidence then at hand, but which has been highly disconfirmed by more recent empirical findings, [so] we would have to say that originally the explanatory account was a correct explanation, but that it ceased to be one later, when unfavourable evidence was discovered'.

[4] I am not concerned here with all of the well-known problems with the D-N account. They include the eclipse problem, Bromberger's flagpole example, the barometer problem, the moon and the tides problem, the syphilis and paresis problem, the hexed salt problem and the birth-control pills problem. See Salmon (1989, pp. 46–50) for more details. Below the drinking-water-every-day example illustrates the problem of explanatory irrelevancy, which is pertinent to the last two problems.

age of 150. In the D-N model, the premises should correctly explain (c) in the sense that supposing (a) and (b) are true, it follows *a priori* that (c) is true, regardless of whether other empirical information enters. However, drinking water is not normally regarded as a reason for fatality, and thus does not correctly explain that Mr. X died before he was 150 years old. This shows that legitimate D-N style explanations do not guarantee that they are correct explanations.

Wishing to avoid both the D-N account and viewing the concept of explanation as a fundamental one without further clarification, Achinstein instead proposes a content-link account of explanation, which he maintains suits the requirements of the concept of potential evidence. He distinguishes between potential explanation and correct explanation. A proposition p, if true, is a potential explanation if and only if p is a correct explanation (Achinstein, 2001, p. 148, 2010a, p. 31). In this sense, the correct explanation is a derivative of potential explanation, the latter of which should be fleshed out first.

For Achinstein, an explanation is a propositional relationship between the explanans (*content-giving proposition* p) and the question raised by the explanandum (*content-question* Q).[5] Suppose there is a question of interest: why did Peter's car have a breakdown? The form of an answer to a why-question can always be expressed as: the reason is _____ (the blank should be filled in with the content of reason). Accordingly, the reason that Peter's car broke down is that, say, the battery was faulty, where 'reason' is dubbed a *content-noun* N and 'the battery was faulty' is a sentence giving the content to the content-noun N. Moreover, it can be seen that this why-question presupposes a variety of propositions including 'Peter had a car', 'Peter's car broke down' and 'Peter's car broke down for some reason'. A *complete presupposition* of a question is then defined as 'a proposition that entails all and only the presuppositions of that question' (Achinstein, 2010a, p. 27). We may stipulate that the three presuppositions noted collectively exhaust all of

[5] For more details, refer to Chapter 2 in Achinstein (1983).

the presuppositions of the question ‘Why did Peter’s car have a breakdown?’ and it seems that the third entails the first two. If so, then the proposition ‘Peter’s car broke down for some reason’ is a complete presupposition of this question. By omitting ‘for some reason’ and adding ‘the reason that’ and ‘is’ followed by a blank, we obtain a *complete answer form* for the question ‘Why did Peter’s car break down?’: The reason that Peter’s car broke down is _____.

With the notions outlined above, Achinstein (2010a) gives a definition of a complete content-giving proposition:

> *p* is a *complete* content-giving proposition with respect to question Q if and only if
> (a) *p* is a content-giving proposition (for a concept expressed by some noun N);
> (b) *p* is expressible by a sentence obtained from a complete answer form for Q (containing N) by filling in the blank;
> (c) *p* is not a presupposition of Q. (Achinstein, 2010a, p. 27, italics in original)

Q is a *content-question* if and only if there is a complete content-giving proposition p to answer Q. Therefore, ‘The reason why Peter’s car broke down is that the battery was faulty’ is a complete content-giving proposition relating to the content-question ‘Why did Peter’s car break down?’. Now we have the essential ingredients for Achinstein’s account of explanation: a complete content-giving proposition p provides a potential explanation for Q (or its indirect form q ‘the car broke down’) and if p is furthermore true, then p provides a *correct* explanation of Q (or q) (Achinstein, 2001, p. 163). If ‘the reason why Peter’s car broke down is that the battery was faulty’ is true,[6] it provides a correct explanation of why Peter’s car broke down. By contrast, the proposition that ‘the battery was faulty’ alone (without ‘the reason why Peter’s car broke down is that‘) does not necessarily provide a correct explanation for the question ‘Why did Peter’s car break down?’, because it is not a complete content-giving proposition relative to the question. ‘The

[6] Achinstein (2001, p. 166) permits a substitution of ‘cause’ for ‘reason’, as in the case where the cause of Peter’s car’s breakdown is the faulty battery.

reason why Peter's car broke down is that the battery was faulty' guarantees that the breakdown was a result of the faulty battery, whilst even if the proposition that 'the battery was faulty' is true, the breakdown may stem from other reasons. Thus, Achinstein's account of explanation can avoid propositions in a legitimate form of explanation, which permit irrelevant reasons to be correct explanations.

As noted above, whilst the D-N model is prone to permission for irrelevant reasons presented as correct explanations, Achinstein's content-link account of explanation seems to forestall this criticism. However, the D-N model meets the requirements for correct explanations by virtue of being objective, contextually insensitive, and not circular. Let us examine whether Achinstein's model meets all three requirements. As regards objectivity and contextual insensitivity: the fact that the reason Peter's car broke down is that the battery was faulty provides an explanation of why Peter's car broke down, which is neither dependent upon Peter's or anyone else's beliefs and knowledge, nor is it subject to a more profound understanding of the mechanisms. Achinstein's definition of explanation is not circular and can be used to define the notion of evidence, because content-giving propositions and content-questions are not defined by any notion of explanation nor by any notion of evidence. Admittedly, the concepts of reason and explanation are intimate; however, Achinstein's definition of explanation is still not circular in this respect. Although Achinstein does not explicate the truth conditions for 'the reason is ___' when defending his account of explanation, his definition of correct explanation does not rely on 'reason' and provides no conditions for determining the truth of claims containing 'reason' such as 'the reason why Peter's car broke down is that the battery was faulty'. Accordingly, Achinstein's definition of explanation is objective, non-contextual and non-circular.

Note that Achinstein's content-link account of explanation does not block normal use of the explanans. The fact that 'the battery was faulty' (i.e. a content-giving proposition) is still acknowledged in his account to provide a correct explanation of Q 'Why did Peter's car break down?', only if the former

can be transformed into a *complete* content-giving proposition with respect to Q. In case the proposition 'the battery was faulty' cannot be furnished with a complete content-giving form, it is possible that when the proposition 'the battery was faulty' and the question Q are laid out, the proposition to answer Q is true but it does not correctly explain Q. Seeing as there may be one reason or another that provides a correct explanation for Q, the link between the proposition and Q is not strong enough to tie both into an explanatory relationship. By contrast, it is impossible that the proposition 'the reason why Peter's car broke down is that the battery was faulty' is true but it does not provide a correct explanation for Q. Within Achinstein's framework, it is not necessarily the case that given only Q and the explanans with respect to Q, the explanans does not correctly explain Q.

## 2.2 Four Concepts of Evidence

Achinstein (2010a, 2014) argues that his theory of evidence can accommodate the four evidence concepts: *epistemic-situation*, *subjective*, *potential* and *veridical* evidence. He claims that these concepts are used in scientific practice and daily lives and must be taken into account when we develop an account of evidence.

Prior to the exploration of the four concepts, an important conceptual distinction should be made between *justification* and *good reason to believe*, the latter of which is concerned with the core of Achinstein's theory of evidence, namely potential evidence. One has justification for believing H (viz. an inference that H is true is rational) with respect to a specific epistemic situation in which any rational agent would believe that H is true. By contrast, something constitutes a good reason for H if it is sufficiently strong for one to believe that H is true from the perspective of an omniscient God. To see the distinction between justification and good reason, consider the following situations.

(a) Justification does not imply good reason:
Suppose an investigator A, from a piece of duct tape that contained

K's DNA and was left in the crime scene, constructed the profile of a killer and is justified in believing that K is the killer. Any experienced investigators in this epistemic situation would come to the same conclusion. However, improbable as it might seem, K's DNA, unbeknown to A, had been transferred accidentally onto the duct tape. This indicates that the contaminated piece of duct tape is not a good reason for believing that K is the killer who investigator A mistakenly believes K to be. Something justifying a belief does not necessarily amount to something being a good reason.

In a different case, a bloodstain is a good reason to suppose that victim V did not commit suicide but was murdered. However, an experienced judge J, with the testimonies heard and the samples gathered by the investigator A, who neglected the bloodstain, under the carpet, containing someone's DNA, formed a judicious judgement and concluded that V died by suicide, and acquitted the defendant K. Although the bloodstain escaped A's and J's notice, the belief that V died by suicide is regarded as justified, because a well-informed epistemic agent, according to the available information, would believe the same hypothesis as the judge. A relevant good reason may be overlooked even if a belief is justified to be true, but is in fact false.

(b) Good reason does not imply justification:

Pursuing the above example, suppose that a bloodstain collected at a crime scene containing someone's DNA provided a strong indication of and thus a good reason for the identity of the perpetrator, and that the death of V occurred before the 1980s, and therefore before the invention of DNA profiling. Since A and J lacked a reliable technique for identifying DNA that allowed for the matching of the bloodstain and K, they could not be justified in believing that K was the perpetrator, merely on account of the bloodstain. Just as a good reason need not be regarded as a good reason for someone to believe.

> Moreover, knowledge of good reason does not imply justification. Suppose footage caught on a surveillance camera shows a homicide and the murderer's face, which is regarded as a good reason to believe that M is the murderer. Initially, the footage justified the police's belief that M was the murderer, but afterwards they found out that M had a twin brother M', who had in fact committed the crime. Though the police no longer have justification for this belief, they still know and believe that the footage provides a good reason that M is the murderer.

Now let us turn to the four concepts of evidence that Achinstein distinguishes. The first concept of evidence is designated as *epistemic-situation* evidence. This is relativised to a situation limited to certain epistemic conditions due to, for example, measurement techniques or background knowledge. If anyone is in the same epistemic situation or given the same epistemic premises (not necessarily good reasons), they should be justified in believing the same hypothesis. From this point of view, epistemic-situation evidence is *impersonally* objective. Epistemic evidence is something like the duct tape that makes investigator A's belief justified or the testimonies and the samples that enable Judge J to reach the judgement that any rational agents would agree to if they had an identical body of evidence.

The second kind of evidence is *subjective* evidence. Whether E is evidence is determined by one's beliefs and community judgement: E is someone's or some community's evidence. Subjective evidence E for a hypothesis H for a person or community G is what is judged good evidence for H by G's beliefs and judgement. The fact that E is judged to be evidence in this community does not guarantee that a person who has E has a good reason or is justified in believing H. This can be demonstrated with reference to a previous court example. The spectators in the court may believe that the accused is guilty mainly on the basis that the accused has a previous conviction for assault. The previous criminal record is subjective evidence. The spectators' judgement relies on their personal beliefs, or even the stereotypes, and lacks a real sense of good reason and justification. Good reason and justification

are based on rationality, which should not be influenced by the standards that vary with different psychological states or social contexts.

By contrast, the last two concepts of evidence are concerned with a *completely* objective notion of evidence, namely, good reason to believe. As correct explanations, potential evidence and veridical evidence are objective and contextually insensitive: both are independent of whether or what we know and not relativised to any epistemic situation (Achinstein, 2010a, pp. 4–5).

> *Potential evidence* requires that
> (a) E does not depend upon anyone's beliefs about H, E or anything about the relationship between H and E;
> (b) E is true;
> (c) E is evidence for H even if H is false;
> (d) E does not entail H.
>
> *Veridical evidence* requires not only what potential evidence requires but also the truth of the targeted hypothesis.

In the scenario discussed above, when investigator A is claimed not to have a good reason that the suspect is the killer, this is due to the erroneous information used to infer the killer K. The erroneous information, by Achinstein (1995, p. 458), is generated from a flaw in the procedure, called *evidential flaws*. Despite the absence of a clear definition of what constitutes evidential flaws, the definition can be characterised as follows:

> The procedure yielding E is evidentially flawed if and only if E is not sufficiently informative about the truth of H (i.e. E does not provide a good reason for H).

More specifically, this insufficiency is mainly due to the mismatch between so-called evidence and its target hypothesis. For example, suppose specific representative values individually contained in two datasets for two variables

are used to compute a Pearson correlation coefficient, obtaining 0.9. This information sufficiently supports a hypothesis that there is a strong correlation between the two variables; hence the procedure yielding the information is not flawed. But if the information about these values is taken as evidence to infer that the two variables are causally related, the procedure yielding the information becomes flawed: the evidential claim is false that the correlation coefficient is (potential) evidence that both variables are causally related. As other reasons (e.g. a common variable that causes the two variables) can explain the correlation, the correlation coefficient is not sufficiently informative about the truth of the causal hypothesis. Accordingly, whether or not E is evidence is relativised to H (plus background knowledge): once H varies, E may or may not be evidence for H.

Note that for the explanationist theory, *bona fide* evidence, whether potential or veridical, can by no means be yielded by a procedure that is flawed. Evidence statements, if true, objectively support a hypothesis, regardless of our beliefs and knowledge. While the truth of (presented) evidence is required in the explanationist theory, it is not the case that the truth of presented evidence entails, or even strongly suggests, that the presented evidence is genuine. Let us consider the duct tape example again. K's biological trace left on the duct tape is real (or the claim about this information is true), but it does not necessarily mean that it is evidence of K being the killer. Since the duct tape does not actually provide sufficient evidence that K is the killer, the procedure yielding this evidence is flawed. Although whether E is evidence for H is independent of us, perhaps from the perspective of Achinstein, all scientists can do is approximate genuine evidence as closely as possible.

If the procedure yielding the information at hand were not flawed, then this item of information would be potential evidence. In addition to the procedure that is not flawed, if the hypothesis under investigation is true, the information used to infer the hypothesis is veridical evidence. Since the footage recording information E about M being the murderer is not flawed, E is

potential evidence.[7] E is not veridical evidence unless M, rather than his twin brother M', is the murderer.

Achinstein's definition of evidence, which I outlined at the outset, refers to the concept of potential evidence. Although scientists aim to seek veridical evidence, it is rarely obtained in the typical absence of knowing that H is true and that H and E are correctly explanatorily connected. Instead, potential evidence is a relatively achievable goal. The formal definitions of the four concepts of evidence can be given with his definition of potential evidence:

> *E is potential evidence for H given B* if and only if
> (a) P(Exp(H,E)/E&B) > 1/2;
> (b) E and B are true;
> (c) E does not entail H.

The notion of veridical evidence demands more:

> *E is veridical evidence for H given B* if and only if E is potential evidence for H given B, in addition to the conditions:
> (d) H is true;
> (e) there is an explanatory connection between H and E.

The notion of epistemic-situation evidence can be defined as:

> *E is epistemic-situation evidence for H given B* if and only if anyone in the same epistemic situation would believe that E is true and have justification for believing that E is veridical evidence for H given B.

[7] A formal way that fits Achinstein's definition of evidence shows that the footage (E) is potential evidence that M is the murderer (H). Let B represent background knowledge about the reliability of the surveillance camera and about the birth rate of identical twins etc. and therefore let us assume P(H/E&B) = 0.95. As will be proved later, P(Exp(H,E)/E&B) = P(Exp(H,E)/H&E&B) x P(H/E&B). Since in the murder case, if H, E and B are true, H correctly explains E, that is, that M is the murderer (in fact M is not) correctly explains the murderer's face in the footage. We obtain P(Exp(H,E)/H&E&B) ≈ 1, and thus P(Exp(H,E)/E&B) ≈ 1 x 0.95 = 0.95, which is much greater than 1/2.

Finally, the definition of subjective evidence is as follows:

> *E is subjective evidence for H given B* if and only if a person or community believes that E is veridical evidence for H given B.

Achinstein's notions of evidence cover two principal evidential relationships: objective support (OS), which is independent of one's belief and concerned with the plausibility of H given E, and subjective acceptance (SA), which pertains to one's justification in believing in H given E. But his notions of evidence are silent about how E guides us in further investigation H, which is the third principal evidential relationship, guidance (GD). In what follows, I will use the notion of potential evidence, which is the core of Achinstein's theory of evidence, to discuss and evaluate his theory.[8]

## 2.3 Good-Reason-to-Believe and Empirical Assumptions

Achinstein (2014) examines the five standard theories of evidence, namely, the hypothetico-deductive (H-D) account, the error-statistical account, the satisfaction view, the subjective Bayesian theory and the objective Bayesian theory. He concludes that they are not satisfactory candidate theories of evidence, since they do not meet one or two assumptions that a theory of evidence should meet, i.e. the *good-reason-to-believe assumption* and the *empirical assumption*.

> **Good-reason-to-believe assumption:**
> E is evidence for H given B only if P(H/E&B) > 1/2.

> **Empirical assumption:**
> Whether E is evidence for H given B is a matter of empirical fact beyond E and B's being empirical facts.

[8] Rehg (2009, p. 85) holds a similar view, claiming that '[t]he category of potential evidence constitutes the realist core of Achinstein's normative model of evidence'.

The two assumptions show that what makes the giving of good reasons and the justification of a hypothesis work is that the evidence raises the probability of the truth of the hypothesis to a sufficient degree and that the evidential relationship is built on empirical investigation. With respect to the good-reason-to-believe assumption, potential evidence and veridical evidence, as noted above, are objectively good reasons to believe hypotheses to the extent that the 'degree of reasonableness is an objective, non-physical, normative fact determined by … physical and mathematical facts' (Achinstein, 2001, p. 97). Moreover, Achinstein's definition of evidence, P(Exp(H,E)/E&B) > 1/2, entails P(H/E&B) > 1/2, which will be proved in Section 2.5.

How much is good enough?[9] Achinstein maintains that E is evidence for H given B only if, given B, E provides a good reason to believe H, and that E is a good reason to believe H given B only if P(H/E&B) > 1/2, which is sufficiently high for the threshold. The function of the threshold is in ensuring that it is not the case that E is evidence for H and ~H simultaneously. If the threshold is set at not greater than 1/2, an unfavourable result might occur. To illustrate this point, consider Achinstein's (2001, p. 7) own coin example. Flipping a fair coin does not lean towards the probability of landing heads nor the probability of landing tails as the fair coin equally supports both, which equals not providing a good reason for either. However, if the threshold is set greater than 1/2, E begins to play an evidential role in supporting a hypothesis rather than its negation.

It is worth remarking here that Reiss (2015) argues that a piece of putative evidence can play a role in indicating the truth of hypothesis $H_1$ and the truth of its rival hypothesis $H_2$. For example, a correlation between two variables (E) can be evidence not only for a causal relationship between the two

[9] Achinstein opposes slightly increased probability as evidence for a hypothesis. In the words of Achinstein (2010a, p. 38): '[A]lthough the fact that I am entering an elevator increases my chances of being in an elevator accident, it is not evidence that this will be so, even a little bit of evidence, since by itself it fails to provide any reason to believe this hypothesis'.

variables ($H_1$) but also for selection bias ($H_2$).[10] In Reiss's sense of evidence, even if $P(H_1/E) < 1/2$ and $P(H_2/E) < 1/2$, which do not meet the good-reason-to-believe assumption, E is (direct) evidence both for $H_1$ and $H_2$ insofar as E is expected supposing that either $H_1$ or $H_2$ is true. Reiss's definitions of evidence will be further discussed and examined in Chapter 4. The correlation, Reiss may claim, is evidence for either rival hypothesis simultaneously. However, in Achinstein's framework, if E is evidence for $H_1$, then it cannot be evidence for $H_2$, which is incompatible with $H_1$, since Achinstein (2001, p. 7) holds that 'if e is a good reason to believe h, then it cannot also be a good reason to believe not-h or some proposition incompatible with h'.[11]

Conciliation can more or less be made available, perhaps, in terms of further disambiguation of 'evidence for what hypothesis'. Suppose the hypothesis of interest ($H_3$) now is that there are some non-coincidence reasons that two variables are correlated (including, for example, causal relationship and selection bias). Let us assume that there is a mechanism linking the two correlated variables and that given this mechanism and correlation E, the probability that there is an explanatory connection between E and $H_3$ is so high as to exceed Achinstein's threshold of 1/2. In this case, Achinstein and Reiss should both agree that correlation E is evidence for $H_3$, when relativised to a hypothesis in which the correlation arises from sheer coincidence; for example, the correlations between Venetian sea levels and British bread prices does not imply causation (Sober, 2001). Despite indicating that there are some non-coincidence reasons, in which the correlation indicates $H_1$ and $H_2$ equally, the correlation may not evidentially support $H_1$ and $H_2$ equally within the candidates for non-coincidence reasons. Thus, for Reiss, correlation is evidence of a causal relationship and selection bias simultaneously in a relatively loose sense, whereas, Achinstein and Reiss should agree, once the evidential settings are carefully specified,

[10] Selection bias is an error which occurs when subjects are not allocated at random to treatment groups and control groups. The outcomes of the study result in part from the systematic differences in characteristics between both groups. Hence the causal conclusion cannot be drawn from the outcomes (Last, 2001, p.166).
[11] Lower case nomenclature is used in the original.

correlation is evidence for one but not for the other, or is not evidence for either.

This empirical assumption is proposed, mainly because Achinstein counters the view that whether E is evidence for H is *a priori*. Achinstein (1995) articulates the *a priori thesis* as follows:

> The only empirical fact that can affect the truth of evidential claims of the form 'e is evidence for h' (or 'e confirms h more than h', or 'e confirms h to degree r') is the truth of e. All other considerations are *a priori*. (Achinstein, 1995, pp. 448–449)

The *a priori* thesis has been assumed in a host of theories of evidence, for example, the H-D account of evidence, the satisfaction view of evidence and Carnap's objective Bayesian theory of evidence, all of which I shall discuss later. In short, presented with H and the truth of E, it is fully decidable *a priori*, via logical and/or mathematical calculations, whether E is evidence for H.

Note that Achinstein (2010a, p. 42) does not declare that 'all evidential statements are empirical'. On the contrary, he acknowledges that there are certain well-specified situations where the claim 'E is evidence for H' is decidable *a priori*. The birthday problem noted earlier offers a case in point. Suppose there are fifty people in a room (E). Let H represent at least two people who share the same birthday. Since P(H/E) is approximately 0.97, E is evidence for H *a priori*. Achinstein (1995) states that

> [i]f we insist that at least some (very basic) evidential claims are not subject to [unexpected flaws which falsify evidence statements], then we can say, more guardedly, that any evidential claim—and these will be numerous—whose truth-value can be affected by the existence of evidential flaws is an empirical claim.[12] (Achinstein, 1995, p. 461)

[12] See also: 'Although there are some a priori objective evidence claims, for the most part, objective evidence claims are empirical' (Achinstein, 2005, p. 48).

Although he does not make it explicit what 'very basic' means, fully specified situations such as the birthday case should, at least, be part of evidential claims invulnerable to flaws.

Achinstein is somewhat inconsistent here insofar as his empirical assumption states that whether E is evidence for H is a matter of empirical fact. I therefore modify the assumption as follows:

> **Empirical assumption (*):**
> Whether E is evidence for H given B is *mostly* a matter of empirical fact beyond E and B's being empirical facts.

Strategically, one can remain neutral about (*) while refuting the *a priori* thesis with a single evidential claim withdrawn on the basis of further empirical facts that shows that the procedure yielding the evidence is flawed. A drug example will be provided later, as a case in point.

In Achinstein's view, evidential relations are objective to the extent to which they are independent of us, with no bearing on whether we can know them *a priori*. The pair of concepts, *a priori* versus *a posteriori*, does not necessarily co-occur with the pair, objective versus subjective. When a claim is *a priori*, it means that such a claim can be known prior to experience, hence an epistemic topic. Objectivity, by contrast, is an ontological topic, which is concerned with dependence on or independence from us. If a claim is objectively true, it may or may not be true *a priori*. The logical truth that P entails P or the mathematical truth that 1+1 = 2 is objective and known *a priori*. By contrast, a claim that DNA samples can identify biological relationships between parent and child is *a posteriori* true and independent of our beliefs. The reverse may hold as well: if a claim is *a priori* true, it may or may not be objectively true. A form of a *subjective a priori thesis* is as follows: 'The only empirical facts that can affect the truth of evidential claims of the form 'e is evidence for h for subject S' (or 'e confirms h more than h′ for S', or 'e confirms h to degree r for S') are: (a) the truth of e; and (b) the fixed values of S's conditional degrees of belief (such as P(h, e))

(Rowbottom, 2013, p. 2824, lower case nomenclature in the original). This *a priori* thesis is subjective, in that the evidential claim that E is evidence for H is dependent on individuals.

Achinstein and many *a priori* theorists agree that the evidential claim that E is evidence for H is objective.[13] That is, whether E is evidence for H (or E confirms H) does not depend upon anyone's beliefs or knowledge (Achinstein, 1995, p. 448). For example, the fact that a certain DNA sample examined by DNA paternity testing shows that two male individuals having a biological relationship between parent and child is evidence that they are father and son, without regard to what anyone believes or knows. No matter whether the claim that the DNA is evidence for the biological relationship is *a priori* or not, depending upon different theories of evidence, this claim is objective rather than subjective. In addition, Achinstein and many *a priori* theorists agree, or can in principle agree, that whether E is evidence for H rests partly on the empirical fact of E's truth. If no such fact exists concerning the DNA sample, it is false, or at least questionable, that the claim about the sample is evidence that one is the father of the other. This illustrates that the truth of evidential claims should be presupposed even on *a priori* accounts of evidence.[14]

---

[13] Not all *a priori* theorists agree that evidential claims are objective. Some subjective Bayesians, for example, Bruno de Finetti (1937 pp. 106–107), take probabilities as degrees of belief, whose values are fixed. Suppose $P(H/E_1)$ and $P(H/E_1\&E_2)$ are accessible purely *a priori* within one's mind. After receiving an additional piece of information from experience, someone updates the degree of belief about H from $P(H/E_1)$ to $P(H/E_1\&E_2)$. In this update, $P(H/E_1)$ is not refuted and instead, as the evidence mounts up, is simply replaced by $P(H/E_1\&E_2)$. The values of $P(H/E_1)$ and $P(H/E_1\&E_2)$ themselves remain unchanged. This entire updating process conforms to Bayes' theorem.

[14] Achinstein (1995, p. 448) asserts that 'normally a claim of the form "e is evidence for h" (or "e confirms h to degree r") entails or presupposes that e is true. … But if the truth of "e is evidence for h" requires the truth of e, then since e is empirical (theorists in question are concerned only with empirical evidence), the a priori thesis is obviously false from the outset'. To avoid this obviously trivial falsehood, he articulates the *a priori* thesis, with the requirement of the truth of E, noted above in the main text. Having said this, *a priori* theorists may or may not subscribe to the view that E is evidence for H only if E is true. For example, a proponent of the H-D account might simply say that if H entails E, E is evidence for H, irrespective of the truth of E. That two people are genetically related given background knowledge entails that their DNA samples have a particular pattern in common. For the H-D account, before DNA was discovered (or constructed), DNA samples were still evidence for genetical relationships, no matter whether DNA had been discovered or not. By analogy, consider statements with a universal quantification, such as 'all men are mortal, which can be paraphrased as 'if S is a man, S is mortal'. There can be no existential import in this universally quantified statement, that is, this statement does not imply the existence of men.

In spite of the agreements both parties may reach, Achinstein (1995) contends that the *a priori* thesis is erroneous: it cannot account for cases where a claim that was previously taken as evidence turns out to be found yielded by a flawed procedure and therefore can no longer serve as evidence. According to the *a priori* thesis, once claim E is, or is labelled as, evidence for H, however many empirical facts stream in, it will not change the role or the label of being evidence. If E is true and is evidence for H *a priori*, there need not be further empirical inquiry into whether or not E is so. This resembles logical or mathematical truths: they are knowable without empirical investigation. Sometimes one may be mistaken about them, but they will not turn out to be false, even if one believes that they are. The original and the (*) empirical assumptions, despite also deeming an evidential relation ('E is evidence for H') objective, both hold that E could mistakenly be labelled as evidence, which may be revealed by further empirical facts.

To illustrate, consider Achinstein's (1995) own example of a study that was designed to find whether a new drug D relieves symptoms S and which obtained a positive result:

> $E_1$: 1000 patients presenting with S were given D, with 950 of them having their S relieved.

There are a hypothesis of interest and background knowledge:

> H: Hilary's symptoms S will be relieved after she takes D.
> B: Hilary, with symptoms S, takes D.

---

This view bears on the concept of empty domain and free logic. For further discussion, see Leonard (2002, Chapter 2). However, as far as the concept of evidence in common use is concerned, it seems not problematic when one says that a DNA sample is, or would be, evidence for their genetic relationship but there is no such DNA sample. Yet it is bizarre when one says that the DNA sample is evidence for their genetic relationship but there is no such DNA sample. In the latter situation, when saying E is evidence for H, we presuppose the existence of E.

At first glance, $E_1$ seems to be strong evidence for H. The overwhelming majority of patients given D in the study were relieved of S, so it is expected that Hilary's symptoms S are very likely to be relieved by drug D. At this stage, if *a priori* theorists adopt, for example, Rudolf Carnap's confirmation function c* based on the empirical and logical factors, they would conclude that $E_1$ confirms H to the degree 0.92, and that $E_1$ is evidence for H insofar as $P(H/E_1)$ is sufficiently high or $P(H/E_1) > P(H)$ (=0.375) (Achinstein, 1995, pp. 472–473).

Nevertheless, upon allowing for the new information, we will find the ostensibly strong evidential relationship between $E_1$ and H to be debunked.

> $E_2$: A companion study of equal size of patients with S who were not given D reported that 990 of them experienced relief of S, even more quickly than in the first study.

Contrasting the new item of information $E_2$ with the preceding information $E_1$, it is indicated that the drug D is not causally efficacious in relieving symptoms S. If D were clinically effective, the percentage of relief in the treatment arm should be greater than that in the control arm. In effect, 95% of the patients taking D feel relieved, but 99% of the patients not taking D had the same effect. There is another explanation for H (e.g. if H were true, H because of the immune system resilient to D) that is more likely than $E_1$'s explaining H, if H were true. In this regard, the study yielding $E_1$ is flawed by virtue of not being sufficiently informative about H. Information $E_1$ is, in the light of $E_2$, shown not to be evidence or good evidence for H.

The first study does not include and is not compared to the control group, which can give a baseline clinical assessment to test whether drug D has causal efficacy for symptoms S. Evidently, $E_1$ fails to serve as evidence for H when $E_2$ is factored in. This additional empirical information $E_2$, Achinstein argues, leads us to the conclusion that the *a priori* thesis should be repudiated: this thesis maintains that $E_1$, once being true and identified as evidence for H, remains evidence for H perpetually on the basis that $E_1$ and

H are in an evidential relationship *a priori*, unchanged by any empirical facts, which clearly does not make sense in the drug example. For this reason, Achinstein (1995) insists that the empirical assumption that whether or not E is evidence for H, given B, is a matter of empirical fact beyond E and B's being empirical facts should be presupposed, or at least applied for most cases, in any theory of evidence.

In defence of the possibility of the *a priori* thesis, Rowbottom (2013) argues that cases in support of the empirical assumption can also be explained from the *a priori* perspective by changing background knowledge B.[15] In doing so, he considers a way of laying bare necessary background information, instead of burying this in elliptical evidential claims.[16] The background knowledge can be expressed as: the study referred to in $E_1$ was not flawed, which Rowbottom (2013, p. 2827) coins the *working assumption*. The basic idea is that E and B are manoeuvrable and jointly constitute evidence — evidence is not necessarily restricted to E itself.

To see how this idea works, consider the possible expressions of the drug example along the lines of Achinstein's and Rowbottom's thoughts respectively. For Achinstein, the procedure yielding $E_1$ is considered flawed, because $E_2$ comes in exposing the insufficiency of $E_1$ (and B jointly) as evidence for H; $E_1$ was evidence for H but becomes non-evidence for H. The case Achinstein wishes to make is that whether or not $E_1$ is evidence is determined by empirical information other than $E_1$. One may defend the *a priori* thesis by asserting that given that the truths of $E_1$ and $E_2$ are empirically decidable, it is *a priori* true that $E_1$ is evidence for H and $E_1$&$E_2$ is not

[15] Rowbottom's (2013, p. 2828, italics in original) position is as follows: '[T]here is no evidence from the way that science is done (in the sense of scientific method) which tells in favour of *either* view of evidence, a priori or empirical. Conversely, I also hold that neither view of evidence has any direct consequences for scientific method'.

[16] Achinstein (2001, p. 10) anticipated this kind of defence of the *a priori* thesis and replied: 'There are cases … where enough information is packed into the *e*-statement to make the claim that *e* is evidence that *h* a priori. But these cases are the exception, not the rule. Nor am I denying that it is possible to transform an empirically evidential claim into an a priori one by incorporating a sufficient amount of additional information of a sort that might be used in defending the empirically evidential claim. But even if this is possible, that will not suffice to alter the empirical character of the original evidential claim, or demonstrate that the original claim is incomplete until this transformation occurs'.

evidence for H, or the degree of support that $E_1$ confers upon H is fixed at $r_1$ and the degree of support that $E_1$ and $E_2$ confer upon H is fixed at $r_2$. However, this line of defence is not satisfactory. It cannot explain that once $E_1$ (for example, to the degree 0.92 according to Carnap's favourite confirmation function c*) is obtained and is regarded as evidence, we still have a motivation for seeking new information such as $E_2$ (Achinstein, 1995, pp. 452–453). For Achinsein, the *a priori* thesis is therefore wrong.

There is another equally convincing interpretation in Rowbottom's analysis. *A priori* theorists actually regard background knowledge as working assumptions B*, i.e. the procedure yielding $E_1$ is not flawed. Rowbottom (2013) remarks that:

> Clearly, the advocate of the a priori theory will say, *this* conditional probability is positive, and indeed equal to unity. After all, [H] is entailed by the conjunction of [$E_1$ and B*]. (Rowbottom, 2013, p. 2827)

From the quote above, we can see that Rowbottom attaches an entailment condition to an *a priori* theory of evidence if the theory is intended to meet Achinstein's challenge. At the very least, $E_1$, coupled with B*, sufficiently shows the truth of H. There is no E as evidence alone for H, so on this reading, the *a priori* thesis, with a new addition of B*, is understood as:

> The only empirical fact that can affect the truth of evidential claims of the form 'E, given B*, is evidence for H' is the truth of E and B*, where E and B* in tandem guarantee the truth of the evidential claims. All other considerations are *a priori*.

A story different from Achinstein's can be told. In Rowbottom's interpretation, the reason that $E_1$ is not evidence is that B contains no working assumptions. No matter whether $E_1$, coupled with B, guarantees or sufficiently ensures the truth of H, it is undecidable *a priori* whether or not $E_1$ is evidence for H. Nevertheless, it is reasonable to claim that $E_1$, coupled with B*, is evidence for H. In the drug example, B* may represent perfectly designed and perfectly conducted trials for the treatment and control groups. Even so,

according to Rowbottom, the truths of $E_1$ and B* are a matter of empirical fact and the way to determine their truths is not so different from the way that scientists delineated by Achinstein determine the truths of E and an evidential claim that E is evidence for H. Both ways call for empirical investigations.

From the *a priori* viewpoint, similar to the logical and mathematical truths, the truths of evidential claims are never refutable. The claim that $E_1$, coupled with B*, is evidence for H is *a priori* true, thereby remaining unchanged and persisting. It is impossible that there be another piece of evidential information left out showing any flaw in $E_1$ coupled with B*. All the necessary evidential information about $E_1$ and B* is integrated into a compact package called 'evidence' that amounts to E in Achinstein's version of the *a priori* thesis. In this regard, E, namely $E_1$&B*, is empirically decidable, and the evidential claim that E is evidence for H is decidable *a priori*. Likewise, $E_1$&$E_2$, coupled with B*' (namely, there are no flaws in the studies that yield $E_1$ and $E_2$), is not evidence for H. As such, the drug example does not violate the *a priori* thesis. If Rowbottom's defence can be applied to every evidential claim, then not only is the *a priori* thesis true, but Achinstein's empirical assumption also cannot hold. Alternatively, on a less radical note, Rowbottom's defence minimally implies that the empirical assumption does not prevail over the *a priori* thesis by virtue of the drug example.[17]

It is unclear what an *a priori* theory of evidence Rowbottom has in mind. The 'working assumption' strategy requires a theory in which E, coupled with B*, entails or at least sufficiently ensures the truth of H. Which existing *a priori* approach enjoys the privilege remains uncertain. The H-D account is formulated reversely: E is evidence for H given B if and only if H&B entails E. Glymour's (1980b) bootstrap theory is not a candidate either: it holds that E

[17] Rowbottom anticipates the possible objection that B* requires further evidence from a non-flawed procedure to ensure B*, which leads to an infinite regress. He replies: '[A]ll this shows is that we have to stop somewhere, which is old epistemological news. This is just a special case of the well-known regress problem. It is not ultimately pertinent to the *a priori* thesis. In asking whether *e* is evidence for *h* we do not require that there is further evidence for *e*. We require only that *e* is true' (Rowbottom, 2013, p. 2827).

is evidence for H relative to theory T if and only if an instance of H is derived from E in concert with T. Carnap's objective Bayesian approach seems promising for this strategy as long as B* is taken into account. In the drug example, even though the degree of confirmation is sufficiently high at 0.92 in Carnap's system, it is unclear whether $E_1$ is evidence for H, since B* is not included in the calculation. At any rate, it may be worth incorporating B* into Carnap's confirmation function c* in terms of logical-linguistic properties of sentences. Perhaps Rowbottom would consider an *a priori* subjective Bayesian approach to evidence whereby P(H/E&B*) is fixed *a priori* within one's mind and updating the degree of belief with a new piece of information E′ amounts to moving from P(H/E&B*) to P(H/E′&B*'), which is also fixed *a priori* within one's mind. Going back to the drug example, if $P(H/E_1\&B^*)$ is sufficiently high or higher than the prior probability of H for researchers, $E_1$ is evidence for H for them. Supposing that the study result is not $E_1$ but $E_1$': 1000 patients presenting with S were given D, with only 50 of them having their S relieved, if researchers believe that $E_1$' is not sufficiently high nor higher than the prior, then $E_1$' is not evidence for H.

The debate is not easily settled. Achinstein and Rowbottom individually provide an alternative explanation for the drug example. However, granted that Achinstein's counter-example to the *a priori* thesis is intuitively appealing, the fact that the *a priori* thesis is refuted is one thing and the fact that the empirical assumption generally holds is another. The empirical assumption (consider (*) here) requires more: whether E is evidence for H, given B, is mostly a matter of empirical fact beyond E and B's being empirical facts. Rowbottom (2013, pp. 2831-2833) argues that the empirical assumption does not square with evidential reasoning in a particular circumstance, namely artificial intelligence, where algorithms, parameters and probability distributions are pre-set, pre-written and installed into the programmes that are employed to conduct empirical investigation such as scientific discovery. The programmes adjust the probabilities of hypotheses in response to new information. Whether E, coupled with B, is evidence for H is the *a priori* ramification. It should, nevertheless, be acknowledged that the case of artificial intelligence does not directly refute the empirical

assumption: it only indicates one inapplicable setting for this assumption. However, positive reasons for the empirical assumption are also wanting. The upshot is that this assumption had better be regarded as *prescription* rather than *description*.

## 2.4 Examining Five Theories of Evidence

Following Achinstein's analysis (2014), let us bear in mind the good-reason-to-believe assumption and the empirical assumption and examine the five theories of evidence in sequence. The H-D account of evidence holds that E is evidence for H given B if and only if (a) H in combination with B entails E and (b) E and B are true. It fails both the good-reason-to-believe assumption and the empirical assumption. To see why the H-D account may violate the good-reason-to-believe assumption, suppose that there is a fabricated law called Lepler's first law, which *mistakenly* states that the planets revolve around the sun in square orbits and that this law entails the fact that the sun exists. The existence of the sun, according to the H-D account, counts as evidence for Lepler's first law; however, it is not evidence for Lepler's first law in that P(Lepler's first law / the existence of the sun) is almost zero. The truth condition of (a) is independent of any empirical matter and can be determined *a priori* and thus it also does not meet the empirical assumption.

The error-statistical account of evidence developed by Deborah Mayo (1996) states that data E from test T is evidence for hypothesis H if and only if H has passed a *severe test* T with a result of E. Passing a severe test is determined by two criteria: *fit* and *severity*. To meet the 'fit' criterion, P(E(T); H) is not low or at least greater than P(E(T); ~H), where E(T) denotes the data E yielded by test T and P(E(T); H) denotes the probability of E(T) under the assumption that H is true. To meet the 'severity' criterion, T would yield a result that fits H less well than E if H were false, construed as P(E(T); ~H) being low.[18] Mayo wishes to draw researchers' attention to the control and

[18] E(T), instead of E, is used here to emphasise the key role of severe testing. For example, if the severity criterion is understood merely as P(E; ~H) being low, which may subject the criterion to an unfavourable situation: a rigged hypothesis can maximally fit the data, even if

assessment of a test's severity rather than just the fit. Her definition of evidence can be framed as:

> E is evidence for H relative to T if and only if P(E(T); H) > P(E(T); ~H) and P(E(T); ~H) is low.[19]

Priors and posterior probabilities, if any, are not responsible for the identification of evidence in this definition.

Note that the symbol ';' in P(E; H) representing a frequentist probability contrasts with '/' in P(E/H), the notation which Bayesians and other conditionalists typically employ for a conditional probability.[20] For a conditional probability, we need to condition only upon 'random variables or their values' (Mayo, 2018, p. 205); however, frequentists, including error statisticians,[21] maintain that P(E; H) should be interpreted as: the probability of E under the assumption that H correctly describes the data generation procedure. Here H is framed in terms of an unknown, constant parameter that defines the data-generating procedure, so there is no such thing as conditioning on a random variable or event, for example, P(Y = y/X) or P(Y = y/X = x), where X and Y denotes random variables, x and y denotes their values, and X = x and Y = y are events.[22] The frequentist probability is not

---

it is false. A severe test must come into play and rule out ways that the hypothesis may be false, which can be encapsulated in the claim that P(~E(T); ~H) is low.

[19] Here I use Mayo's (2005, p.124, footnote 3) minimal requirement for the fit probability, which will not affect the point I wish to make.

[20] Overall, frequentists and Bayesians construe 'probability' in a distinctive way, which has been described by the statisticians David Draper and David Madigan (1997, p. 18): 'In the frequentist definition of probability, you restrict attention to phenomena that are inherently repeatable under identical conditions, and define probability as a limiting proportion in a hypothetical infinite series of such repetitions. In the Bayesian approach, you imagine betting with someone about the truth of a *proposition* (which can be anything—not just an aspect of a repeatable process—whose truth value is not yet known), and ask yourself what odds you would need to give or receive to make the bet fair'.

[21] Though belonging to the frequentist camp and speak the same probabilistic language as other frequentists, error statisticians are contrasted with classical statisticians, Fisherian or Neyman-Pearsonian (Mayo, 2018, p. 55).

[22] Hypotheses and data are accorded different inferential meanings in accordance with which approach is taken: 'When we reason in a frequentist way, … we view the data as random and the unknown as fixed. When we are thinking Bayesianly, we hold constant things we know, including the data values, … — the data are fixed and the unknowns are random' (Draper and Madiganv, 1997, p. 18).

akin to the formal Bayesian computation. For the sake of argument, let us assume that error statisticians and conditionalists have common ground to continue the discussion on evidence, for which error statisticians would provisionally grant the conditional probability symbol '/'.[23]

To examine whether this theory meets the good-reason-to-believe assumption and the empirical assumption, consider an example modified from Colin Howson (1997). Suppose that a subject undergoes a test that can detect whether one has disease D. If the subject has the disease ($H_s$), the test yields a positive result (E(T)) 95% of the time; therefore $P(E(T)/H_s)$ = 95%. If the subject does not have disease D ($\sim H_s$), the test yields a positive result (E(T)) 1%; therefore, $P(E(T)/\sim H_s)$ = 1%. $P(E(T)/H_s) > P(E(T)/\sim H_s)$ shows that the positive result fits the hypothesis that the subject has D, and $P(E(T)/\sim H_s)$ being low shows that the test yielding the positive result has high severity. The positive result, nevertheless, may not count as evidence for the hypothesis that the subject has disease D. Suppose that the prevalence of D in a certain population is very low, say 1/100,000. This can reasonably be taken as the prior probability, $P(H_s)$ = 1/100,000, meaning that the subject is very unlikely to have D before the test. The posterior probability is then expressed as:

$$
\begin{aligned}
& P(H_s/E(T)) \\
&= P(E(T)/H_s)P(H_s) / [P(E(T)/H_s)P(H_s)+P(E(T)/\sim H_s)P(\sim H_s)] \\
&= (95\% \times 1/100{,}000) / [(95\% \times 1/100{,}000 + 1\% \times (1-1/100{,}000)] \\
&\approx 0.095\%.
\end{aligned}
$$

The positive result counts as evidence in the error-statistical account, but this example blatantly violates Achinstein's good-reason-to-believe assumption.[24] The empirical assumption, however, is satisfied in the error-statistical theory since the test is an empirical matter and $P(E(T)/H_s)$ is also empirically

[23] Mayo (2005, p. 112) in fact took this strategy when claiming that 'we use the conditional probability symbol "/", since Achinstein is mounting a Bayesian criticism'.
[24] This exemplifies the so-called *base-rate fallacy*, which occurs when the prior probability of the disease (the 'base rate') is ignored (Howson and Urbach, 2006, p. 24).

decidable.

It is disputable whether the error-statistical account does not meet the good-reason-to-believe assumption (see Achinstein, 2001, pp. 134–140; Mayo, 2005). Note that Mayo (2005, p.113) eschews prior probabilities, holding that priors 'are nearly always unavailable or irrelevant for scientific contexts'. She doubts whether the argumentative move that equates the priors concerning the population and the priors concerning the individual of interest is defensible. Regarding the foregoing example, despite accepting that P(H) can be objectively obtained through a random sampling from the given population with a particular relative frequency of disease D, 1/100,000, Mayo rejects the assumption: P(H: a person randomly selected from the population has D) = P($H_s$).[25] Mayo (2005) considers such an assumption fallacious, which she calls the *fallacy of probabilistic instantiation*; given P(H) = 1/100,000, it does not follow that P($H_s$) = 1/100,000. From the frequentist perspective, $H_s$ is either true or not, even if *this, not another,* subject is randomly selected from the population; there is no such thing as assigning probability to a *unique* event. By contrast, H concerning *some* subject randomly selected from the population is viewed as a *generic* type of event, or a random variable, such that P(H) is admissible (Mayo, 2018, p. 407). Instead of pursuing the debate,[26] I simply wish to point out that in this kind of well-specified case, many medical scientists nevertheless accept the figures based on the frequencies (e.g. a clinician estimates the post-test probability of a *unique* patient by using sensitivity, specificity and prevalence) (Florkowski, 2008). Those who accept priors concerning the individual would agree that the good-reason-to-believe assumption fails here.

This does not mean that Mayo's conception of evidence is inferior to Achinstein's. In terms of the good-reason-to-believe assumption, Mayo's account looks unsatisfactory, but this is because this assumption has equated a necessary condition for evidence with high posterior probability,

[25] This is called an *empirical* or *frequentist* prior (Mayo, 2018, p. 185).
[26] Achinstein and Mayo's debate ensued (See Achinstein, 2010b; Mayo, 2010).

one that is objected to by Mayo in the belief that the priors regarding unique events are dubious. Mayo (1997, p. 325, italics in original) defends the error-statistical account elsewhere writing: 'Since calculating posterior probabilities via Bayes' theorem requires the introduction of prior probabilities (to an exhaustive set of hypotheses) and since these are unavailable in testing scientific hypotheses regarded as true or false, posteriors are *quite deliberately not* made the goal of NP tests'. Furthermore, in the diagnosis example, P(E(T)/H) / P(E(T)/~H) is extremely high. In Bayesian terms, the high likelihood ratio still boosts the probability of H even with the extremely low prior probability. This conception of evidence is in part captured by Carnap's (1962, p. xvi) *confirmation as increase in firmness*: E (incrementally) confirms H relative to B if and only if P(H/E&B) > P(H/B).

The satisfaction view of evidence is illustrated by Hempel's theory of confirmation and Glymour's bootstrap theory. Glymour's (1980b) bootstrap theory advocates Hempel's core idea that evidence is a positive instance of a hypothesis and resolves some problems that Hempel's theory of confirmation encounters. Theoretical terms for example, are not included in Hempel's theory of confirmation, which is enumerative induction, so that instances cannot be used to confirm a theory containing theoretical vocabulary. The claim 'a raven is black' confirms 'all ravens are black' but in Hempel's account, the claim 'the sun produces certain types of lines in a spectrograph of light' cannot confirm the hypothesis 'all stars contain Helium'. For now, I shall concentrate on Glymour's bootstrap theory and discuss whether it meets the good-reason-to-believe assumption and the empirical assumption.

Glymour's bootstrap theory holds that E is evidence for H relative to theory T if and only if an instance of H is derived from E in concert with the hypotheses (H′, H″, etc.) other than H in T. T is comprised of H, H′, H″ and so forth, which involve theoretical (unobservable) vocabulary. To see whether E confirms a particular H amounts to checking whether other Hs within T, combined with E (observational), derive an instance of H. Appealing only to certain portions of T itself to establish evidential relevance between E and a targeted H, repeating this procedure to other Hs, and eventually confirming T

as a whole are as much as ‘T’s pulling itself up by its own bootstraps’.

Glymour’s (1980b, p. 150) main example from the history of science presents a brief sketch of how this theory operates. Newton’s law of universal gravitation (H) describes the inverse-square property of forces between bodies. In support of the law, Newton considered Kepler’s laws as evidence (E), which, however, do not contain the theoretical term ‘force’. Appealing to his second law of motion (other H in T), which mentioned the term ‘force’ (by stating that the force acting upon a body is the product of its mass times its acceleration), Newton eventually derived the law of universal gravitation from Kepler’s laws combined with the second law of motion. According to Glymour’s bootstrap theory, Kepler’s laws are evidence for Newton’s law of universal gravitation.

However, Achinstein’s (1983, pp. 358–359) example indicates that Glymour’s bootstrap theory does not satisfy the good-reason-to-believe assumption. Consider theory T, which consists of the following two equations, or hypotheses:

$H_1$: A (the total force acting on a particle) = C (the quantity of God’s attention focused on a particle);
$H_2$: B (the product of a particle’s mass and acceleration) = C.

The values of A and B are measured by experiments. According to Glymour’s bootstrap theory, a claim E describing both A and B is evidence for $H_1$ relative to T, on the basis that B and $H_2$ jointly determine the value of C, equal to that of A obtained experimentally and in turn, $H_1$ is derived. But there is no support that P($H_1$/the values of A and B) is greater than half. Nor does Glymour’s theory satisfy the empirical assumption, because an instance of H being derived from E in concert with T is *a priori* decidable.

On the view of the subjective Bayesian theory of evidence, E is evidence for H given B if and only if P(H/E&B) > P(H/B), where a probability is interpreted as the degree of belief and the only constraint is that the degree of belief

should obey the probability axioms. It does not satisfy the good-reason-to-believe assumption. For instance, the fact that Robert buys a lottery ticket is insufficiently strong to be evidence that he will win even though, according to the subjective Bayesian theory, it increases the probability of Robert winning the lottery. The subjective Bayesian theory, however, generally meets the empirical assumption, because a probability is viewed as the degree of belief, which is determined by a matter of empirical fact.[27]

The final standard theory of evidence, the objective Bayesian theory of evidence, does not meet the good-reason-to-believe assumption and the empirical assumption if we adopt Carnap's interpretation of probability.[28] For Carnap (1962), the probability is fixed by the relationships between the logical-linguistic properties of sentences, for instance, H and E.[29] When it comes to an evidential relation, two definitions of evidence can be considered. The first is concerned with Carnap's *qualitative* concept of confirmation, same as in subjective Bayesianism: E is evidence for H given B if and only if P(H/E&B) > P(H/B). The second is Carnap's *quantitative* concept of confirmation, in which E is evidence for H given B if and only if P(H/E&B) attains a certain high degree *r*. The first definition does not fulfil the good-reason-to-believe assumption, just as subjective Bayesianism fails to do. However, there is an example to illustrate that the second definition fulfils the good-reason-to-believe assumption but does so in a trivial manner. Given that people die under the age of 150 (B), the fact that Mr. Y drinks water every day (E), according to objective Bayesianism, counts as evidence that he will die before the age of 150 (H) because P(H/E&B) is equal to one,

[27] This is not always the case. Some, like de Finetti (1937 pp. 106–107), may take an *a priori* subjective Bayesian position.
[28] Objective Bayasians can adopt different interpretations. For example, one is that objective Bayesianism should conform to the three norms: probability, calibration and equivocation (Williamson, 2010). For more details, see Chapter 5 in this thesis.
[29] The logical-linguistic relations can be explicated by the following example. Suppose that there are four collectively exhaustive descriptions (*state descriptions* in Carnap's terminology) of H and E, namely, (H&E), (H&~E), (~H&E) and (~H&~E). The degree of confirmation is expressed in terms of logical measure functions $m(\cdot)$, which assign numbers to each description under the constraint that the sum of the measures should be equal to unity. Assume $m$(H&E) = 1/3, $m$(H&~E) = 1/6, $m$(~H&E) = 1/6 and $m$(~H&~E) = 1/3. Then P(H/E) = *r* is obtained: P(H/E) = $m$(H&E) / $m$(E) = $m$(H&E) / $m$([H&E] v [~H&E]). In this case, the value of P(H/E) = (1/3) / (1/3+1/6) = 2/3.

though P(H/B) already equals one. Carnap's objective Bayesian theory nonetheless fails to meet the empirical assumption, because whether P(H/E) attains a certain level is *a priori,* depending only upon logical-linguistic relations and calculation.

## 2.5 Explanation as Evidential Relevance

Relevance matters when it comes to evidential reasoning. Even if there is an account of evidence that regards drinking water as evidence for dying before the age of 150 and that meets the good-reason-to-believe assumption and the empirical assumption, we may still feel reluctant to call it evidence. This is because the maximum life span for humans does not exceed 150, regardless of whether one drinks water or not. Achinstein thus offers his explanationist theory of evidence that he believes meets the two assumptions and avoids irrelevance of H to E. For convenience, I repeat his theory here: E is evidence for H given B if and only if (a) P(Exp(H,E)/E&B) > 1/2; (b) E and B are true; (c) E does not entail H. Achinstein argues that the high probability condition is a necessary but not a sufficient component to an account of evidence, and thus employs the explanatory connection between E and H to secure the relevance between E and H. This is, in my terms, where a relevance-mediating vehicle comes into play. The explanationist theory employs explanatory connection as the relevance-mediating vehicle by which E is labelled evidence.

In order to appropriately characterise potential evidence, the kind of explanatory connection required should be clarified. Regarding the clarification, two considerations should be applied. For one thing, it is too stringent to require only the common uni-directional explanation in which H explains E (Achinstein, 2001, p. 150). The examples of no heartbeat and the 101st toss of a coin in the opening of this chapter are cases in point. For these reasons, Achinstein urges multi-directional explanatory relationships between H and E, namely, H explains E, E explains H, or another hypothesis explains both.

Furthermore, it is too stringent to require that there *exist* an explanatory connection between H and E (Achinstein, 2001, pp. 150–151). To illustrate this point, consider the example of a treatment. Suppose that treatment T was confirmed to be 95% effective in relieving symptoms S (E) in well-controlled trials. With E, it was predicted that Henry's symptoms S would be relieved after he took T (H). His symptoms S ended up being relieved not because of treatment T, but because his immune system adapted its response. Nevertheless, E should be regarded as (potential) evidence for H regardless of whether or not E does explain H. Like the footage example above, it is sufficient that there is *likely* to be a constant explanatory connection between H and E.

By synthesising the two considerations, Achinstein (2001, p. 151) concludes that '[w]hat is needed for potential evidence is not that there is an explanatory connection between h and e, but that, assuming the truth of both e and h, there probably is such a connection'. That there is likely to be an explanatory connection is formalised with an apparatus shown in the condition (a) P(Exp(H,E)/E&B) > 1/2, as noted at the outset of this chapter. The condition (a) entails that P(H/E&B) > 1/2, the latter of which is a necessary condition of E being good reason for H. Achinstein (2001, p. 153, italics in original) holds that 'from the definition of *explanatory connection*, "there is an explanatory connection between *h* and *e*" entails *h*', which amounts to Exp(H,E) entailing H. It can be shown that P(Exp(H,E)/E&B) = P(H/E&B) x P(Exp(H,E)/H&E&B).[30] The proof is as follows:

[30] P(H/E&B) > 1/2 is entailed with the following proof: given that $p = qr$, $p > 1/2$, $0 \leq q$, $r \leq 1$, we obtain $q = p/r \geq p > 1/2$. By substituting P(Exp(H,E)/E&B) for $p$, P(H/E&B) for $q$, and P(Exp(H,E)/H&E&B) for $r$, we obtain P(H/E&B) > 1/2.

P(Exp(H,E)/E&B)
= P(Exp(H,E)&E&B) / P(E&B)
= [P(H&E&B) / P(E&B)] × [P(Exp(H,E)&E&B) / P(H&E&B)]
= P(H/E&B) × [P(Exp(H,E)&H&E&B) / P(H&E&B)]
[since Exp(H,E) entails H]
= P(H/E&B) × P(Exp(H,E)/H&E&B).

If, given E and B, it is more than 50% likely that there is an explanatory connection between H and E, then, given E and B, E is a good reason to believe H. However, if only P(H/E&B) is high, it, at most, shows that there is *some* good reason that H is true, but not necessarily as a result of E (i.e. E may not be *that* reason). Moreover, the fact that P(H/E&B) and P(Exp(H,E)/H&E&B) are both high does not ensure P(Exp(H,E)/E&B) > 1/2. Suppose P(H/E&B) = 0.8 and P(Exp(H,E)/H&E&B) = 0.6. It follows that P(Exp(H,E)/E&B) = 0.8 × 0.6 = 0.48, which is smaller than 0.5, meaning that E is not a good reason to believe H.

The empirical assumption is met by (a) P(Exp(H,E)/E&B) > 1/2 and (b) E and B are true. The truth of E and of B is an empirical matter. Whether a high probability exists of there being an explanatory relation between H and E is neither a matter of logic relations nor a matter of mathematical relations, so (a) fulfils the empirical assumption.

Equipped with the relationship between explanatory connection and probability, we can understand the distinction between *veridical* evidence and *conclusive* evidence within Achinstein's framework. Conclusive evidence guarantees the truth of H, yet veridical evidence does not necessarily do so (Achinstein, 2001, p. 27). Speaking of conclusive evidence is of use to avoid a misunderstanding, i.e. that veridical evidence requires that H be true and therefore verifies H with certainty. Conclusive evidence can be defined as:

*E is conclusive evidence for H given B* if and only if
(a) E is veridical evidence; and
(b) P(H/E&B) = 1.

Consider the birthday example to illustrate the distinction between veridical evidence and conclusive evidence. If there are fifty people in a room (E), how likely is the hypothesis (H) that at least two people share the same birthday? Very high; P(H/E) equals to approximately 0.97. Suppose no other reasons for H (e.g. both are picked deliberately) are available, it is obtained that P(Exp(H,E)/H&E&B) = 1. Suppose further that there are two people's birthdays on the same day in the room (i.e. H is true) and that there is an explanatory connection between H and E. It follows that E is veridical evidence for H. However, since P(H/E) is not equal to unity, E, as per the definition given above, is not conclusive evidence for H.

## 2.6 Examining Achinstein's Theory of Evidence

Several difficulties with Achinstein's account of evidence should be discussed here. Achinstein (2014, p. 383) asserts that 'Thomson's later experimental results were evidence for Thomson's charged particle hypothesis in all four senses of "evidence"'. We need not know everything about Thomson's experiment to see why Achinstein must be mistaken. It seems impossible for Achinstein to ensure whether the evidence Thomson's experiment presented lacks any flaw and the hypothesis is true, since Larry Laudan's (1981) pessimistic induction argues that once-successful past theories have typically turned out to be false, illustrating that any theory can possibly be falsified unless we are God. Achinstein might insist that Thomson excluded the interfering factors to select the only one possible explanation that electrical effects occur, so this is the only true explanation for the phenomenon. However, Hertz also thought he excluded any interfering factors in his experiment and found no electrical effect, so we do not know whether we have potential evidence and veridical evidence although finding them is our aim.

Perhaps Achinstein just wanted to illustrate the notion of veridical evidence by supposing that Thomson was right, rather than *arguing* that Thomson was in fact right. However, we should be careful to read his relevant sentences,

such as '[W]hat the scientist seeks is veridical evidence. Usually, when a scientist claims that some experimental result is evidence that a hypothesis is true, he can be construed as making a claim using this concept' (Achinstein, 2014, p. 391). It is misleading, and can be true only on this subjective reading: the scientist *believes* that the experimental result is veridical evidence for the hypothesis. Veridical evidence is rarely recognisable particularly when hypotheses of interest involve unobservable entities (e.g. quarks, intelligence and economic inequality), unobserved entities (e.g. extinct organisms and languages), and unobserved processes (e.g. the process of a certain murder). If there is any veridical evidence for the unobserved, it is likely to be obtained only with hindsight. For example, irregularities in Uranus's orbit were observed, which were not explained by Newton's law of gravitation. The irregularities were explained by the hypothesised existence of a perturbing body and were therefore potential evidence for the existence of the planet Neptune. With advances in astronomical technologies, the hypothesis that Neptune exists has been confirmed to be true, at this point the irregularities have been converted from previous potential evidence into veridical evidence.

Another difficulty is that Achinstein (2014, p. 391) argues that his definition of evidence rather than the aforementioned theories of evidence 'can be used to define each of the four concepts of evidence distinguished' because the definition captures the core sense of evidence and thus forms a basis for the four concepts. If we combine Carnap's qualitative and the quantitative concepts of confirmation, we will have a new objective Bayesian account that satisfies both the good-reason-to-believe assumption and the empirical assumption:

> **The Merged Objective Bayesian Account of Evidence (MOB):**
> *E is evidence for H given B* if and only if
> (i) P(H/E&B) > P(H/B);
> (ii) P(H/E&B) attains a certain high level above 1/2.

This definition meets the good-reason-to-believe assumption by virtue of the

condition (ii). The MOB account also does not make trivial cases evidence as seen in the drinking-water example, where P(H/E&B) = P(H/B) although P(H/E&B) = 1, because the account requires condition (i). As for the empirical assumption, conditions (i) and (ii) are both empirically decidable if the *a priori* Bayesian interpretation of probability is not adopted.

The MOB account seems as favourable as the explanationist theory of evidence. To illustrate this point, consider the Monty Hall problem, that is a special case because it is often cited as a stumbling block for theories of evidence (e.g. Fitelson, 2007):

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host [Monty Hall] who knows what's behind the doors, opens another door, say No.3, which has a goat. He then says to you, "Do you want to pick door No.2?" Is it to your advantage to switch your choice? (vos Savant, 1990)

In the following argument, the letters $H_i$ and $E_j$ mean the following:

$H_1$: The prize is behind door 1.
$H_2$: The prize is behind door 2.
$H_3$: The prize is behind door 3.
$E_1$: The contestant randomly picks door 1.
$E_3$: The host opens door 3, which is not the location of the car.

Since the contestant's initial choice is independent of the place of the car, $P(H_1/E_1) = P(H_2/E_1) = P(H_3/E_1) = 1/3$, but when the host's opening door 3 is added in, $P(H_1/E_1\&E_3) = 1/3$, $P(H_2/E_1\&E_3) = 2/3$ and $P(H_3/E_1\&E_3) = 0$.[31] The probability of $H_3$ conditional on $E_1$ becomes zero when $E_3$ is further considered on the basis that $E_3$ eliminates $H_3$. As for the remaining two

[31] The direct calculation with Bayes' theorem proceeds as follows: $P(E_3/H_1\&E_1) = 1/2$ (the prize is behind door 1, so the host can randomly pick and open either door 2 or door 3); $P(E_3/H_2\&E_1) = 1$ (the prize is behind door 2 and the contestant already chooses door 1, so the host must open door 3); $P(E_3/H_3\&E_1) = 0$ (the prize is behind door 3, so it is impossible for the host to open door 3). Therefore, $P(H_2/E_1\&E_3) = P(E_3/H_2\&E_1)P(H_2\&E_1) / P(E_1\&E_3) = [P(E_3/H_2\&E_1)P(H_2\&E_1)] / [P(E_3/H_1\&E_1)P(H_1\&E_1) + P(E_3/H_2\&E_1)P(H_2\&E_1) + P(E_3/H_3\&E_1)P(H_3\&E_1)] = 1 / (1/2+1+0) = 2/3$.

doors, the contestant, on the basis of a new piece of evidence $E_3$, should switch to door 2 to win the car.[32]

The explanationist theory can account for $E_3$'s being evidence for $H_2$. Given that the contestant initially picks door 1 ($E_1$), the prize's being behind door 2 explains the host's opening door 3 as the host must open door 3 with the intention of not exposing the location of the car. Thus, $P(Exp(H_2,E_3)/E_1\&E_3) = P(H_2/E_1\&E_3) \times P(Exp(H_2,E_3)/H_2\&E_1\&E_3) = 2/3 \times 1 = 2/3$, which is greater than 1/2, as required in the definition of evidence in the explanationist theory.

While it may be difficult to compute the probability of an explanatory connection for real-life cases, the apparent difficulty of computing that for $H_1$ and $E_3$ can be cleared up, once the explanatory connection is fleshed out. $P(Exp(H_1,E_3)/E_1\&E_3) = P(H_1/E_1\&E_3) \times P(Exp(H_1,E_3)/H_1\&E_1\&E_3) = 1/3 \times P(Exp(H_1,E_3)/H_1\&E_1\&E_3)$. Seemingly, $P(Exp(H_1,E_3)/H_1\&E_1\&E_3)$ is underdetermined, because when the prize is behind door 1, the host can open either door 2 or door 3 freely, which corresponds to the random opening of door 3. That is, the prize's being behind door 1, relative to the contestant's random picking of door 1, does not seem to explain why the host opens door 3. Where one claim explains another, the explanation does not require that its explanans all be present for the explanatory purpose. What is required is that one member of the explanans set should make a difference to the explanandum, other necessary members in the explanans set being equal. For example, suppose a window is broken (W) as a ball with certain momentum hits it (B) and the window is made of glass (G). B and G in tandem explain W, but usually multiple factors influence the occurrence of W. When other necessary members are present (G in this example), it is not necessary to mention all of them to sufficiently explain an event. Instead, it

[32] These probabilities baffle many, as they think that $P(H_1/E_1\&E_3) = P(H_2/E_1\&E_3) = 1/2$, which indicates that regardless of whether the contestant switches doors, the chance of winning the car remains unchanged because the probability of the car being behind the door opened should be divided and distributed equally between the remaining two options (Falk, 1992). This is correct if the scenario is rewritten: There is an alien with a UFO landing onto the stage when the host opens the door, and this alien facing two unopened doors is asked 'which door do you choose?'. The chance to win the car is 1/2 whichever door is chosen. However, 'that's because [this alien] lacks the advantage the original contestant had—the help of the host' (vos Savant, 1990).

can be claimed that B explains W relative to G. Since *no* reasons are available other than the prize being behind door 1, relative to the contestant's picking of door 1, to explain the opening of door 3, $P(Exp(H_1,E_3)/H_1\&E_1\&E_3) = P(E_3$ because of $H_1$ given $H_1$, $E_1$ and $E_3)$, which is one.[33] It follows that $P(Exp(H_1,E_3)/E_1\&E_3) = P(H_1/E_1\&E_3) \times P(Exp(H_1,E_3)/H_1\&E_1\&E_3) = 1/3 \times 1 = 1/3$.

The MOB account also allows us to claim that $E_3$ is evidence for $H_2$. For its two conditions are met as $P(H_2/E_1\&E_3) = 2/3$, which is greater than 1/2 and $P(H_2/E_1\&E_3) > P(H_2/E_1)$ (=1/3). In addition, as mentioned above, the MOB account meets the good-reason-to-believe assumption and the empirical assumption and avoids trivial cases, so according to Achinstein's characterisation of evidence, no reason favours the explanationist theory over the MOB account with respect to this case.

What is worse is that some intuitively acceptable evidence is covered by the MOB account but is not covered by the explanationist theory. Let us consider the murder of Russia's mad monk, Rasputin.[34] A version of the whole process is described as follows:

> Rasputin was invited to visit Yusupov's home and, once there, was given poisoned wine and tea cakes. When he did not die, the frantic Yusupov shot him. Rasputin collapsed but was able to run out into the courtyard, where Purishkevich shot him again. The conspirators then bound him and threw him through a hole in the ice into the Neva River, where he finally died by drowning. (Encyclopedia Britannica, 2019)

[33] Achinstein does not provide details about how to calculate the probability of explanatory connection. He simply assumes figures to illustrate his point. For example, Achinstein (2001, p. 154, italics in original) writes: 'Suppose that 70% of those with symptoms *S* have relief of those symptoms in a week. Writing *R* for getting relief in a week, and *S* for having symptoms, we have $p(R/S) = .7$. [Footnote omitted] Suppose that 70% of those with *S* who take medicine *M* get relief in a week, so $p(R/S\&M) = .7$. Suppose finally that among those with S who take *M* and get relief in a week, 60% get relief *because* they took *M* (while 40% do so for other reasons), so $p(R$ because of $M/R\&S\&M) = .6$. From the previous theorem p(*R* because of *M*/S&M) = $p(R$ because of $M/R\&S\&M) \times p(R/S\&M) = .6\times.7 = .42$. That is, 42% of those with symptoms *S* who take medicine *M* get relief because of *M*. *M* is effective in relieving *S* only 42% of the time. (And, in fact, taking *M* doesn't increase the chances of relief in a week, since $p(R/S\&M) = p(R/S) = .7$'.

[34] Michael Strevens (2008) uses this case when discussing causal and explanatory relevance.

For simplicity's sake, let us consider only the two actions that the assassins would anticipate when plotting the assassination: shooting and drowning.

H: Rasputin died.
$E_1$: Shooting had been carried out.
$E_2$: Drowning had been carried out.
B: Rasputin had been invited out to Yusupov's home.

Suppose $P(H/E_1\&B) = 1/2$ and $P(H/E_2\&B) = 1$, from which it follows that $P(H/E_1\&E_2\&B) = 1$.[35] Suppose also that drowning would be carried out immediately after shooting, that there was no way to know whether the target, from being shot at to being thrown into the river, was dead or alive, and that the post-mortem could not be conducted due to his body being missing. The assassins should have been completely confident of Rasputin's death when they decided to carry out the shooting and drowning. Intuitively, $E_2$ alone suffices as evidence that Rasputin died as the success rate of murder by drowning is unity. $E_2$ is regarded as evidence, relative to $E_1$ as background knowledge, for the MOB account, because $P(H/E_2\&E_1\&B) = 1$, which is greater than 1/2, and $P(H/E_2\&E_1\&B) > P(H/E_1\&B)$.

However, $E_2$ is not evidence for the explanationist theory, since $P(Exp(H, E_2)/E_2\&E_1\&B) = P(Exp(H,E_2)/H\&E_2\&E_1\&B) \times P(H/E_2\&E_1\&B) = 1/2 \times 1$, which is not greater than half. For the explanatory connection, $P(Exp(H,E_2)/H\&E_2\&E_1\&B)$ is equal to the probability of drowning explaining Rasputin's death, given that Rasputin died and the shooting and drowning had been executed. The question is: how much do the shooting and the drowning individually explain Rasputin's death, given all of the above conditions? Clearly, neither of them can fully explain his death on its own. Rasputin had a fifty-fifty chance of being shot dead, and even if Rasputin survived the shooting, he would end up being drowned. The probability for

[35] Proof: $P(H/E_1\&E_2\&B) = P(H\&E_1\&E_2\&B) / [P(H\&E_1\&E_2\&B) + P(\sim H\&E_1\&E_2\&B)] = 1$, since $P(\sim H\&E_1\&E_2\&B) = P(\sim H/E_2\&B)P(B/E_2)P(E_2)\ P(E_1/\sim H\&E_2\&B) = [1- P(H/E_2\&B)] \times P(B/E_2)P(E_2)\ P(E_1/\sim H\&E_2\&B) = 0 \times P(B/E_2)P(E_2)\ P(E_1/\sim H\&E_2\&B) = 0$.

each situation is half and thus $P(Exp(H,E_2)/H\&E_2\&E_1\&B) = 1/2$, with $E_2$ alone not entailing H. This is a counter-example to the explanationist theory, but not to the MOB account.

It seems that the MOB account is more favourable than the explanationist theory, because it not only meets the two assumptions and avoids trivial cases, but also makes sense of some evidence-related cases when the explanationist theory cannot do so. However, the defence of the objective Bayesian account of evidence may not be decisive. I offer an example here: suppose that there are a total of 1,000,000 lottery tickets and Robert has 900,000 of them (B). Robert obtains another lottery ticket from his friend (E), so P(H: Robert will win the lottery/E&B) > P(H/B) and P(H/E&B) > 1/2. According to the MOB account, the fact (E) that Robert obtains one more lottery ticket counts as evidence that he will win the lottery (H) relative to B. However, it is problematic that E counts as evidence for H here. Achinstein's theory of evidence should exclude the case as evidence on the basis that E and H lack an explanatory relevance between them because the lottery ticket adds a mere extra 1/1,000,000 probability of winning the lottery. The additional lottery ticket does not correctly explain that Robert will win the lottery. Nor does Robert's winning explain his obtaining ticket; nor do both have any reason in common that explains them. The MOB account, rather than the explanationist theory, is subject to this counter-example.

## 2.7 Conclusion

To conclude, without the aim of involving evidence as guidance on further investigation into a hypothesis, Achinstein's explanationist theory of evidence attempts to provide a framework that employs explanatory relations to connect hypothesis and evidence and can inform us about when E is strong enough to be evidence for H. In the purpose-specific spirit, I have evaluated his theory and have identified several of its difficulties: the notions of potential evidence and veridical evidence are hardly epistemically attainable and the MOB account of theory that meets the good-reason-to-believe and empirical assumptions in a non-trivial manner can make sense of other

sources of evidence that do not count as evidence for the explanationist theory. So far I have not discussed the limitations of the explanationist theory. In some settings, its relevance-mediating vehicles of explanatory threshold do not consider what should be labelled as evidence. This issue will be illustrated in Chapter 5 with cases of historical linguistics.

# 3

# Metaphysical Approach: Cartwright's Argument Theory of Evidence

There is no evidence *simpliciter*. Evidence does not exist in a one-place relation in a way that an electron or a tree is recognised by its characteristics. Nor is evidence identified relative only to a hypothesis in a two-place relation, as assumed in specific versions of the hypothetic-deductive (H-D) theory, which stipulates that any claim, without background knowledge, directly deduced from a hypothesis is evidence for the hypothesis. Instead, many philosophical theories of evidence subscribe to a view that evidence is a three-place relation;[36] evidence for a hypothesis is always relative to a *relevance-mediating vehicle*, which I have defined in Chapter 1.

In this chapter, I pay particular attention to Nancy Cartwright's argument theory, in which the mediating vehicle for evidential relevance is a sound argument.[37] Cartwright (2013) holds that a good argument ties some empirical claims together: the claim or claims in the form of the premises and the claim appearing as the conclusion are evidence and hypothesis respectively. From the tool-based view I endorse, I shall identify the strengths and limitations of the argument theory. Upholding the same spirit, I will show that the difficulties that the argument theory encounters mostly arise from its Platonic characterisation of evidence, i.e. that metaphysical relations are built on sound arguments.

[36] I list some three-place theories of evidence: (some versions of) the H-D confirmation (Achinstein, 2014; Sprenger, 2011), subjective Bayesianism (de Finetti, 1937; Ramsey, 1926;), objective Bayesianism (Jaynes, 1968; Williamson, 2010), the bootstrap theory (Glymour, 1980b), the explanationist theory (Achinstein, 2001, 2014), the error-statistical account (Mayo, 2004) and the inferentialist account (Reiss, 2015a).

[37] A sound argument refers to one that is a valid argument and in which its premises are all true.

## 3.1 Dissection of the Argument Theory by RCTs

Cartwright (2013) develops her argument theory in the course of her work on *what randomised controlled trials (RCTs) are evidence for*. To understand what the theory looks like in concrete terms, and to see what her motivation for it is, it is useful to follow her here. Within many special sciences, such as medicine and psychology, the establishment of causal claims rests largely on RCTs, which are normally characterised in these terms: random assignment and blinding in the hope of balance, at least in the mean, of the net effects of other causal factors. Cartwright then asks: under what conditions can RCTs *clinch* causal conclusions? Why do many results of well-conducted RCTs fail to extrapolate to the target setting? Cartwright (2012) argues that the key to answering these questions is to recognise that the foundation for evidential claims is the argument.

The movements of evidence-based medicine (EBM) and of evidence-based policy (EBP) have invested enormous resources and effort in apprising us of a hierarchy of evidence of different quality, which is considered to improve the quality of decision-making. Among these standards, RCTs are widely heralded as the gold standard for evidence that is taken to establish causal claims in health, medicine, social sciences and policy deliberation. These claims are of two kinds: claims about efficacy and claims about effectiveness. A study that establishes *efficacy*, or internal validity, is one that 'confers a high probability of truth on the result of the study' (Cartwright, 2010, p. 60). When a study is internally valid, it, by reducing bias, draws a causal conclusion that the changes in interventions (e.g. a medical treatment) bring about the differences between different groups (e.g. treatment group versus control group). *Effectiveness*, or external validity, relates to 'whether the result that is established in the study will be true elsewhere' (Cartwright, 2010, p. 60). In contrast to internal validity, external validity is of more pragmatic value: evidence from high-quality systematic reviews, meta-analyses of RCTs or more than one low-risk-of-bias RCT is ranked highest in terms of using causal claims established to make

predictions about whether a claim that seems to hold in one study will hold elsewhere.

Investigating the preconditions for establishing causal claims using RCTs and the predictive success of RCTs, Cartwright extends the concept of necessary premises in a valid argument to any relationship between hypothesis and evidence. This is the central thesis of her argument theory of evidence. Evidence is a three-place relation in which a claim E is evidence for hypothesis H relative to an argument A if and only if E figures essentially in A and A is a valid and sound argument for H:

> A well-established empirical claim *e* is evidence for hypothesis *h* relative to a good argument A (or A, A′, A″…) if and only if *e* is a premise in A, which is itself a good argument for *h* (or, is a premise in A′ which is a good argument for a premise in a good argument A for *h*, etc.), where a good argument has true premises and is deductively valid. (Cartwright, 2013, p. 5)

When we reconstruct an argument underpinning evidential relationships, it becomes apparent where the gaps that prevent us from reasoning from evidence to conclusion are located.

In science and policy making, rigour is essential when it comes to evidence. In the argument theory of evidence, rigour matters in two distinct respects: material and formal. This distinction is thought to originate with Rudolf Carnap (1937, pp. 302–303), according to whom the material mode of speech is about the world and the formal mode of speech is about language itself. When this distinction is applied to evidence, the material mode has to do with ‘what facts of Nature there are and what other facts can ensure they obtain’ (Cartwright, 2013, p. 5), whilst the formal mode concerns ‘our hypotheses about what facts obtain and the further hypotheses that provide warrant for them’ (Cartwright, 2013, p. 6). It can be seen that the argument theory emphasises the formal mode in requiring true premises and deductive entailment. However, this thematisation corresponds to facts and their relationships in the material mode; the relationships in the world are fixed

and can, in some cases, be conceived of as causal relationships.[38] It is both — these facts, rendered as true premises, and these material-mode relationships, rendered as valid arguments — that mediate evidential relationships. Both modes are concerned with the metaphysical level in my sense: these fixed relations in the world hold independently of us. These two modes are distinct from the claim that 'an agent is justified in believing in H', which indicates an epistemological attitude.

### 3.2 Argument of Efficacy of RCTs

The argument theory provides a framework for appraising how RCTs show evidence for causal efficacy. RCTs can be internally valid because if the confounding factors were actually balanced between the treatment and control groups, positive results of an RCT imply the causal conclusion that treatment T brings about outcome O in some subset of the study population. Confounding factors are ones that cause O via causal pathways except the pathway(s) along which T causes O. In an RCT, confounding factors must be controlled if we are to be assured of the causal conclusion that T causes O, drawn from the fact that the probability of O in the treatment group is greater than in the control group.

To put this point in my terms, what the argument theory demonstrates is that a relevance-mediating vehicle that enables the causal conclusion to be drawn from evidence of an ideal RCT can be an argument. Suppose an ideal RCT is performed in which treatment T and outcome O take binary values (e.g. ±), where T is the cause under test (treatment) and O, the putative effect (outcome). By 'ideal' I mean that the treatment and control groups are balanced with respect to the net effect of all other causal factors. As such, certainly this is far from any real experiment when the best we can hope for is balance in the long run, and imbalance is to be expected in any real RCT.

[38] The standard philosophical view is that the causal relata are *event*s, which has been challenged by a number of philosophers. Some champion *facts* as the ideal candidate for the causal relata (Bennett, 1988; Mellor, 1995), but this is not the point here, so I make no further enquiry into the causal relata.

Suppose that the RCT result shows that the probability of O (outcome) in the treatment group (treatment T is given) is greater than in the control group (treatment T is not given). The result can be obtained with a sufficient sample size and a statistical chi-square test, from the data of the RCT (the frequency of the occurrence of O in the treatment group is greater than in the control group). However, as almost any premise whose foundation need not be continually tracked down, let us assume that the evidence under discussion is the RCT result given by a statistical test rather than the organised data from the RCT. The hypothesis of interest is that T causes O in population σ.

Let $K_i$ = a population that is causally homogeneous with regard to O except for T and its downstream effects.[39] In other words, $K_i$ represents a particular combination of confounding factors ($C_1$, $C_2$, … ,$C_n$). Call the study population σ, in which the $K_i$'s are sub-populations ($K_i \subseteq \sigma$).

**Argument of Efficacy:**

P1. For any $K_i$, $P(K_i/T\&\sigma) = P(K_i/{\sim}T\&\sigma)$. [by the balance assumption of an ideal RCT][40]

P2. $P(O/T\&K_i) > P(O/{\sim}T\&K_i)$ iff T causes O in $K_i$, and T causes O in some σ iff T causes O in some $K_i \subseteq \sigma$. [by the probabilistic theory of causality]

P3. E: the frequency of the occurrence of O in the treatment group (T&σ) is greater than in the control group (~T&σ) and the sample size is large enough. [the RCT result]

P4. $P(O/T\&\sigma) > P(O/{\sim}T\&\sigma)$. [from P3 and the RCT result shown statistically significant, say, by a chi-square test]

P5. For some $K_i \subseteq \sigma$, $P(O/T\&K_i) > P(O/{\sim}T\&K_i)$. [from P1 and P4][41]

---

[39] A causally homogeneous group is one whose variables have the same values for each individual in the group. Suppose the causally relevant set in the study population σ includes only two confounding factors $C_1$ and $C_2$. $K_i$ consists of an exhaustive set of subpopulations that are causally homogeneous with regard to O except for T and its downstream effects in σ. Each $K_i$ can be one of these: $\{C_1, C_2\}$, $\{C_1, {\sim}C_2\}$, $\{{\sim}C_1, C_2\}$, $\{{\sim}C_1, {\sim}C_2\}$.

[40] The balance assumption: $\forall K_i$ are distributed identically in the treatment and the control group. That is, '[i]n an ideal RCT each $K_i$ will appear in both [study] wings with the same probability' (Cartwright 2010, p. 64).

[41] Proof: 1. $P(O/T\&\sigma) > P(O/{\sim}T\&\sigma)$ P4
2. $P(K_i/T\&\sigma) = P(K_i/{\sim}T\&\sigma)$ P1

Con. H: T causes O in σ. [from P2 and P5]

As long as these assumptions are met, from an ideal RCT, the conclusion 'T causes O in the test population σ' can be deduced from the RCT result, which is labelled evidence relative to other premises and the hypothesis.

On their own, RCTs are not enough to draw causal conclusions. For RCTs to establish causal claims about the study population, a *clincher* argument is required, which calls for not only the positive results of RCTs but also some additional preconditions, to reach a causal conclusion. Cartwright (2007) distinguishes 'clincher' arguments from 'voucher' arguments as different kinds of warrants for (causal) claims. Clinchers have an argumentative structure wherein, as the argument theory requires, the premises deductively entail the causal claims. For example, Cartwright argues, along the lines I have just sketched above, that a positive result, i.e. $P(O/T\&\sigma) > P(O/\sim T\&\sigma)$, of an ideal RCT exemplifies a clincher for the causal claim 'T causes O in the study population'. The truth of the conclusion drawn from an ideal RCT is guaranteed given the fulfilment of the preconditions (P1, P2 and P3). According to Cartwright (2007, p. 14), the types of methods of inquiry that can *clinch* causal claims include 'econometric methods, Galilean experiments, probabilistic/Granger causality, derivation from established theory, tracing the causal process [and] ideal RCTs'. Vouchers, by contrast, speak in favour of claims without establishing them with deductive validity. The following methods typically establish evidence that *vouches* for causal claims: H-D method, 'qualitative comparative analysis, or looking for quantity and variety of evidence' (Cartwright, 2007, p. 12). Contrary to clinchers, vouchers often come with a broad scope of application but lack such deductive certainty.

---

3. $\sim\exists K_i[P(O/T\&K_i) > P(O/\sim T\&K_i)]$ AIP
4. $\forall K_i[P(O/T\&K_i) \leq P(O/\sim T\&K_i)]$ from 3
5. $\forall K_i[P(O/T\&K_i)\ P(K_i/T\&\sigma) \leq P(O/\sim T\&K_i)\ P(K_i/\sim T\&\sigma)]$ from 2&4
6. $\sum P(O/T\&K_i)\ P(K_i/T\&\sigma) \leq \sum P(O/\sim T\&K_i)\ P(K_i/\sim T\&\sigma)$ from 5
7. $\sum P(O/T\&K_i\&\sigma)\ P(K_i/T\&\sigma) \leq \sum P(O/\sim T\&K_i\&\sigma)\ P(K_i/\sim T\&\sigma)$ from 6
8. $P(O/T\&\sigma) \leq P(O/\sim T\&\sigma)$ from 7
9. $[P(O/T\&\sigma) > P(O/\sim T\&\sigma)]$ & $[P(O/T\&\sigma) \leq P(O/\sim T\&\sigma)]$ from 1&8
10. $\exists K_i[(P(O/T\&K_i) > \exists K_i P(O/\sim T\&K_i))$ from 3–9 IP

Lacking deductive certainty, do vouchers provide some lesser degree of evidential support? Cartwright (2007, p. 12) doubts this, and claims that '[t]hat is hard to say since the relation between evidence and conclusion in these cases is not deductive and there are no general good practicable "logics" of non-deductive confirmation, especially ones that make sense for the great variety of methods we use to provide warrant'. Clearly any voucher can be turned into a clincher by adding the additional premises needed to form a deductive argument. For instance, if an RCT fails to satisfy the preconditions for ideal studies, the RCT can at best vouch for a causal conclusion; it will be a clincher only if these ideal conditions are met. So Cartwright must more or less intend the distinction between clinchers and vouchers to be: methods generally come with a set of conditions that say what is for them to be carried out in the ideal. The hypotheses clinchers support can be deductively inferred from the results plus the assumption that the ideal conditions hold.

As noted, in order for an RCT to be a clincher, it must meet ideal conditions. We can rarely, if ever, expect that any particular RCT succeeds in providing grounds for claiming that the causal claim has been clinched. However, despite the result that P(O/T) > P(O/~T) arising from a well-conducted RCT, we are sometimes reluctant to acknowledge that the result has successfully established a causal claim. In the *British Medical Journal*, Leonard Leibovici (2001) reported an RCT that John Worrall (2007a) described as *impeccably randomised* and *double blind* and that corroborates that there is a correlation between remote, retroactive intercessory prayer and length of stay in hospital. However, as Worrall points out, it is questionable whether the causal conclusion elicited from such an RCT is admissible.

The case was well summarised by Worrall. As he describes it, the medical information about 3,393 inpatients who caught blood infections at the Rabin Medical Centre in Israel between 1990 and 1996 contains '(i) those patients' mortality; (ii) their length of stay in hospital; and (iii) the duration of the fevers they had suffered' (Worrall, 2007a, p. 1005). In 2000, a study was carried out

retrospectively to know whether a remote, retroactive prayer can have a positive effect on (i)–(iii). The inpatients were randomly assigned to two groups, with 1,691 to the treatment group and 1,702 to the control group. There was no significant baseline imbalance between the two groups, which had been checked with known 'main risk factors for death and severity of illness' (Worrall, 2007a, p. 1005). The study was double blind; the patients and the doctors treating them had no way of knowing to which group they had been assigned. Prayers were made on behalf of the treatment group, and while the result showed that there was no statistically significant difference in mortality rates between the two groups, their length of stay in hospital and duration of fever were significantly reduced, which implies that prayer has a miraculous power to change the past.[42]

Federica Russo and Jon Williamson (2007) suggest a strategy for blocking this type of specious inference: they use evidence of difference-making in a well-conducted RCT alongside evidence that there is a mechanism connecting cause and effect.[43] Apparently the RCT of the effects of retroactive prayer is an exception. The Russo-Williamson Thesis (RWT) is expressed as follows:

> In order to establish A is a cause of B in medicine one normally needs to establish two things. First, that A and B are suitably correlated—typically, that A and B are probabilistically dependent, conditional on B's other known causes. Second, that there is some underlying mechanism linking A and B that can account for the difference that A makes to B. (Clarke *et al.*, 2014, p. 343)

With no conceivable physical mechanism explaining the correlation in this case,[44] the claim that the miraculous power of prayer brought about a shorter

[42] There was a difference, though not statistically significant, between the treatment group and the control group in terms of morality, with 28.1% and 30.2% respectively (p-value = 0.4). By contrast, the treatment group saw a statistically significant drop in length of stay in hospital (p-value = 0.01) and the duration of their fevers (p-value = 0.04) (Leibovici, 2001).
[43] Clarke et al. (2014) claim that well-devised and well-conducted RCTs can, in principle, provide evidence of difference-making and evidence of a mechanism in tandem. From the perspective of the argument theory, every well-done RCT provides evidence of a mechanism only via the assumption that this RCT has established the causal relationship between T and O. By adding the assumption that there can be no causation without a connecting mechanism, we can draw the conclusion that there is a mechanism.
[44] Leibovici (2001, p. 1451) holds a similar view: 'No mechanism known today can account

length stay in hospital is physically implausible. But a religious person who believes that God is so omnipotent as to be able to change the past as well as the future might assert that there is in fact some mechanism at work. Even so, at least for those who grant that there is no conceivable mechanism, they will not believe that retroactive prayer is an effective remedy. Furthermore, even supposing the RCT, had it been ideal, had revealed that there must have been a reasonable mechanism at work, and supposing this experiment had approximately, as best it possibly can, attempted to satisfy the ideal conditions (e.g. double-blinding, randomised design, etc.), a causal claim would not follow deductively, because even our best efforts cannot assure the ideal conditions are met to form a sound argument. Indeed, imbalance is to be expected in any single run of the experiment.

### 3.3 Argument of Effectiveness of RCTs

The argument theory is well suited to expose another insufficiency of RCTs as the claims made on the basis of RCTs call their external validity into question. The argument theory sheds light on the restricted function of RCTs in the application of evidence to practical uses in such contexts as education schemes, policy deliberation and medical evaluation. The vast majority of theories of evidence hardly address the challenges posed by evidence for use. Before criticising the effectiveness of RCTs, Cartwright (2012, 2013) lays out the strategy whereby RCT advocates attempt to predict an outcome in the target setting wherein a policy or treatment were to be applied.

In cases where evidence for use is intended to generate plausible predictive claims supported by causal claims established by RCTs, a deductive relation, Cartwright (2012) maintains, is required between a sufficient amount of the right kinds of evidence and an intended claim about the future.

---

for the effects of remote, retroactive intercessory prayer said for a group of patients with a bloodstream infection'. Nonetheless, Leibovici concludes that now that remote, retroactive intercessory prayer has been shown to be effective, undertaking further studies to clarify its mechanism would be worthwhile.

Cartwright (2013, pp. 14–15) articulates a set of sufficient conditions for an argument that is a relevance-mediating vehicle for causal claims established to attain external validity: the *effectiveness argument*. Ideally, the outcomes of RCTs can be transported to the target setting by virtue of this argument.

In order to understand this argument, we must first know about *support factors*. The notion of support factors can be elaborated in terms of J. L. Mackie's (1965) INUS condition (Insufficient but Necessary parts of a condition which is itself Unnecessary but Sufficient). INUS conditions characterise causality in the relations in which, say, *x* is a cause of *y* in the sense that *x*, combined with support factor 1, support factor 2 and so forth, is sufficient to produce *y* and in this set every single factor including *x* itself is necessary. No proper subset of the factors is sufficient to bring about *y*. The cause *x* along with its support team is not a necessary condition for producing *y* as there are other causal sets that are sufficient to produce *y* as well.

To see how INUS operates, consider the ways one can get from New York to London. Air or maritime transport allows a passenger to reach her destination. If she chooses the air transport mode, she will have to walk and take land vehicles (e.g. train, bus, taxi, etc.) to get to an airport before arriving in London by plane. Having a flight booked is not enough on its own to take her to London: walking and taking land vehicles are all necessary to make her arrival in London happen. However, the air option is not an exclusive option. She can switch to the maritime transport, which is also sufficient to take her to the destination if enough of the other necessary conditions are met (e.g. walking, taking vehicles, swimming, etc.). Returning to our concern with RCTs, it is rare that treatment variables or policy variables (X) are sufficient on their own to bring about the outcome (Y). The realisation of the outcome calls for collaboration between these variables and their support team.

Although RCTs are widely deemed to be the gold standard of evidence for causal claims and to guarantee policy effectiveness, they cannot, without

support factors, guarantee that the outcome of an RCT in the study setting will work in a new target setting (Cartwright, 2012). The argument theory, as discussed below, makes explicit the problem that RCTs can establish only one premise in a valid argument that relies on many other premises.

Suppose now that there is a causal claim that T causes O in the study population σ (in Cartwright's (2013) terminology, *it works there*) and we wish to know whether the hypothesis is true that T will make a positive contribution to O for some individuals in the target population (*it will work here*).

**Argument of Effectiveness:**

P1. E: T causes O in the study population σ. [from the causal claim established in the RCT(s)]

P2. If T plays a causal role in the production of O in the study population, T can do so in the target population. [same causal role]

P3. The support factors necessary for T to make a positive contribution to O in the target population are present for some individuals in the target population.

Con. H: T will make a positive contribution to O for some individuals in the target population if T were to be implemented.

Those who believe in the effectiveness of RCTs speculate that in the target setting, the treatment will work in the same manner as in the RCTs and the outcome will not be very different from that of the test setting. In the target setting, they have assumed both the robustness of the causal power of the treatment (i.e. T can cause O in the new settings if it can in a study setting) and the recurrence of the support factors that make the treatment effective in the study setting. Formal results show that the transportability from one setting where causal relationships are identified to a target setting is feasible only under certain conditions (cf. Bareinboim and Pearl, 2013). For example, if the average treatment effect (ATE) is to be the same in a target as had been measured in an RCT on a different population, two tasks must be

accomplished: (1) if the cause plays a causal role in the original study setting, it can do so in the target setting, and (2) the expectation (average) of the support factors must be the same in both. Just as a valid, sound argument is required when licensing the ATE from the RCT, a valid argument and the truth of the assumptions (1) and (2) are required to serve as evidence for the same ATE in the target (Cartwright, 2012).

Cartwright's argument theory calls into question the strategy by which advocates argue for RCTs as a basis for establishing external validity. Real RCTs are likely to suffer from various sources of known and unknown errors that undermine their efficacy. Cartwright (2013) holds that even if ideal RCTs are available, the outcomes will not guarantee that decisions based on them will bring about the expected results elsewhere, what she refers to as the 'from-there-to-here problem'. I reformulate this problem as follows:

> **Same causal principle shared by treatments:** treatments in the study population and the target population are governed by the same causal principles. This ensures that if T plays a causal role in the study, it can do so in the target.
>
> **Same support team:** treatment does not result in the same effect size unless the same expected value for the support factors is obtained in both populations.

As Cartwright (2012) points out, causal principles, for the most part, are local and fragile, in part due to the way causal principles work, depending on the coordination of the components of the underlying causal structure. If the arrangement of the causal structure is altered even slightly, the causal principle may not apply. If a causal principle is observed to hold in one setting, it does not imply that it will operate in a new setting, beyond locality.

The case of the Family Nurse Partnership (FNP) in the UK illustrates the failure of the external validity of RCTs. The FNP programme was developed in the USA almost 40 years ago, it was imported to and adapted in the UK in

2006, and it was carried out in the UK from 2009, with the aim of 'affect[ing] risks and protective factors within prenatal health-related behaviours, sensitive and competent care-giving, and early parental lifecourse' (Robling *et al.*, 2016, p.148). In the US, several RCTs showed considerable benefits of FNP for 'improv[ing] birth outcomes, cognitive and socioeconomic development, use of preventive health care, and reduc[ing] potential abuse (eg, injuries, ingestions, and emergency department attendances) and maltreatment' (Robling *et al.*, 2016, p.147). However, the Building Blocks RCT, led by Dr. Michael Robling (2016), which aimed to assess the effectiveness of FNP in the UK, showed that in the short term, the FNP programme was not as effective in the UK as it was in the US.

Beginning in 2009, the Building Blocks RCT enrolled and tracked over 1,600 participants (mothers-to-be aged 19 or younger across England) until their child reached the age of two. These mothers-to-be were randomly assigned to the FNP group involving up to 64 structured home visits, with approximately half of the mothers receiving support from FNP as well as 'usual' care from health and social services and the other half only receiving 'usual' care. The results showed no evidence for any 'additional short-term benefit for our selected primary outcomes (smoking in pregnancy, birthweight, emergency hospital attendance and admission for the child, and subsequent pregnancy)' (Robling *et al.*, 2016, p.147).[45] On the contrary, the study reports that there was 'an incremental cost for FNP of £1,993 per participant' (Robling *et al.*, 2016, p.151). In the end, the study suggests that a follow-up evaluation of the longer-term effectiveness of FNP was needed.

After analysing the data, the authors of the report suggested that the difference in the short-term effects of FNP between the US and the UK might

[45] However, the research revealed some small positive secondary outcomes including 'intention-to-breastfeed, maternally reported child cognitive development (at 24 months only), language development using a modified maternal-reported assessment (at 12 and 18 months) and using a standardised assessment (the Early Language Milestone; at 24 months), levels of social support, partner-relationship quality, and general self-efficacy' (Robling *et al.*, 2016, p.150) and showed that '[r]ates of child safeguarding concerns documented in primary care records were higher for FNP clients. There were no other differences found….' (Robling *et al.*, 2016, pp. 150–151).

have resulted from several factors, such as differences in the enrolment criteria and the extent of the care provided (Robling *et al.*, 2016). Compared to the situation in England, women with relatively more risk factors were enrolled in the FNP programme in the US, which might have led to comparatively fewer disadvantages as well as more heterogeneity (Robling *et al.*, 2016, p.152). Additionally, supportive health and social services, including 'community based family doctors, midwives, and public-health nurses, and, in most trial sites, specialist teenage pregnancy midwives' were more accessible for young mothers in the UK than in the US (Robling *et al.*, 2016, p.152). So the effect of the FNP intervention may have been diluted by these commonly provided resources in the UK. It is evident that the positive effects of FNP trials in the US did not validate the effects of FNP in the UK. The same support factors such as similar enrolment conditions and health and social services should line up with the FNP programme in the UK so that the outcome of the target setting would appear to work in a fashion similar to the original study setting.

Advocates of RCTs would presumably object that the increasing number of RCTs that broaden the application scope of a causal relationship established from them can enhance the reliability of a prediction. However, Cartwright (2013, p. 16) criticises this strategy on the basis that a prediction based on enumerative induction is not a kind of robust inference. Overall, as Cartwright points out, enumerative induction relies on whether the property of interest spans all inductive objects. For example, a great number of white swans observed from many different places in the UK would entitle us to be justified in believing that all swans are white under induction by simple enumeration. Similarly, advocates of RCTs can argue that a great deal of samples of a causal claim 'T makes O happen' established from RCTs can entitle us to claim that T makes O happen everywhere, whereby they further predict that T will make O happen in the target setting. The further inferential move relies on whether the property of interest is shared by the inductive base and the target hypothesis. According to Cartwright (2013, p. 16), while electron charge is generalisable and bird colour sometimes is, causal properties

require more additional premises to lead us from ‘it works there’ to ‘it will work here’.

Cartwright characterises the relation between evidence and hypothesis as that between true premises and the conclusion they entail in a deductive argument. The structure of the arguments is analogous to a pyramid, in which hypothesis A at the top layer of an argument pyramid is the conclusion supported by its premises on the lower layers ($A_1'$, $A_2'$, $A_3'$ …), and each of these premises by further arguments (Cartwright and Hardie, 2012, pp. 16–20). It can be seen that these lower-layer premises ($A_1'$, $A_2'$, $A_3'$ …) should be self-evident (e.g. logical or mathematical truth) or well-grounded (e.g. sensory experience in normal circumstances); if one of them is not (say, $A_1'$), then A1′ also needs a valid, sound argument to support it. The premises ($A_{11}''$, $A_{12}''$, $A_{13}''$ …) of this more bottom-down argument should be shown to be true. The process continues until the premises are self-evident or well-grounded, or we are just willing to take them for granted.

### 3.4 Limitations, Difficulties and Strengths

In what follows, I point out the major limitations of the argument theory and its inherent difficulties.[46] As regards its limitations, Cartwright (2014) makes a fundamental distinction between evidence and justification: a theory of evidence is concerned with the conditions under which the data lend support to a hypothesis, whereas a theory of knowledge involves the conditions under which we are justified in believing claims. The argument theory just tells us that if you have such-and-such premises that form a sound and valid argument, each of the premises is labelled evidence; what it is not responsible for is when we are justified in believing that they are true. The argument is objectively there on the metaphysical level; epistemically, the

[46] Reiss (2015, pp. 48-52) has argued that the argument theory is unsatisfactory in that it does not satisfy all of his four desiderata for a theory of evidence. As I have outlined in Chapter 1, I endorse a meta-thesis of pluralism: a theory of evidence should be evaluated against the goal it sets out for itself. That means that even if the argument theory fails to meet some of Reiss’s desiderata, it can still be a ‘partially’ satisfactory, contrary to ‘fully’ satisfactory in Reiss’s sense, theory if accomplishing its goal.

truth of the premises relies on the judgement of practising scientists. In actual practice, the requirement of true premises in the argument theory is difficult to meet: the premises setting out empirical claims are themselves fallible and can rarely be *proved* true. If advocates of the argument theory wish to bestow practical value on the theory, they could maintain that the truth of the premises ($A_1'$, $A_2'$, $A_3'$ …) can be guaranteed by the bottom-layer premises ($A_{11}''$, $A_{12}''$, $A_{13}''$ …), which constitute a valid argument for them. As I have mentioned, the truth of premises is guaranteed if they are self-evident or well-grounded. It is rare, if ever, that self-evident claims that constitute a valid argument on their own, are regarded as evidence. Moreover, it is controversial whether the truth of so-called well-grounded claims is verified by the reliable mechanisms (e.g. sensory experience) as their foundation. Lack of self-evident or well-grounded claims as premises can lead to an infinite regress. The absence of guidance on knowing that the premises of an argument are true is one of the limitations of the argument theory.

A solution could lie in a different interpretation of what constitutes well-grounded claims. Foundationalism should be replaced with coherentism, whereby the truth of the premises is ensured by the temporarily assumed truth of premises which could otherwise be confirmed by other arguments possibly drawn from different disciplines. John Norton (2014, pp. 686–688) refers to this reciprocal support as *scaffolding,* likening it to the temporary supports that are used when a variety of types of stone buildings such as domes, arches and cathedrals are being built and that are removed when the buildings are constructed in such a way as to be self-sustaining thereafter. In the same vein, the premises of an argument acquire material warrants that are domain-specific but may be drawn from different disciplines.

Another limitation is that the argument theory hardly explains the principles behind how facts are adduced as evidence in the historical sciences, particularly regarding historical linguistics. I will visit this issue in Chapter 5. Now let us turn to the difficulties with the argument theory. They arise from its overly liberal definition of evidence: it requires merely logical relations between claims to identify alleged evidential relations, without requiring that

the premises evidentially transmit their warrant to the conclusion.[47] Specifically, it allows any true claim to be evidence for itself (e.g. A is evidence for A), for a conjunction to give evidence for its operand (e.g. A&B is evidence for A), for the claim obtained by *modus ponens* (A→B & A is evidence for B), or for a logical truth (e.g. A is evidence for B v ~B). However, I will argue that these kinds of evidence alleged by the argument theory do not seem to be evidence nor are they evidence in many of the circumstances mentioned above. One reason against their qualification for being evidence could be that they have no bearing on causality.[48] Yet it is debatable whether evidential warrant embodied in the premises must be causal or can be something else; this is beyond the scope of this study. Another reason I will rely on instead is the failure to transmit evidential warrant by purely logical relations.

Cartwright (2013) defends the alleged evidential relation between any arbitrary fact and a logical truth, in which, in her account, the fact will count as evidence:

> Still I wouldn't advise spending much to buy information about other facts to warrant a logical truth. If you know a claim is a logical truth you don't need to buy information about other facts to warrant the claim. And if you don't know that the claim is a logical truth, you will have trouble warranting that the claim is implied by the fact you buy. Still, if I don't know h is a logical truth but I am assured that if e then h, then e is surely worth learning. (Cartwright, 2013, p. 7)

Cartwright (2013, p. 8) also holds that A&B is certainly evidence for A. Along with the above quoted reasoning, she might argue: if you do not know A with the intention of knowing A, it may be worth an investment in information about A&B, since A&B can of course provide information about A.

---

[47] It can be argued that the argument theory also defines evidence too narrowly. Logical relations do not cover all types of evidential relations. However, in the tool-based view I endorse, if the argument theory is modified to provide a sufficient, rather than necessary, condition for evidence, it does not need to cover other types of evidential relations.

[48] Cartwright makes a distinction between evidence and explanation; the latter may carry causal connotations while the former does not. Cartwright (2013, p. 8) argues that a man's taking birth control pills is evidence that he does not get pregnant and that those who do not regard it as evidence confuse 'the task of providing that a fact obtains with the task of explaining why it obtains'.

Conversely, if you have known A, it is not worth buying information about A&B to know A; it is not advisable to waste money. In short, for Cartwright, the fact that A&B is evidence for A is not generally useful for us because we generally have to learn A and B individually in order to learn A&B.

Yet I argue that even if non-causal relevance relations could support evidential claims, purely logical relations are, *in some cases*, not genuine evidential relations. These cases occur when we have already known the conclusion (e.g. A) before or at the same time as we have information that is ready to infer the conclusion (e.g. A&B). In such cases, the so-called evidence does not work for us. Suppose one hundred people who jumped from the top of the Shard, the tallest building in London, died. Would Mr. X die if he did the same thing? Yes, it can be inferred that he is very likely to die as one hundred people have taught the lesson with their lives. But if he was already included in the 100-person sample, it seems inappropriate to assert that one hundred deaths are evidence that Mr. X would die if he jumped from the top of the Shard. The difference in the availability of evidence between the two scenarios stems from whether or not information about the conclusion supports the premise.

Clearly, in cases where A is not information available preceding evidential inference or where A&B is warranted by theories other than the occurrence of A, A&B can be evidence for A. For instance, the information about Mr. X being in the 100-person death list is not accessed beforehand, so it is evidence for his death. Additionally, according to special relativity, all light propagates with speed no more than $c = 3.00\times10^8$ m/s. This information about the maximum speed of light is evidence for a claim that the speed of this light propagating in a vacuum is no more than $3.00\times10^8$ m/s, even if this light is not actually measured in terms of its speed. In other cases, as noted earlier, where A is already known before or at the same time as A&B is known, A&B is not evidence for A. Likewise, A is not evidence for itself insofar as it is not possible that A remains unknown before A is known. Arguably, contrary to the argument theory, A&B is not always evidence for A, and A is not evidence for itself.

Evidence should allow us to infer from observation to unobservable or observable but yet-to-be-observed entities, events or regularities. It is not the case that evidence is used to infer a claim that is already confirmed by observation from other claims about observed events or phenomena. We do not need evidence to establish what is already known. Transmission from the observed to the observed may involve many kinds of epistemic activities other than evidential inference; explanation is one of them. Why did the massive earthquake devastate much of the area? Suppose the hypothesis that the area was destroyed is known to be true. Although there are many plausible explanations for this destruction (e.g. soil liquefaction), there is no need to gather evidence for the truth of the occurrence of the destruction unless we need evidence to know whether the soil liquefaction did exacerbate the destruction. Hence, for purely logical relations, say, A&B deduces A, now that A&B is observed, A has been observed, and thus A&B is not evidence for A, though the argument theory should admit A&B is. The same applies to the case that A is not its own evidence.

There is another situation that should be considered, that is, whether (A→B) & A is evidence for B. Here I aim to find a relatively general strategy to object to taking evidential relationships as purely logical relations. To put it in a broad context, logical relations, in combination with true premises, sometimes fail to transmit evidential warrant. In what follows let us consider Michael McKinsey's (1991) argument for the incompatibility of first-person authority and externalism about mental content to illustrate this concern:

(P1) I believe that water is wet. [from privileged first-person authority][49]

[49] First-person authority means that 'we have a distinctively first-personal and specially authoritative way of knowing that we ourselves have that [mental] property, when we do have it, without needing to conduct any detailed empirical investigation either of the environment and our relation to it or of our internal cognitive architecture' (Davies, 1998, p. 322). First-person authority allows only sincere claims. 'On any account, first-person authority holds only for avowals that are sincere, in the sense of not deceitful, and if we allow that non-conscious deceit is possible, then we must allow that some avowals may be

(P2) If I believe that water is wet then I (belong to a community of speakers some of whom) have had contact with water. [from strong mental content externalism]

(Con) Therefore, I (or some members of the speech community) have had contact with water.

Externalism about mental content holds that 'whether a person (or other physical being) has that property depends, not only on conditions inside the person's skin, but also on the person's environment and the way that the person is embedded in that environment' (Davies, 1998, p. 322). McKinsey (1991) argues that for externalists, first-person authority (P1) and the externalist thesis (P2) can be known or warranted *a priori*, i.e. I do not need to look at the external world to know or warrantedly believe both premises. If (P1) and (P2) are known or warranted *a priori*, these jointly entail that (Con) is known or warranted *a priori*. In other words, since knowledge of (P1) and of (P2) is non-empirical, knowledge of (Con) follows in a non-empirical manner. This means that on externalism, we can know something about the external world without looking at it. To say that (Con) is knowable *a priori* would be a preposterous conclusion that externalists would be unwilling to accept, since it implies that even without empirical investigation we can know that we live in a water world. McKinsey therefore puts forward the incompatibility thesis: either externalism is untenable, or we have no privileged access to our mental content.

Martin Davies's (1998) and Crispin Wright's (2000) novel solution to this incompatibility problem is to expose *warrant transmission failure*.[50] This is not suggesting that a valid argument's conclusion is not warranted when its premises are warranted. Evidential warrant cannot be transmitted from the true premises of a valid argument to its conclusion if the truth of the conclusion is a precondition for having evidence for one of the premises. In

insincere, and hence non-authoritative, even though no conscious deception is intended' (Frankish, 2004, p. 224).

[50] What makes their solutions novel is that the preceding solutions attempted to refute (P2). See Brueckner (1992) for traditional solutions.

other words, the premises of a valid, even sound argument fail to warrant the conclusion, if on the supposition that the conclusion were false, the evidence warranting one of the premises would not be evidence. This situation is viewed as warrant transmission failure because the premises do not provide any new information about the truth of the conclusion.

As Wright points out that, while not incompatible, a *cogent* argument and a *valid* argument are related but distinct. A cogent argument transmits the warrant of its premises to the truth of its conclusion, while a valid argument may fail to do so 'where there was warrant for the premises in the first place because the conclusion was *antecedently* warranted' (Wright, 2000, p. 141). A valid argument obeys the principle of 'closure of warrant'; a cogent argument obeys the principle of 'transmission of warrant' (Wright, 2000, p. 140). The principle of closure of warrant is weak in that it ensures only that whenever the premises possess a warrant, the conclusion also possesses a warrant; however, the warrant for the conclusion might have been obtained elsewhere. The principle of transmission of a warrant ensures that in addition to the conclusion in possession of a certain warrant, the warrant for the conclusion comes from the warrant of the premises.

Wright's (2000) remarks are worth quoting at length:

> Transmission of warrant need not be an absolute characteristic of a valid argument. It may be that a particular argument is such that one type of possible ground, $w_1$, for its premises is transmissible—can yield a novel reason for accepting the conclusion when taken in conjunction with recognition of the validity of the inference—while another, $w_2$, is not, but can only be possessed in the first place by a thinker whose information already includes warrant to accept the conclusion.
>
> Intuitively, a transmissible warrant should make for the possible advancement of knowledge, or warranted belief. A warrant is transmissible, more specifically, when we may envisage a logically non-omniscient but otherwise perfectly rational subject coming to believe a proposition for the first time in a way which depends on their recognition *both* of the validity of the inference in question and of their possession of warrant for its premises. (Wright, 2000, p. 141, italics in original)

Returning to McKinsey's incompatibility argument, evidence that supports

(P1) is my introspective experience, which presupposes (Con) that we have had contact with water if we are externalists, who believe that introspective experience rests on our interaction with water. It becomes apparent that (P1) and (P2) fail to transmit epistemic warrant to (Con) insofar as knowledge about the water environment is, on strong externalism about the mental content expressed in (P2), the precondition of self-knowledge that I believe that water is wet. Valid deduction does not necessarily warrant evidential transmission success.

Now let us apply this conclusion to the argument theory. If we are to use E as evidence for H relative to argument A, we must at least have reasons for E, and these reasons cannot presuppose H. A valid and sound argument can still be utilised as a relevance-mediating vehicle as long as it avoids non-evidential-warrant-transmission claims as its premises.

When does deductive entailment successfully transmit warrant to the conclusion? Consider that the claim 'if it rains the ground gets wet' and the claim 'it rains' together deductively entail the claim 'the ground gets wet'. The premises do not necessarily rely on the fact that the ground gets wet because the fact that it rains can be observed by other routes, such as infrared images of tall, cold clouds that indicate rainfall, or seeing raindrops falling from the sky.[51] So if the premises gain warrant from elsewhere, the truth of the claim 'the ground gets wet' is evidentially well-supported. Along this line, some trivial logical relations (e.g. any true claim is evidence for itself or for a logical truth or A&B is evidence for A), legitimate as evidential relations in the argument theory, cannot bestow an empirical warrant of the premises upon the conclusion.

Although it is not obvious whether A transmits warrant for the logical truth Bv~B and thus whether A is evidence for Bv~B,[52] the cases mentioned

[51] The conditionals, such as 'if raindrops fall from the sky, then the ground gets wet', are obtained partly from associations and partly from scientific knowledge. Such conditionals are, though fallible, reliable.
[52] Davies (1998) acknowledges that it is possible that not all necessary conditions count as preconditions, which unblocks the possibility of knowledge by inference.

above should suffice to draw a conclusion that not all valid, sound arguments function as moulding evidential relations. Perhaps Cartwright could maintain that the argument theory aims only to characterise what evidence and evidential relationships there are at the metaphysical level: what facts speak for (or give rise to, if understood causally) a fact or what premises ensure a conclusion. She could assert that the warrant transmission failure illustrated above is irrelevant; after all, 'warrant' is an epistemic notion. The epistemic claim that we are warranted in believing X on the basis of Y is not the same as the metaphysical claim that X is evidence for Y. Since the two kinds of claims are distinct, it can be reasonable to make different demands of them. This reply could deflect the criticism that *epistemic* warrant should be transmitted from premises to the conclusion, since the argument theory is not suited to tackle epistemological issues. This echoes the view that a theory of evidence should be assessed on a purpose-specific basis, as indicated in Chapter 1.

I have two comments in response to Cartwright's possible reply. Cartwright could argue that claims about evidential relevance must be true, though the truth of these relevance claims is not necessarily epistemically attainable. If claims about evidential relevance did not have empirical grounds, the reasoning would be called "proof", (typically seen in logic and mathematics) the credibility of which derives from axioms and theorems and requires no empirical content of the premises.[53] Furthermore, one of the virtues of the argument theory is the requirement that evidential arguments be sound, which provides researchers with guidelines for ensuring that each premise

[53] Achinstein (2014, p. 390, italics in original) also holds that 'entailment would be *proof* not evidence', which accounts for the reason for the third condition of his theory of evidence, namely (c) E does not entail H. Lipton (2004, p. 5) holds a similar view that '[i]nductive inference is thus a matter of weighing evidence and judging probability, not of proof'. Here, two situations should be carefully distinguished. For logical and mathematical proof, premises are true in accordance with axioms but not with empirical investigation, and thus do not carry evidential meaning. However, once empirical content is bestowed upon logical or mathematical premises, such premises may function as evidence. A logical example is the 'if it rains the ground gets wet' example, as noted above. A mathematical example can be illustrated with a birthday example. Suppose there are 367 people in the room (E), we can guarantee that there must be a pair having the same birthday (H) on the basis that there are 366 days (including the 29th of February) in a year. In this case, E is evidence for H, though E entails H.

they use has support from its bottom-layer sound argument and that there is no inferential gap between premises and conclusion. This indicates an epistemic attitude: an expectation that a theory of evidence should work for us. If Cartwright wants the argument theory to have this virtue, she has to solve the difficulty of epistemic warrant transmission. Otherwise, the argument theory would offer at most a Platonic ideal of evidence and evidential relationships. This virtue of the argument theory, i.e. it makes clear what we need to know, is not much of a virtue if an argument demands us to know what we have scant access to.

By requiring deductive entailment, the argument theory admits trivial logical relations that transmit no substantial evidential warrant to a hypothesis. Nevertheless, in spite of this difficulty, the argument theory captures the intuition that a piece of evidence will not count as evidence if it is found to be irrelevant to the hypothesis. Relevance, according to the argument theory, hinges on the truth of the premises and the formal relation of deductive entailment. To illustrate, consider an example where the suspect's fingerprint on the duct tape, combined with support factors such as his motive for the murder and his physical presence at the crime scene, is believed to be strong evidence that he committed the murder. But his fingerprint on the duct tape no longer counts as evidence that he murdered the victim if one premise of the original argument for the hypothesis that he is the murderer is found to be false (e.g. the testimony to his presence at the scene turns out to be false) or the argument is found to be invalid (e.g. he was set up to leave his fingerprint on the duct tape). Evidential relevance may not persist if the original argument is found to be invalid or unsound subsequently.

It is worth remarking that the argument theory of evidence is concerned with metaphysical issues, involving *facts* that ensure the truth of a claim in the material mode and *reasons* that ensure the entitlement to believe in in the formal mode. In the next chapter, I will consider Reiss's inferentialist account of evidence that, conversely, emphasises the role of *judgements* in inferences from evidence to hypothesis at the epistemological level.

## 3.5 Conclusion

I conclude this chapter by recapitulating my proposal In Chapter 1. I have recommended recognising the value of partial, namely purpose-specific, theories of evidence; they solve certain philosophical and practical problems about evidence and should be assessed with respect to their scope and purpose. I have distinguished three distinct evidential relationships, namely objective support, subjective acceptance and guidance, and have proposed that these replace the desiderata for any theory of evidence. Desiderata require theories to conform wholly, while the identification of evidential relationships maps where a theory of evidence is located for the sake of further evaluation. In this chapter, I scrutinise the argument theory of evidence in accordance with this proposal.

Cartwright (2014) describes the aim of the argument theory of evidence with the following remarks:

> To figure out whether *e* is evidence for *h*, the Argument Theory guides you to look for good arguments connecting *e* and *h*. Of course it doesn't tell you how to tell if an argument is good. But that's not in its job description. Coming up with an argument is part of the ordinary normal science job of scientific discovery. To check that it is valid, perhaps one needs a good logician or a good mathematician. To tell if the premises are true, we employ the normal methods available in the paradigm in which we work for assessing the kinds of claims the premises make. (Cartwright, 2014, p. 111)

The above quote reflects one of the major limitations of the argument theory: looking for a valid argument and checking the truth of the premises are the tasks for which scientists, and not philosophers, are responsible. I have examined the argument theory in this spirit, arguing that its difficulties, including its inability to ensure the transmission of epistemic warrant and to exclude metaphysical evidential relationships irrelevant to our purpose, arise from a radical demand for deductive entailment.

However, the argument theory has several strengths. It can remind us of inferential gaps, such as when we attempt to use the results of RCTs to extrapolate to other settings. It can also explain why something previously accepted as evidence may no longer serve as evidence when some other premise has been shown to be flawed or the argument is invalid.

# 4

# Pragmatist Approach: Reiss's Inferentialist Account of Evidence

Forty years ago, Clark Glymour (1980a) famously ruled hypothetico-deductivism (H-D) 'hopeless'. Paradoxically, a host of contemporary biomedical and social scientists regard the hypothetico-deductive (H-D) model of evidential reasoning as a legitimate methodology in science.[54] H-D may not be utterly futile if additional criteria are supplemented. Philosophers endeavour to identify epistemic virtues, including predictability (e.g. Lakatos, 1976; Hitchcock and Sober, 2004), simplicity (e.g. Quine and Ullian, 1978; Kelly, 2007) and explanatory power (e.g. Lipton, 2004). Whether these attempts, particularly those concerning epistemic virtues, are successful or not is controversial.

This chapter focuses on the 'add-on' strategy of eliminativism (e.g. Bird, 2010a), which avoids the debate between realists and anti-realists, in which epistemic virtues are perceived as pragmatic, and truth-irrelevant, virtues by anti-realists. Eliminativism can be regarded as having been integrated with H-D into Deborah Mayo's (1996) error-statistical account of evidence. As discussed in Chapter 2, it concerns the reliability of test procedures in terms of controlling errors by computing specific numerical values concerning data given particular hypotheses. The error-statistical account is injected with an H-D-like concept, namely *fit*, and reflects eliminativism in the concept of *severity*. However, this account eschews prior probabilities, which can sometimes be profitably employed in medical sciences or otherwise appropriate settings. For example, the incidence rate (base rate) of a disease functioning as priors, in practice, can contribute to evidential information.

Contrasting with Mayo's quantitative approach, in this chapter I consider a new version of the H-D theory of evidence, which yields categorical yes/no

[54] For details, see Reiss (2015b, pp. 60–61).

answers to hypotheses without assigning numerical values to them. This new version substitutes the deductive component with the *expectation/loose implication* component combined with eliminativism, that is, Julian Reiss's (2015a, 2015b) inferentialist account of evidence, which sets out its mission as fitting well with the evidence descriptions of biomedical and social science practices.

I begin by laying out three critical constituents of Reiss's inferentialist account: *supporting evidence*, comprising direct and indirect evidence; *warranting evidence*; and *context*, ones that demonstrate its mediating vehicles for evidential relevance to be *expectation/loose implication* and *elimination of possibilities*. Subsequently, I will show that direct evidence provides a framework for us to understand how claims or objects are thought relevant to evidence. This concerns *guidance* (GD), one of the evidential relationships I outlined in Chapter 1. This theory also has a component related to what I have called *subjective acceptance* (SA), viz. warranting evidence. To complement the framework, I also formulate the notions of inferential consistency and inferential irrelevance. Finally, in terms of strengths, difficulties and limitations of purpose specificity from the perspective of the tool-based view I endorse, I will argue: (1) that the strength of the inferentialist account is explaining why randomised controlled trials (RCTs) are deemed to be the gold standard of evidence and why a variety of sources of evidence are deemed legitimate in the biomedical and social sciences;[55] and (2) that its difficulties and limitations stem from the fact that supporting and warranting evidence may be too coarse-grained to accommodate quantitative evidential reasoning in the scientific sphere.

## 4.1 Two Paradigms: Experimental vs. Pragmatist

Following the epidemiologist Mark Parascandola's (2004) distinction between two paradigms of reasoning from evidence in the biomedical and social sciences, Reiss (2015a, 2015b) aims to articulate both the *experimentalist*

[55] For a more detailed exposition of RCTs, see Chapter 3.

and the *pragmatist* paradigm (especially the latter). According to Reiss, the experimentalist paradigm takes RCTs as the gold standard of evidence, and it posits that outcomes produced by other methods should be appraised with reference to how much resemblance these methods bear to the gold standard. However, under the pragmatist paradigm (a relatively vague concept in Parascandola's paper), a hypothesis, according to pragmatic criteria, is supported by multiple sources of evidence,[56] not limited to RCTs. Reiss puts a great deal of effort into clarifying the so-called pragmatic criteria. In the belief that the pragmatist paradigm spans diverse and manifold other disciplines,[57] Reiss contends that this paradigm does better justice than the experimentalist paradigm to the complexity of biomedical practice and social research, especially regarding the ways that researchers gather evidence and reason from the evidence to establish generic causal claims (i.e. causal relations between even-types such as 'smoking causes lung cancer'). In order to spell out the pragmatist paradigm, Reiss sets forth a new theory of evidence that he calls the *inferentialist account of evidence*.

Reiss points out that in the biomedical and social sciences, RCTs implement the experimentalist paradigm by virtue of being widely deemed the gold standard of evidence 'in all the domains labelled "evidence-based", which

[56] This is a subtle literal difference between my usage and Reiss's, which is essential to highlight. I use 'multiple *sources* of evidence', instead of Reiss's (2015a, p. 341) 'diverse *bodies* of evidence' in the quote 'scientific claims are inferred, using pragmatic criteria, from diverse bodies of evidence that may but need not include experiments'. As stated in Chapter 1, I endorse a view that there is no such thing, unless labelled via relevance-mediating vehicles, intrinsically being evidence. In other words, on this view, RCTs are not necessarily evidence, unless they are already *for* a certain hypothesis *under* certain conditions. However, the phrase 'diverse bodies of evidence' seems to indicate the opposite: RCTs are already there as a body of evidence awaiting us to use pragmatic criteria to infer hypotheses. I believe that my rewording is consistent with Reiss's notions of evidence and will not affect the points I wish to make.

[57] The arena for discussing the pragmatist paradigm is set in the biomedical and social sciences. Reiss, nonetheless, believes that this paradigm can be extended to other disciplines, as can be seen from the following quotes: 'I will focus on scientific domains where randomized experiments can be and are frequently employed. This includes the domains mentioned above but excludes all those domains where controlled experiments are effectively epistemic engines, such as large parts of physics and chemistry and basic/in vitro research in the biomedical sciences. I shall also exclude historical sciences such as cosmology, astronomy, astrophysics, geology, palaeontology, and archaeology. I do believe that the proposed account can be extended, but I will leave the extension to future work' (Reiss, 2015a, p. 342) and 'I do think that the basic idea of the account given here applies mutatis mutandis to other domains and other types of hypothesis, but here I will only be concerned with the biomedical and social sciences (Reiss, 2015b, p. 59).

include parts of medicine, dentistry, nursing, psychology, education, social policy, and criminal justice, but also parts of economics' (Reiss 2015a, p. 341). The methods deployed under the experimentalist paradigm are therefore applications of Mill's methods. RCTs, or possibly other methods conforming to such a paradigm, operate reliably when several shielding conditions are met, such as the blocking or controlling of confounding factors to ensure that a variation in outcome stems solely from the adjustment of interventions or treatments. From the viewpoint of the experimentalists, these methods are 'intrinsically reliable and therefore epistemically basic' (Reiss 2015a, p. 358) and are thus the benchmarks against which the evidential status of other methods can be measured. Naturally, experimentalists have to address the question: to what extent do other methods used to establish evidential claims resemble these orthodox methods?

Natural experiments can acquire a status of 'credible' by mimicking these orthodox methods. Natural experiments are described as being judged as the 'second best' method in the RCT-dominated paradigm (Reiss, 2015c, p. 374) and they resemble RCTs in some essential respects, which can be illustrated in terms of the three hallmarks of an RCT:

> (1) The response of experimental subjects assigned to receive a treatment is compared to the response of subjects assigned to a control group.
> (2) The assignment of subjects to treatment and control groups is done at random, through a randomizing device such as a coin flip.
> (3) The manipulation of the treatment—also known as the intervention—is under the control of an experimental researcher. (Dunning, 2012, p. 15)

Natural experiments are a special form of observational study. Both typically bear the first hallmark: conventional observational studies compare 'outcomes for units bearing different values of independent variables (or "treatment conditions")' (Dunning, 2012, p. 16), whilst natural experiments compare results from treatment and control groups. The second hallmark nevertheless differentiates natural experiments from conventional observational studies. Observational studies lack the second hallmark: the

treatments are assigned neither randomly nor experimentally intentionally. By contrast, natural experiments, in which confounding factors may not be controlled or balanced to the same extent as they are in RCTs, are still capable of providing compelling evidence for hypotheses of interest by (approximately) randomly assigning subjects to treatment and control groups, and thus bear the second hallmark.

However, as opposed to RCTs, natural experiments lack the third hallmark; they obtain their data not from the intervention, but from '"naturally" occurring phenomena—actually, in the social sciences, from phenomena that are often the product of social and political forces' (Dunning, 2012, p. 16). Examples of natural experiments include Galiani and Schargrodsky's' (2010) Argentinian land-titling study, Angrist and Evans' (1998) study on the effect of family size on labour supply and Angrist's (1990a, 1990b) study on the effects of military service on later income. Although the third hallmark is absent in natural experiments, they mimic RCTs in terms of the first and second hallmarks, which enables their reliability to be accounted for by the experimentalist paradigm.

According to Reiss (2015a), despite its rigorous demand for evidence quality, the experimentalist paradigm is uninformative about why evidence of various kinds, other than those pieces obtained by RCTs or natural experiments, is acknowledged as legitimate in scientific practice. Under this paradigm, it remains unclear in what ways certain observational studies other than natural experiments count as evidence for hypotheses, in particular the studies which seem to be far from these Mill's methods-based studies. For instance, formal consensus is evidentially legitimate, albeit not to a high degree of hierarchy of evidence, in evidence-based medicine (Clarke *et al.*, 2014). Contrarily, within the experimentalist paradigm, formal consensus, remarkably dissimilar from the gold standard, is difficult to view as evidence.

Even if the experimentalist paradigm could verify the legitimacy of various kinds of evidence, it hardly makes sense that they are mostly considered subsidiary or inferior, particularly in the medical sciences. Reiss (2015a, p.

359), for example, casts doubt on the paradigm's ability to cohere with such a hierarchy, even by virtue of how much resemblance a kind of evidence bears to the gold standard of evidence. For illustration, consider a version of hierarchy of evidence recommended by the US Preventive Services Task Force (USPSTF) (2017).[58] In this evaluation system of the quality of evidence in medicine, from highest to lowest quality, the types of evidence are enumerated as:

I. Properly powered and conducted RCT; well-conducted systematic review or meta-analysis of homogeneous RCTs
II-1. Well-designed controlled trial without randomization
II-2. Well-designed cohort or case-control analysis study
II-3. Multiple time-series, with or without the intervention; results from uncontrolled studies that yield results of large magnitude
III. Opinions of respected authorities, based on clinical experience; descriptive studies or case reports; reports of expert committees

It is difficult for the experimentalist paradigm to account for, say, why uncontrolled studies with large effects are ranked higher than expert reports, since neither is structurally similar to RCTs nor can either be adapted to suit the experimentalist paradigm.

The pragmatist paradigm, by contrast, uses a different mode of justification. Under this paradigm, RCTs are not regarded as the gold standard, and diverse sources of evidence work in tandem to eliminate alternatives (e.g. biases), as we shall see in Sections 4.2 and 4.3 — for example, one piece of evidence is used to rule out biases that cannot be eliminated by another piece of evidence. The paradigm may, in most cases, acknowledge certain types of evidence, such as well-designed and well-conducted RCTs, as the strongest, or the most reliable form of, evidence, not because of what they

[58] Other hierarchy systems include but not limited to the Australian National Health and Medical Research Council (NHMRC)'s (1999) hierarchy, the Oxford Centre for Evidence-Based Medicine (OCEBM)'s (2011) hierarchy and the World Health Organization (WHO)'s hierarchy. The kinds of evidence listed in these systems may vary slightly but are classified hierarchically in terms of similar levels.

are, nor simply because they conform to Mill's methods. The advocates of the pragmatist paradigm instead attribute the emergence of the strongest evidence to their propensity to satisfy specific conditions. The primacy of the RCTs can be acknowledged in the pragmatist paradigm, but for reasons deeply entrenched in a methodological foundation shared with other sources of evidence. In what follows, I will introduce the inferentialist account of evidence, which Reiss proposes to articulate the methodology of evidential reasoning within the pragmatist paradigm.

## 4.2 Supporting Evidence

Reiss's ambition in developing the inferentialist account of evidence is to characterise the ways scientific claims are, and should be, inferred. This account is meant to articulate the pragmatist paradigm both descriptively and reconstructively and allows for the role of pragmatic criteria and diverse sources of evidence.[59] For Reiss, the term 'evidence' conflates two notions: supporting evidence and warranting evidence. Supporting evidence is a 'mark or symptom' of the truth of a hypothesis and warranting evidence is a 'reason' for the truth of a hypothesis (ranging from weak reason to guarantor) (Reiss 2015b, p. 60). Supporting evidence is a fundamental concept on which warranting evidence relies and which can be further classified into two types: direct evidence (evidence *for* a hypothesis) and indirect evidence (evidence *against* alternative hypotheses). In explaining how direct evidence and indirect evidence work, Reiss draws upon the '*hypothetico'* component of the hypothetico-deductive (H-D) account of evidence without the burden of the '*deductive'* component.

The deductive component burdens the H-D account of evidence with a number of difficulties. The standard H-D account states that a claim E is evidence for a hypothesis H, given relevant background knowledge B, just in

[59] A philosophical account aims to describe scientific practice, meanwhile providing a rational reconstruction of the criteria working scientists employ.

case H in combination with B entails E.[60] However, the H-D account of evidence has been considered fraught with defects (see, e.g. Achinstein, 2001; Glymour, 1980a; Lipton, 2004; Norton, 2003; Reiss, 2015b; Woodward, 2011). I do not belabour all of the problems with the H-D account and instead focus on the relationship of inheritance between the H-D account and the inferentialist account. Specifically, I would like to draw attention to three substantial problems (Problems A–C) concerning the component of deductive entailment, abandoned by the latter account. In what follows, after an introduction of a substitute component for deductive entailment, I shall show to what extent the inferentialist account can address these problems.

On the one hand, deductive entailment fails to guarantee that E is evidence for H. For one thing, in *Problem A*, trivial implications would not be regarded as evidence in scientific practice. Suppose there is a fabricated law called Lepler's first law, which mistakenly states that the planets revolve around the sun in square orbits and that this law entails the existence of the sun. The existence of the sun, according to the H-D account, counts as evidence for Lepler's erroneous first law. Alternatively, in *Problem B*, since any claim deductively entails itself, it should be evidence for itself. Nevertheless, one should not regard a claim as evidence for itself, in that if the evidence were identical with the hypothesis that it purports to support, it would be pointless to infer a hypothesis whose truth is already known beforehand.[61] On the other hand, in *Problem C*, evidence is not restricted to any claim that can be deduced from a hypothesis. As the previous discussion illustrates, well-designed and well-conducted RCTs, cohort studies, case-control studies,

[60] The requirement of the truth of E is included in some versions of the H-D account. For instance, Achinstein (2001, p. 147) incorporates the truth of evidence into the basic H-D condition. But this requirement is not found in many other versions (see, e.g. Woodward, 1983; Bird, 2010b).

[61] One can, with the following example, doubt the 'already-known' reason. Suppose H: She ate the cake that was in the kitchen, $E_1$: H, $E_2$: There are some cake crumbs left on her fingers, and $E_3$: it was not possible that any people other than her had been in the kitchen. $E_2$ itself is normally regarded as evidence indicating H. But if we know $E_3$ before knowing $E_2$, does this mean that $E_2$ is no longer evidence in this situation in the same way $E_1$, identical to H, is not evidence for H? Obviously not, though $E_2$ is of no use in indicating H, given knowing a sufficient reason $E_3$. For we can find a counterfactual situation in which $E_2$ were to be evidence for H, if $E_3$ would not be present, whilst we cannot find any like this for H ($E_1$). This distinction can be further made sense of by the notion of *evidential warrant transmission failure* that I have elaborated on in Chapter 3.

multiple time series, large-effect observational studies, expert opinions and case reports are all accepted as sources of evidence, despite their results not being deductively entailed by hypotheses.[62]

In spite of the numerous problems, many biomedical and social scientists still take the H-D method as a standard model of evidential reasoning (Reiss 2015b, pp. 60-61). According to Reiss's diagnosis, the tension consists in deductive entailment as the standard of assessment of evidence; the deduction requirement for evidential relationships should be abandoned and replaced with what we may call *loose implication*. Reiss (2015b, p. 61) subsequently advances a framework that resurrects 'the spirit of the hypothetico-deductivist theory but does not suffer from the problems and counter-examples that beset its logical positivist formulation'. In a similar spirit of the H-D method (particularly regarding its 'hypothetico' component), Reiss (2015b, p. 63) proposes two types of supporting evidence that scientists use to reason from evidence to hypothesis:

> **Direct evidence:** $E_d$ is direct evidence for a hypothesis H if and only if $E_d$ is a pattern in the data that one is entitled to expect, supposing that H is true.

> **Indirect evidence:** $E_i$ is indirect evidence for a hypothesis H if and only if $E_i$ is a pattern in the data that is incompatible with what one is entitled to expect, supposing that one of H's competing hypotheses H′, H″, H‴ and so on is true.

From the above definitions, an evidential relation between H and E is an inductive, rather than deductive, relation, i.e. H need not imply E logically, or

[62] Reiss illustrates this point by listing an array of evidence for causal claims. Reiss (2015b, p. 349) argues that a causal hypothesis that 'C causes E', even with background knowledge, does not logically entail claims about certain patterns, such as a correlation between C and E, a change in E after an intervention on C, C's constituting an INUS (Insufficient but Necessary parts of a condition which is itself Unnecessary but Sufficient) condition for E, a continuous process from C to E, or a mechanism linking C and E. Nonetheless, as Reiss (2015b, pp. 345-346) points out, claims about these kinds of patterns can be evidence for a causal hypothesis.

vice versa. Reiss (2015a, p. 346) construes evidential relations as those that hinge upon 'our understanding of how the world works', and he adopts a broad view on the nature of evidence, not restricted to statements — suggesting that evidence is a pattern in the data that, if a hypothesis is true, scientists are entitled to expect to obtain. Under the definitions of supporting evidence above, patterns may, if understood broadly enough, embody statements, entities, beliefs, figures, indexes and so forth. I am not concerned here with the taxonomy of things that can be evidence. If we follow customary usage in the philosophy of science, Reiss's terminology 'expect to obtain patterns' can be understood as 'loosely imply', statement-to-statement relations. Statements, to some extent, can cover those pattern-related things by making statements *about* them. Since deduction is a relation between statements/propositions, the way of understanding makes explicit the distinction between deductive entailment in H-D and loose implication in Reiss's conception of evidence on a syntactic-semantic level.[63] I will therefore use 'expect' for epistemic contexts and 'loosely imply' for logical-linguistic contexts for the rest of the chapter.

One way to understand Reiss's definitions of supporting evidence is as follows: if $E_d$ is a statement loosely implied by H, $E_d$ directly supports H by signalling the *inferential possibility* of H. If $E_i$ is a statement that is compatible with H but goes against expectations from H's competing hypotheses, $E_i$ indirectly supports H by undermining the plausibility of H′, by making other inferential possibilities implausible. By 'inferential possibility', I mean that a statement is inferentially relevant to its target via loose implication in a particular context. Inferential possibility is not the same as logical possibility, the latter of which is realised simply via non-contradiction. Here I use 'inferential possibility' to highlight the contrast between possibility and plausibility; in short, for the inferentialist account, inference begins from

[63] Another reason for me not to adhere to the term 'pattern' used in the definitions of evidence is that Reiss takes, for example, murder weapons as patterns. Philosophically, anti-realists can assert that objects, to the greatest extent, are inferred through our perception, whilst intuitively, the claim that objects such as murder weapons are patterns seems odd. To avoid such controversy, 'loose implication' is sometimes an adequate substitute.

possibility and eventually ends at plausibility, as I shall explain.

The distinction between the two types of supporting evidence can be understood as a distinction between what I call a *projection strategy* and an *elimination strategy*. The former is to catch relevant, or inferentially possible, hypotheses with an H-D style of fishnet called ‘loose implication’, whilst the latter is to sift the plausible hypotheses from the rest through a sieve called ‘eliminativism’. In more detail, the strategy of direct evidence is captured by the standard H-D account of evidence that holds that ‘H *entails* E if E is evidence for H’ and can be modified as ‘H *loosely implies* E if E is evidence for H’ and asks ‘What patterns would one be entitled to expect if H were true?’. If H loosely implies E, H is supported by E. Direct evidence aids in informing us of what claims are relevant provided that H is true. For instance, if it is hypothesised that two events are related causally, they are expected to be correlated. Suppose that an event A and another event $C_1$ are correlated and there is another event $C_2$ also correlated with A. These correlations would provide direct evidence for the hypothesis that $C_1$ and $C_2$ are both causes of A.

Being an indicator of inferentially possible hypotheses, direct evidence itself does not suffice to indicate that a hypothesis of interest is plausible compared to competing hypotheses. Direct evidence could allow competing hypotheses that coexist: if $H_1$ loosely implies $E_d$, $H_1$ is supported by $E_d$, whilst if $H_2$ also loosely implies $E_d$, $H_2$ is also supported by $E_d$. The indirect evidence strategy helps here, in that it has to do with the elimination of alternative hypotheses. Suppose there is something compatible with (not necessarily inferentially compatible) $H_1$ but incompatible with $H_2$, this is called, according to Reiss, indirect evidence for $H_1$ by eliminating $H_2$.

Direct and indirect evidence can be exemplified by a case where there are several possible causes for my toothache. It could result from dental caries, periodontitis, cracked teeth or trigeminal neuralgia, whereas the toothache is, in Reiss’s terminology, direct evidence for any of these hypotheses of the causes of the toothache. After a comprehensive examination, my dentist

ruled out tooth nerve-related causes, such as tooth decay, periodontitis and a cracked tooth, because he found that the painful tooth had been treated with root canal therapy, which indicated that the nerves of the tooth had been removed. This is indirect evidence that my toothache was caused by trigeminal neuralgia, because the removal of the tooth nerves is incompatible with the hypothesis that any dental caries, periodontitis or a cracked tooth exists. It follows that whilst direct evidence supports multiple hypotheses, indirect evidence plays a role in eliminating alternatives.

Note that Reiss seems to conflate 'E is expected if H is true' with 'H explains E' and 'H and E are incompatible' with 'H cannot explain E'. Carefully distinguishing 'expectation/loose implication' and 'inferential incompatibility' from 'explanation' and bringing the concept of *inferential irrelevance* into play, can avoid the problem of trivial implication (Problem A above). These conflations can be found in many contexts: 'If a genetic factor were appealed to in order to *explain* this observation, there would have to have been a mutation in males first and a few decades later in females, a pattern that had not previously been observed' (Reiss, 2015a, p. 353, my italics), 'These confirmed a dramatic increase in lung cancer risk among smokers but could not be *accounted for* by Berkson's paradox' (Reiss, 2015a, p. 353, my italics), 'However, the misclassification hypothesis cannot *explain* micropatterns in the data' (Reiss, 2015a, p. 353, my italics) and so forth. These sentences indicate both that if H explains E, then E is expected supposing that H is true, and that if H cannot explain E, then E fails to be evidence for H, which, according to Reiss, would mean that H and E are incompatible. If 'incompatible' were understood in a strictly logical sense, the definition of indirect evidence would not accommodate what I quoted above, because, for example, Berkson's paradox is logically compatible with the epidemiologic trend in lung cancer among smokers.[64]

[64] Put simply, Berkson's paradox is a general pattern where 'observations on a common consequence of two independent causes tend to render those causes dependent, because information about one of the causes tends to make the other more or less likely, given that the consequence has occurred' (Pearl, 2000, p. 17).

The way of attempting to cash out 'expect/loosely imply' and 'incompatible' in terms of 'explain' has some initial appeal, but upon close examination turns out to be problematic. Reiss's definition of indirect evidence is defined as a pattern of the data (E) that is incompatible with what is expected, supposing that H′ (a competing hypothesis of H) is true (E′). But he occasionally claims that '[i]ndirect support is given by patterns in the data that are *incompatible with the truth of an alternative hypothesis*' (Reiss, 2015b, p. 63, italics in original), pointing out that E is incompatible with H′, not E′. The first use of the term 'incompatible' is taken as 'logically incompatible', meaning that E and E′ cannot both be true. For the second use, I construe 'incompatible' as *inferentially incompatible*, as E and H′ are not necessarily logically incompatible. Although Reiss does not explicitly define 'incompatible′ in his account, this term in the second use cannot simply be formulated as: H and E are inferentially incompatible if and only if H cannot explain E. Whilst the inferential incompatibility between H and E does imply that H cannot explain E (i.e. if H can explain E, H and E should be inferential compatible), the converse does not hold. For example, a hypothesis that K is the murderer ($H_k$) explains neither that K has an alibi ($E_1$) nor that K has human DNA ($E_2$). $H_k$'s failure to explain $E_1$ shows inferential incompatibility between $H_k$ and $E_1$, whilst $E_2$'s being unexplainable in the light of $H_k$ arises from *inferential irrelevance*, upon which I will expand below.

Problems also arise if 'expect/loosely imply' is conceptualised as: E is expected supposing that H is true (namely, H loosely implies E) if and only if H explains E. Although it follows from H's explaining E, that H loosely implies E, it need not be the case that H loosely implies E only if H explains E. To illustrate, consider a classic example of causation. Let us suppose $H_{ap}$: The atmospheric pressure drops, $H_s$: A storm occurs and $E_b$: A barometer falls. Whilst $H_{ap}$ causally explains $E_b$ and thus loosely implies $E_b$, $H_s$ only loosely implies $E_b$ not by virtue of explaining $E_b$. Loose implication covers more situations than does explanation, that is, explanation is logically stronger than loose implication, or expectation. Arguably, appealing to 'explain' does not fully explicate 'expect/loosely imply' and 'inferentially incompatible' for the inferentialist account.

To help formulate ‘expectation/loose implication’ and ‘inferential incompatibility’ as well as possible, my suggestions are as follows.

> **Expectation/Loose implication:** E is expected supposing that H is true (i.e. H loosely implies E) if and only if E, if H is true, is likely enough to occur relative to a particular context, i.e. $P(E/H\&C) > \alpha$, where C is a related context and α is the standard of occurrence relative to E, H and C.
>
> **Inferential incompatibility:** H and E are inferentially incompatible if and only if E is unlikely enough to occur relative to a particular context, supposing that H is true, i.e. $P(E/H\&C) < \varepsilon$, where ε is the standard of non-occurrence relative to E, H and C.
>
> **Inferential consistency:** H and E are inferentially consistent if and only if the occurrence of E, supposing that H is true, is between likely enough and unlikely enough relative to a particular context, i.e. $\varepsilon \leq P(E/H\&C) \leq \alpha$.
>
> **Inferential irrelevance:** H and E are inferentially irrelevant if and only if E, relative to a particular context, is equally likely to occur regardless of H (or H, relative to a particular context, is equally likely to be true regardless of E), i.e. $P(E/H\&C) = P(E/C)$, or equivalently $P(H/E\&C) = P(H/C)$.

These notions can be defined in either qualitative or quantitative ways, as shown above; the former is what Reiss may maintain for the inferentialist account and the latter is easier to capture at a glance. The context renders the standard highly domain-sensitive by which inferential judgments are made. Context is an essential element of the inferentialist account, and I will come back to this in Section 4.4. Note that the universal lower bound on α is required to be greater than 1/2, on the basis that if E is expected, supposing that H is true, then E should be more likely to occur than ~E. For example, if

a coin toss is expected to come up heads ($E_h$) supposing that the coin is heads-biased ($H_b$), then $P(E_h/H_b) > P(\sim E_h/H_b)$, which amounts to $P(E_h/H_b) > 1/2$. Similar logic applies to the universal upper bound on ε. If that the coin is heads-biased ($H_b$) and that the tossed coin lands tails ($E_t$) are inferentially incompatible, at least $P(E_t/H_b) < P(\sim E_t/H_b)$ and thus $P(E_t/H_b) < 1/2$.

To see whether the definitions by reference to 'likely' work, reconsider the murderer example, in which $H_K$: K is the murderer, $E_1$: K has an alibi and $E_2$: K has human DNA, added with $E_3$: K's DNA is left on the murder weapon and $E_4$: K was the victim's friend. If in a particular context (e.g. legal or investigative), $E_3$ is judged to the extent that it is likely enough to occur for $H_k$, $H_k$ loosely implies $E_3$. Since $E_1$ would be extremely unlikely supposing that K committed the murder relative to the same context, $E_1$ and $H_k$ are inferentially incompatible. If K were the murderer, there would be neither any assurance nor denial that K had a friendship with the victim; $E_4$ and $H_k$ are inferentially consistent. Lastly, since the murderer is human, $H_K$ is equally likely regardless of $E_2$; $E_2$ provides no information about the truth of $H_K$ and is inferentially irrelevant to $H_k$. Integrating inferential consistency and irrelevance into the notions of evidence can complement the inferentialist account.

Inferential compatibility is distinct from logical/physical compatibility. H and E are logically compatible if and only if it is not the case that the conjunction of H and E is necessarily false, and H and E are physically compatible if and only if it is possible that H and E are true in accordance with physical laws. In these senses, inferential compatibility is (logically) stronger than logical/physical compatibility by virtue of implying the latter two types of compatibility. That is, if H and E are inferentially compatible, they are logically/physically compatible, but not vice versa. For example, the weapon is a pattern to be expected if K is the killer, which is logically/physically compatible with the hypothesis that the person died by accident. However, the accident hypothesis and the weapon found are inferentially incompatible, since the weapon is not expected to appear, supposing that the accident hypothesis is true.

So far I have established three points. The first is to argue that expectation/loose implication and inferential incompatibility cannot be fully conceptually captured by explanation. The second is to suggest that expectation/loose implication, inferential incompatibility, inferential consistency and inferential irrelevance can be fleshed out in terms of qualitative and quantitative relationships between H, E and context. The third is to distinguish inferential compatibility from logical/physical compatibility. In what follows I delve into the notion of supporting evidence more deeply.

Armed with loose implication which differs from deductive entailment, can Reiss's notion of supporting evidence be immune to the difficulties mentioned above, namely, restricted sources of evidence (Problem C), self-deduction (Problem B) and trivial implications (Problem A), those that the H-D method encounters? As regards Problem C, from Reiss's point of view, since a causal claim loosely implies a claim about correlations,[65] relations of invariance and processes, these implications, which do not count as evidence in the H-D account, are admissible as direct evidence for the causal claim.

However, in broad terms, a claim deductively derived from a hypothesis is expected to emerge on the supposition of the hypothesis, so Reiss's notion of supporting evidence seems to face the same difficulties as the H-D method, including self-deduction, like the claim that E is evidence for itself, and trivial implications, like the case of Lepler's first law. Reiss (2015a, p. 346) gives a simple reason against self-deduction as evidence (Problem B), claiming that 'any statement entails itself, but no self-respecting biomedical or social scientist would take the truth of a hypothesis as support for itself'. His reason is unconvincing, as it shows that the inferentialist account does not fulfil the purpose it sets out, which is, in part, to account for real practices,

[65] Correlation neither is nor 'logically' implies causation but can be evidence for causation (Russo, 2015a). Nor does causation 'logically' imply correlation. These claims do not contradict the claim that causation 'loosely' implies correlation, as loose implication is inductive rather than deductive.

and it also shows that the H-D advocates can also take advantage of this reason to defend the H-D method in this respect.

A possible way to avoid the self-deduction difficulty is to appeal to the definitions of supporting evidence. Reiss does not confer legitimacy to self-deduction, presumably on the basis that evidence is a pattern in the data we are entitled to expect to find under the supposition of H. He can just hold that in the case of self-deduction, H does not constitute a pattern in the data if H is true — or conversely, if E is evidence for H, then E must not be identical with H and thus there is no self-deduction. Otherwise, Reiss should bite the bullet, admitting that if under the supposition that H is true, H itself is an expected pattern, then any whimsy can be evidence for itself. This, however, would be absurd.

The appeal to the notion of supporting evidence is not viable. The main contention can be that supposing H is true, we are entitled to expect a pattern exactly the same as the description in H; does this not mean that if H entails itself, then H is evidence for itself? The crux of the problem lies in a lack of a clear characterisation of 'patterns in the data' and 'data'. Despite not proffering such a characterisation, Reiss (2015b) gives a clear-cut example that instantiates a distinction between the two concepts:

> I say 'patterns in the data' instead of 'data' because scientists aren't normally entitled to expect specific data sets. That a coin is biased towards heads entitles us to expect that there are more heads in a series of tosses but not 17 out of 20. The same is true of causal hypotheses. (Reiss, 2015b, p. 76)

Let us consider this example: $H_b$: The coin is biased, $E_h$: In a series of tosses, the coin comes up heads far more often than tails and D: There are 17 heads in 20 tosses. Reiss's idea is as follows. A particular number of the heads recorded as the data D is not expected if $H_b$ is true, whilst $E_h$, a pattern in the data, is expected if $H_b$ is true. Accordingly, 'a pattern in the data' represented by $E_h$ and 'data' by D are clearly distinguished.

Reiss seemingly has to admit that the notion of supporting evidence still confronts the problem of self-deduction when the following example is considered. Suppose there is a hypothesis ($H_h$) identical to $E_h$. To allow maximum latitude for Reiss to avoid the problem of self-deduction, we may agree to accept that whether something is a pattern in the data is relativised to a hypothesis. It can be argued that $E_h$ is a pattern in the data relative to $H_b$ but not relative to $H_h$. If so, $H_h$ is not direct evidence for itself. However, it is hard to conceptualise another pattern, other than $E_h$, that is expected to be found supposing that $H_h$ is true. Nor can D be direct evidence for $H_h$. Supposing that $H_h$ is true, whilst we would assert that we are entitled to expect that there are two heads in two tosses (D'), we would hesitate to say that we are entitled to expect a specific data set such as D, as there are many possible sets of data that constitute heads outnumbering tails in 20 tosses. Theoretically, any claim between $E_h$ and D in terms of the abstract-concrete extent can be a candidate for direct evidence for $H_h$, but it is hard to conceptualise such a claim. Considering these issues, judging whether something is a pattern in the data or is just the data remains perplexing.

A possible reply is that in the coin example, since $E_h$ is 'what is perceived' rather than 'what actually exists' captured by $H_h$, $E_h$ is not $H_h$ itself and thus is not evidence for itself. This reply, however, involves persistent and unresolved debates in metaphysics, epistemology and the philosophy of mind. Perhaps a straightforward solution to the self-deduction difficulty would be one mentioned in Chapter 3. That is, scientists do not need evidence to establish what is already known; $H_h$ is not evidence for itself insofar as it is not possible that $H_h$ remains unknown when $E_h$ is known. This solution can be implied by Reiss's notion of *context*, which I will visit in Section 4.4. More importantly, this solution would not benefit the H-D account: unlike expectation/loose implication, which is sensitive to context, the background knowledge B involved in the H-D definition does not change the fact that a statement entails itself. Thus, Reiss need not draw upon real scientific practices themselves to explain the illegitimacy of self-deduction as evidence.

As regards trivial implications (Problem A above), appealing solely to loose implication or expectation is not enough to solve this problem. Entailed existential claims are typically not taken as evidence for causal hypotheses. However, for the inferentialist account, if an existential claim is entailed by a causal hypothesis, then the existential claim constitutes direct evidence for the hypothesis by being implied by the hypothesis. The existence of the sun should count as a pattern to be expected under the supposition of Lepler's first law and thus serve as direct evidence for Lepler's first law. Upon discussing the problem with the H-D method, Reiss (2015b, p. 40) argues that '[c]ausal hypotheses do, however, entail existential claims …. the existential claim is not relevant to the truth of a causal hypothesis. Hypothetico-deductivism is therefore not a good theory of support'. Whether Reiss's reason here is cogent or not hinges upon what he means by 'relevant'. As noted earlier, the relevance that he intends is inferential relevance, which cannot be fleshed out merely with the notion of loose implication or expectation.

In order to insulate the inferentialist account from the problem of trivial implication, the notion of inferential irrelevance I have outlined can be invoked. A note of caution must be issued: an existential claim can sometimes be evidence for hypotheses. To illustrate, consider a case of *Helicobacter pylori* (*H. pylori*), a bacterium that colonises the gastric and duodenal mucosa. Peptic ulcers had been widely believed to be primarily caused by stress and spicy food ($H_{ss}$) before the early 1980s when *H. pylori* was identified ($E_{hp}$) and hypothesised as (and actually is) a major causative factor of peptic ulcers ($H_{hp}$) (Johnson *et al.*, 2007). According to the inferentialist account, the existence of *H. pylori* ($E_{hp}$) is supporting evidence for $H_{hp}$ by virtue of being expected to appear supposing that $H_{hp}$ is true. Moreover, $E_{hp}$ is inferentially relevant to $H_{hp}$, because $E_{hp}$ is more likely to appear under $H_{hp}$ than without. Indeed, the history of how $H_{hp}$ was confirmed is not straightforward: $H_{hp}$ was well supported by other sources of evidence. The findings were that those with *H. pylori* are more likely to have ulcers than those without, i.e. P(ulcers/*H. pylori*) > P(ulcers/~*H. pylori*), and that patients' recovery covaries with their taking antibiotics (Thagard, 1998, p. 66). Also,

although it was widely held that the stomach was too acid for a bacterium to survive, a piece of mechanistic information subsequently became available about *H. pylori*'s capability of neutralising stomach acid by secreting ammonia (Thagard, 1998, p. 70). These pieces of evidence jointly justify $H_{hp}$, which is concerned with the notion of warranting evidence (see Section 4.3). The aforementioned sun example is, by contrast, where inferential irrelevance comes into play: although the existence of the sun may be direct evidence for every competing hypothesis, it is inferentially irrelevant to Lepler's first law, no matter whether we consider Lepler's first law or not. The upshot is that the existential claim, if embodied within all of the competing hypotheses, can be disregarded as non-evidence insofar as it is not relevant to the inference. Trivial implication does not pose a problem for the inferentialist account.

It is noteworthy that the unobservable can serve as evidence if it satisfies the requirements of evidence considered in the inferentialist account. Reiss's definitions of supporting evidence apply patterns, which are not necessarily observable. The toothache example above applies the observable as evidence,[66] including the painful tooth and the records of root canal therapy.[67] However, in some cases the unobservable can function as evidence. For instance, Reiss (2015b, p. 62) claims that '[c]orrelations, similarly, are best thought of as theoretical relations that can be estimated using one or another measure (such as the Pearson correlation coefficient) but which is not observable as such…'. The marks of causal relationships (e.g. correlations) may be unobservable, but this does not necessarily mean that these unobservable marks cannot be expected from the data and thus cannot be evidence. Rather, they, through inferences, gain the entitlement to indicate the truth of the hypothesis on the grounds that they are constructed or established by virtue of the observable.

[66] In this case, whether my toothache is observable or not is debatable. Whilst it is observable to me, it is not observable to my doctor unless some kind of behaviourism is adopted that reduces 'toothache' to certain kinds of behaviour.
[67] The fact that the nerves have been removed is also observable.

An ordinary example is when we see a person standing outside shivering on a snowy day. We are initially likely to infer that he is feeling cold, which is, to some extent, evidence that his clothes cannot keep him warm or that he needs to come inside and warm up. The claim that he feels cold, if unobservable for us, is not groundless: it is inferred, on the basis of the observations and functions as evidence for further hypotheses. Similar concerns apply when correlations, theoretical constructs, stem from the observations of the corresponding increase (or decrease) among two variables, which can be recorded on a spreadsheet. While the data recorded in the spreadsheet is evidence for a correlation, correlations, though unobservable, can be evidence for causal hypotheses. It can be seen that the notion of patterns used to define supporting evidence accommodates both the observable and the unobservable as evidence.

### 4.3 Warranting Evidence

Having discussed supporting evidence, we can now move on to warranting evidence, which results from combining direct evidence and indirect evidence. While the strategy of direct evidence modifies the standard H-D method, the strategy of indirect evidence employs eliminativism. If there is a body of evidence in favour of $H_1$ and there are other pieces of evidence for eliminating alternative hypotheses $H_2$ and $H_3$, we have a good reason to infer $H_1$. Reiss (2015b, p.73) distinguishes four levels of warrant: proof, strong warrant, moderate warrant and weak warrant.

| | |
|---|---|
| **Proof:** | All relevant competing hypotheses are eliminated. |
| **Strong warrant:** | All salient competing hypotheses and some non-salient are eliminated. |
| **Moderate warrant:** | Most competing hypotheses, some of which are salient, are eliminated. |
| **Weak warrant:** | Some competing hypotheses, none of which are salient, are eliminated. |

Note that Reiss here uses the term *salient*. Salient alternatives are 'alternatives for which there exists direct support, [that] contribute more to the strength of the warrant than non-salient alternatives because a true alternative is more likely to leave traces in the data than a false alternative' (Reiss, 2015a, pp. 357–358). 'Salient' is used to mark out competing hypotheses that are supported by additional direct evidence beyond the direct evidence shared between the target hypothesis and its competing hypotheses.

It should also be noted that Reiss does not ascribe more evidential warrant to $H_1$ than to $H_2$ merely because $H_1$ acquires more supporting evidence than $H_2$. For Reiss, supporting evidence is construed as an 'indicator of the truth" rather than typically seen as the 'reason for the truth"; it only tells us 'what kinds of facts we have to collect in order to evaluate a hypothesis' (Reiss, 2015a, p. 343) and supports a hypothesis 'without yet constituting a reason to infer the hypothesis, even a weak one' (Reiss, 2015a, pp. 342–343). Supporting evidence does not even require that the must-collect facts be *true* propositions,[68] so if $H_1$ has more supporting evidence than $H_2$, it is not necessary that there be more facts in support of $H_1$ than $H_2$. Warranting evidence focuses on a hypothesis that can best survive the process of elimination and thus gives a good reason to believe. Competing hypotheses, if not eliminated, do not have less warrant than the focal hypothesis, even if they have less supporting evidence.

Put differently, even assuming that there are more facts supporting $H_1$ than $H_2$, both can be equally warranted unless one of them is eliminated, because more direct evidence just tells us more about what would be relevant to the hypothesis. For the inferentialist account, the only legitimate way to assess

[68] The non-requirement for the truth does not appear to accord with sound common usage: $E_d$ and $E_i$ in the definitions of supporting evidence are not required to be true. A way to avoid this discord is, I suggest, to distinguish between *potential* and *real* supporting evidence in an analogous way that Hempel (1965) or Achinstein (2001) made a distinction between potential and correct explanation. Therefore, one may claim that E is potential supporting evidence just in the case it meets either of the definition of direct or indirect evidence, and that E is real supporting evidence just in case it is true beyond being potential supporting evidence.

the warrant of the focal hypothesis is by assessing to what extent its competing hypotheses, sharing the same direct evidence with the focal hypothesis, are eliminated by indirect evidence, and not by assessing by how much additional direct evidence it obtains insofar as the additional direct evidence, if consistent with competing hypotheses, does not eliminate them. The role that direct evidence plays in the assessment of warrant is determining whether competing hypotheses other than the focal hypothesis are salient or not. However, failure to falsify some competing hypotheses does not amount to a failure to distinguish the evidential strength between them. The latter is what the inferentialist account is silent about and is what I will explore in Section 4.5.

Importantly, the epistemic notion of proof here is distinctively different from the mathematical or logical notion of proof. The latter notions of proof possess deductive certainty, that is, provided that the premises are true, the conclusion following from them must be true. As far as evidential support is concerned, the epistemic proof is inductively strongest without possessing absolute certainty. This kind of proof is fallible even if all competing hypotheses have been eliminated. This is because all possible competing hypotheses may not be collectively exhaustive (i.e. the combination of H's competing hypotheses under consideration does not necessarily fully cover the negation of H), or competing hypotheses that have been ruled out should not have been eliminated. The different levels of non-proof warrant are distinguished in terms of how many salient hypotheses are eliminated on the basis of indirect evidence. Put simply, if epistemic proof is obtained by virtue of the elimination of all competing hypotheses, then we have the strongest, though still fallible, reasons for a causal claim 'C causes E', for example. But if only strong (moderate/weak) warrant for this causal claim is obtained, then we only have reasons for the causal claim of 'C is very likely to cause E' ('C may cause E'/'C might cause E').

The preceding example of my toothache is a case in point, indicating that the strength that warranting evidence gains hinges upon the extent to which salient and non-salient hypotheses are eliminated. My dentist initially

suspected that my toothache was caused by a decayed tooth ($H_d$), a cracked tooth ($H_c$), periodontitis ($H_p$) or trigeminal neuralgia ($H_t$). The hypothesis ($H_d$) is then directly supported by the fact that he, deducing from X-rays, found that there was indeed a cavity. $H_d$ is accordingly a salient hypothesis. By contrast, he did not find any crack on the tooth or any problem with the soft tissues of my gum. $H_c$ and $H_p$ are non-salient hypotheses, as they do not have direct support, except for my suffering from a toothache. $H_c$ and $H_p$ are not the only non-salient hypotheses when my dentist could not exclude other non-salient hypotheses, such as myofascial pain, psychogenic toothache, brain tumour and any other hitherto known non-dental causes. $H_d$, $H_c$ and $H_p$ are all eliminated by the finding that the nerves of the tooth had been removed. $H_t$ is thus strongly warranted because the only salient competing hypothesis $H_d$ and non-salient competing hypotheses $H_c$ and $H_p$ have been ruled out. The dentist finally suggested that I see a neurologist for further neural examination. This strong warrant does not arise from *a priori* relations between H and E, as stated in the H-D account, but rather relates to a matter of a posteriori investigation.[69] From the viewpoint of the inferentialist account, only through eliminating competing hypotheses empirically can evidential warrant be bestowed upon E.

As noted in Section 4.2, the inferentialist account can be invulnerable to the trivial implication problem (Problem A). The existence of the sun is a pattern that scientists are entitled to expect to see in the data if Lepler's first law is true, but it is a trivial piece of evidence because of its irrelevance to the hypothesis. Warranting evidence offers another way of avoiding Problem A. Trivial implications do not serve as warranting evidence or are ignorable in the body of warranting evidence, insofar as they hardly provide indirect evidence against alternative hypotheses. For instance, even granting that the existence of the sun is also direct evidence for Kepler's first law, it cannot differentiate Kepler's first law from Lepler's first law, hence contributing

[69] Reiss's conception of evidence is that evidential reasoning lacks universal schemas that are applicable everywhere as it is material and idiosyncratic to different settings (cf. Norton, 2003). In the light of this conception, his inferentialist account of evidence holds that whether E is evidence for H is a matter of empirical facts rather than a priori relations described in the H-D account.

nothing to determining the levels of evidential warrant. Existential triviality can be direct evidence for and thus inferentially irrelevant to every alternative hypothesis; if so, it is of no evidential importance. Compared to an H-D advocate, which offers only one kind of evidence via deductive entailment and has to acknowledge that existential triviality is evidence for a hypothesis, Reiss can painfully accept that existential triviality is direct evidence, as well as denying that it is indirect evidence or warranting evidence.

Having sketched relevance-mediating vehicles in Chapter 1, here I want to specify the relevance-mediating vehicles in the inferentialist account. Relevance-mediating vehicles can be understood as configurers setting up materials or claims in particular ways so that some of the materials or claims are labelled evidence. Put succinctly, evidential relations do not already exist but emerge from a setting. From the previous discussion, labelling an established fact 'direct' evidence is to establish its relevance to a hypothesis in question, and labelling it 'indirect' evidence severs the putative relevance of old patterns to an alternative hypothesis.[70] The relevance, i.e. inferential possibility, is established by loose implication and severed by the negation of loose implication. It can be seen that the relevance-mediating vehicles at work are loose implication and elimination of possibilities.

Let us illustrate this with the example of the toothache. The claim that I had a toothache (i.e. direct evidence) is inferentially relevant to the hypothesis that my toothache was caused by trigeminal neuralgia. The claim that my tooth had been treated with root canal therapy severed the relevance of my toothache to the hypothesis that my toothache was caused by a decayed tooth. Warranting evidence, by virtue of supporting (with direct evidence) and eliminating (with indirect evidence) inferential possibilities, ensures that the remaining hypothesis/hypotheses in question attain certain plausibility, such that we have reason to believe them. The hypotheses that my toothache was

[70] It is unclear, for Reiss, whether the relevance between an alternative hypothesis and direct evidence remains unchanged (i.e. whether the original direct evidence is no longer direct evidence for the alternative hypothesis), if the alternative hypothesis is eliminated by indirect evidence. If the answer is negative, the removed relevance can still be restored if the indirect evidence is refuted.

caused by a decayed tooth, that my toothache was caused by a cracked tooth, and that my toothache was caused by periodontitis, were all ascribed to inferential impossibility in the light of the indirect evidence (i.e. the claim that my tooth had been treated with root canal therapy), which rendered plausible the hypothesis that my toothache was caused by trigeminal neuralgia.

### 4.4 Context

Having spoken of supporting evidence (consisting of direct and indirect evidence) and warranting evidence, in order to illuminate the vague concept of ‘patterns in the data’ (or ‘loose implication’ instead) to some extent, Reiss appeals to another requirement for the notion of evidence, namely, contextual features.

We realise that supporting and warranting evidence merely provide an abstract structure not indicative of the meaning of loose implication. That is, there is no tangible content about what direct evidence should be expected under the supposition that a certain hypothesis is true, nor about in what alternative hypotheses the same piece of direct evidence is expected, nor about when alternative hypotheses are ruled out by indirect evidence. In this sense, claims loosely implied by a hypothesis of interest may vary depending upon subjective judgement. This is not the way that scientists conceive, understand and adduce evidence: certain types of claims are hardly to be considered evidence in the first place. Nor is it the way scientists generate hypotheses: any claims that loosely imply a claim about a particular pattern can arbitrarily form numerous alternative hypotheses.

If we were to take ‘loosely implies’ without further specification, this could mean that numerous impractical claims come out as evidence or hypotheses. Before presenting Reiss’s solutions to these problems, let us first consider what kinds of constraints are needed on ‘loosely implies’.

Constraints ought to be introduced and placed upon the notion of supporting evidence in the inferentialist account to avoid or at least curb the surplus candidates of evidence. It seems that for the inferentialist account, if we do not demand that patterns be genuine or statements about them be true, that will give us too much of what is taken to be evidence, as well as too many hypotheses to consider. Humans are creatures with a tendency to seek patterns, even false ones, in random and unconnected pieces of information that we believe to be true. For instance, certain patterns in lottery numbers may appear in gamblers' minds. Patterns may be individually subjective, which is an insecure foundation for science, if they are used as evidence. For example, a claim made from the imagination about images obtained by electron microscopy of an unknown variant virus that can cause the trigeminal nerve to transmit pain signals should count, if any, as evidence for the hypothesis that my toothache is caused by trigeminal neuralgia. This is because the hypothesis, premised on the imagination, loosely implies that my trigeminal nerve was infected by the unknown virus. This would bring about an unfavourable situation: even if there is no such virus, it is still direct evidence for the hypothesis, meaning that any non-existing or false patterns can be marks for certain hypotheses. Even where it is hard to distinguish between true and false patterns, contested patterns, according to the notion of direct evidence, should count as marks for hypotheses. For example, a toasted cheese sandwich that emerged with a pattern of the Virgin Mary was sold for $28,000 on eBay in 2004, a pattern which might be regarded as direct evidence for the existence of God. Arguably, constraints are needed to prevent so-called patterns generated from the imagination or excess subjectivity.

One option to curb the surplus candidates of evidence is to demand that statements about evidence be true. It is not an uncommon demand, as evidence is required to be true in Achinstein's explanationist theory of evidence (see Chapter 2), Cartwright's argument theory of evidence (see Chapter 3) and many other accounts of evidence. However, within the schemas of direct evidence and indirect evidence used in the inferentialist account, whether E is true or false remains unknown. Without a constraint

upon the truth of evidence (recall that evidence can be assumed to be statements about pattern-related things), statements about whims, fantasies or other subjectively identified patterns can all count as direct evidence for certain hypotheses, provided that the hypotheses loosely imply the claims about these undesirable things.

Constraints ought also to be imposed upon the selection of hypotheses; otherwise, direct evidence can support numerous so-called salient hypotheses that appear absurd.[71] For example, a hypothesis that an unknown variant virus caused my toothache is supported by the claim that I had a toothache, since someone believes that this hypothesis loosely implies the latter claim. This means that taking the unknown virus hypothesis into account and comparing it with other commonly known hypotheses (e.g. the hypotheses that the toothache was caused by a decayed tooth, that it was caused by a cracked tooth or that it was caused by periodontitis) can be unproblematic, which is not well-suited for the real practices of medical diagnosis. Also, the problem of supported hypothesis surpluses can be exposed by modifying Achinstein's example for the present purpose. Suppose my car did not start this morning (E). I hypothesise that at 2:07 last night 5 boys and 2 girls replaced 18.9 gallons of petrol in my tank with the same amount of water (H). By the definition of direct evidence, under the supposition of H, I am entitled to expect to obtain E. However, there is an unbridgeable gulf between H and E, and E is 'far too meager a reason to believe the very specific hypothesis [H]' (Achinstein, 2010a, p. 14). In order to prevent absurd hypotheses, constraints are required.

In his notions of evidence, Reiss resorts to context in order to impose the constraints mentioned above on a distinctive characteristic, namely loose implication, or patterns in the data. Although the definitions of supporting

[71] Achinstein (2001, pp. 147–149) expresses parallel criticisms against the H-D view on the basis that the H-D view permits numerous 'crazy' hypotheses. High probability and explanatory connection, he suggests, are required to curb surplus hypotheses. A similar concern can be found in Wesley Salmon's (1966, p. 115) remark: 'The basic trouble with the hypothetico-deductive inference is that it always leaves us with an embarrassing superabundance of hypotheses. All of these hypotheses are equally adequate to the available data from the standpoint of the pure hypothetico-deductive framework'.

evidence, whether direct or indirect evidence, are loose, none of them should be discarded. As noted above, constraints are required to narrow the set of hypotheses and constrain the set of evidence for a given hypothesis: on what conditions are we entitled to restrain the set of candidate hypotheses and rule out competing hypotheses? In what circumstances are we entitled to claim 'a hypothesis loosely implies a statement to possibly be evidence' or entitled to expect a pattern in the data? When context is brought into play, arbitrarily conceivable patterns and hypotheses can be reduced in number.

Context dictates the application scope of supporting evidence, based on 'background knowledge about how the world works, the nature and purpose of the inquiry, and certain normative commitments' (Reiss 2015a, p. 349).[72] It is my dentist's background knowledge, in constraining the set of hypotheses and evidence, which enables him to hypothesise that my toothache could be caused by trigeminal neuralgia. The first full description of trigeminal neuralgia dates back to 1773, before which presumably one would not suspect the existence of this type of nerve pain (Prasad and Galetta, 2009). It is also my dentist's background knowledge that enables him to take into consideration and dismiss the possibility of a decayed tooth, a cracked tooth and periodontitis. Contextual features conformed to by medical practitioners prevented my dentist from making an unlikely claim that an unknown virus was evidence for my toothache, and from taking into consideration a hypothesis that my toothache was caused by an unknown virus. Moreover, my dentist's purpose of inquiry was not to conduct pathological research into the causes of my toothache, but to effectively relieve the pain, so causes other than the ones commonly seen were not taken into consideration at the beginning of the diagnosis. My dentist initially suspected that the cause of my toothache was a decayed tooth, presumably because the treatment for a

[72] Different purposes of an inquiry can result in different means of evidence collection. Even for the same purpose, the truth of a hypothesis is, in practice, not the sole consideration, and normative commitments (e.g. value judgements and cost-benefit analyses) should be considered. Contextual features other than background knowledge can be illustrated by a quote from a medical journal paper: 'The question is how much evidence is needed to move from research to practice, when the matter is life saving interventions in poor settings. The yardstick for decision making should take into account the risks and benefits in the local conditions, not those of an ideal situation' (Potts et al., 2006, p. 702).

decayed tooth is relatively less time-consuming and is less likely to bring a medical malpractice suit. These considerations indicate his value judgements (i.e. normative commitments). Thus, contextual features can serve as constraints on loose implication, enabling evidence and hypotheses under investigation to match our inquiry purposes and value considerations.

In addition to constraining the scope of evidence and hypotheses, context conversely enables the formation of evidence. A simple example would be selecting the more likely suspect among X and Y, according to a sketch of the criminal. There are two hypotheses to entertain: $H_1$: X is the criminal and $H_2$: Y is the criminal. Originally, the statements about the striking facial features (e.g. a big mole above the lip, thick arched eyebrows, etc.) in the portrait are a far cry from the statements loosely implied by each hypothesis about who committed the crime, simply because both of the suspects and the portrait are not alike. However, when the police officers are informed that there are no other suspects at the crime scene, this new piece of contextual information (C) enables the portrait to become evidence (E) that the suspect with a rather small mole above the lip and thick straight eyebrows is most likely to be the criminal. In sum, contexts can function as constraints on evidence and hypothesis.

## 4.5 Tool Assessment of the Inferentialist Account

Now it becomes clear why not all RCTs can be viewed as the strongest evidence unless they are well-designed and well-conducted insofar as (almost) all competing, other than causal, hypotheses are ruled out (e.g. selection bias).[73] Poorly designed or poorly conducted RCTs are weakly warranted, due to their inability to exclude many competing hypotheses. For instance, the Grading of Recommendations Assessment, Development and

[73] Selection bias is an error that occurs when the allocation of treatment groups and control groups is not randomised. The outcomes of the study result in part from systematic differences in characteristics between both groups. Hence the causal conclusion cannot be drawn from the outcomes (Last, 2001, p.166). Selection bias is 'eliminated in an RCT because the procedure of random allocation means that the experimenter cannot affect the arm that particular patients are assigned to' (Worrall, 2007b, pp. 453–454).

Evaluation (GRADE) defines four levels of evidence quality: very low, low, moderate and high, which roughly correspond to the four levels of warrant, and the GRADE system also stipulates three factors: large effect, dose-response relation and plausible confounders or biases that may make us underestimate or overestimate the observed effect size — to upgrade levels of evidence quality, and five downgrading factors — risk of bias, inconsistency, indirectness, imprecision and publication bias (Guyatt *et al.*, 2011).

In the light of Reiss's inferentialist account, the operation of all of the factors, whether they upgrade or downgrade, hinges upon the increase or decrease in their ability to rule out competing hypotheses. For example, the convergence of the results between observational studies plays a role of indirect evidence in reasonably ruling out the possibility of sheer chance ($H_{sc}$), but the results are still direct evidence for both causal relations ($H_{cr}$) and selection bias ($H_{sb}$). By randomisation, an RCT rules out $H_{sc}$ and $H_{sb}$; an RCT is not only direct evidence for $H_{cr}$ but is also indirect evidence against $H_{sc}$ and $H_{sb}$. Thus, the inferentialist account can explain that RCTs are ranked higher than observational studies in the hierarchy of medical evidence.

Moreover, Reiss's inferentialist account remains open to a variety of sources of evidence, sometimes even as evidentially strong as well-designed and well-conducted RCTs. This source of evidence can be seen in observational studies (i.e. non-interventional), particularly when the size of effect is too large to be ignored. It is acknowledged that in many cases, observational studies are unreliable; using them as evidence for causal claims is likely to lead us to an erroneous conclusion due to selection bias and other sources of bias.[74] These biases either make false causal claims or conceal real causal relationships, as the confounding factors are not be controlled or balanced between the treatment group and the control group. However, in

[74] Other sources of bias include: reporting bias, detection bias, performance bias and indication bias (Viswanathan et al., 2013).

some exceptional cases, the evidence produced can be graded as high-quality or even strongest, since the effect size of observational studies vastly outweighs the conceivable bias of confounders. For instance, there is no need to carry out an RCT to show that anyone dropped from an aircraft in flight would die and that those with parachutes are very likely to survive (Howick, Glasziou, & Aronson, 2009, p. 186). Suppose it is recorded that 1 in 10,000 people survived jumping without a parachute while 9,900 in 10,000 people equipped with a parachute stayed alive. The odds ratio of survival with a parachute to survival without a parachute is (9,900 / 100) / (1 / 9,999) = 989,901, the difference being too large to be explained away by confounding factors.[75]

Through the notions of supporting evidence and warranting evidence, the legitimacy of the observational studies on the parachute use as an effective means of death prevention can be justified. Let $H_1$ be the hypothesis that the parachute can prevent death and $H_2$ be the hypothesis that selection bias exists regarding the number of the survivors dropped by parachute. The relevant selection bias here is as follows: 'individuals jumping from aircraft without the help of a parachute are likely to have a high prevalence of pre-existing psychiatric morbidity. Individuals who use parachutes are likely to have less psychiatric morbidity and may also differ in key demographic factors, such as income and cigarette use. It follows, therefore, that the apparent protective effect of parachutes may be merely an example of the "healthy cohort" effect' (Smith & Pell, 2003, p. 1460). Here the 'healthy cohort' effect arises, when a particular group's mortality is lower than that of another group as a result of the pre-existing differences between the groups (e.g. psychiatric morbidity). These differences can cause a different proportion of parachute use in each group. This effect produces a spurious association between parachutes and reduced mortality.

[75] Reiss (2015a, p. 355) holds a similar view: 'Large effects can be a great help to the elimination of alternative explanations because alternatives become intolerably implausible'.

However, the magnitude of the protective effect of parachutes is not attributable to plausible confounders such as selection bias. Suppose that $H_1$ and $H_2$ are the only two hypotheses to be entertained in light of existing background knowledge. Direct evidence here is the difference in the numbers of deaths between those jumping with a parachute and those jumping without. This piece of direct evidence supports $H_1$ and $H_2$ equally, by virtue of being explained by both $H_1$ and $H_2$. Yet, the substantial gap in the numbers of deaths constitutes indirect evidence, indicating a divergence between both, with $H_1$ implying, while $H_2$ being incompatible with (i.e. not being able to explain), this substantial gap (provided that background knowledge indicates selection bias is numerically relatively small). There can surely be other alternative hypotheses containing other confounders, but the observed effect size still overpowers the joint effects of conceivable confounders.

A similar example is general anaesthesia. The observed effect size of general anaesthesia 'has swamped the combined effects of any plausible confounders … [and] are unlikely to be accountable by selection bias, placebo effects or reporting bias' (Howick, Glasziou, & Aronson, 2009, p. 187). The lack of RCTs does not prevent the assurance of the effectiveness of general anaesthetics in terms of rendering patients unconscious.[76] The large effect size counts as evidence in the inferentialist account by virtue of being loosely implied by the hypothesis that general anaesthesia is effective in bringing about unconsciousness.

However, the inferentialist account has its own difficulties and limitations. By difficulties, I mean to what extent an account does not accomplish the objectives it sets out for itself. By limitations, I wish to define the scope of its application by considering where else an account does not perform well. As noted above, one difficulty with the inferentialist account is self-deduction as

[76] Similarly, it is implausible for the effectiveness of certain treatments, such as 'Heimlich manoeuvre, cardiac defibrillation and parachutes to prevent death', to be tested in RCTs, but their effectiveness is confirmed by other sources of evidence (Howick, Glasziou and Aronson, 2009, p. 186).

evidence (Problem B). This difficulty applies to a particular situation: something satisfies the conditions for evidence but does not count as evidence. Another type of difficulty arises because the account does not cover all of the evidence used in the scope it is intended for. With respect to the inferentialist account, its employment of patterns expected in the data, or loose implication, may be too coarse-grained to block alternative low-probability hypotheses, which would otherwise be disregarded at the initial stage of entertaining plausible hypotheses.

To understand the difficulties of the second type, consider a medical vignette in which a knowledgeable clinician, upon observing the bulbar symptoms of her patient, begins to contemplate possible diagnoses: from relatively common neurological disorders to myasthenia gravis (MG), a rare disease whose prevalence in the UK is approximately 15 patients per 100,000 inhabitants (Spillane, Higham and Kullmann, 2012).[77] The clinician finally decides to exclude MG, not in the belief that the MG hypothesis is false, but due to a low pre-test probability, given that the patient has no other known symptoms and no relevant medical history, derived from the very low probability of MG occurrence. There is no pattern in the data incompatible with the MG diagnosis. Probabilities can be helpful and are in fact employed in practice when we think about what candidate hypotheses should be considered and which ones can reasonably be excluded (tentatively).[78]

A riposte might be that contextual features can help in the MG case or other similar sufficiently investigated settings. As in the toothache case where my dentist's background knowledge enables him to avoid considering unlikely hypotheses, contexts can prevent the clinician from taking MG into consideration. Reiss could argue that the low-probability MG hypothesis can still be accommodated within the qualitative framework of the inferentialist

[77] Some philosophers deem this type of reasoning fallacious and call it 'fallacy of probabilistic instantiation' (e.g. Mayo 2010, p. 196). For example, it is fallacious that we infer from the MG prevalence to the probability of a particular individual having MG. But it does not fit in well with medical practice.
[78] Whether this example is classified as hypothesis generation (i.e. coming up with hypotheses) or hypothesis confirmation (i.e. judging which hypothesis is more plausible) does not affect the thrust of the argument.

account, when cost/benefit considerations, belonging to contextual features (see Reiss, 2015b, p. 72), play into the determination of alternatives as relevant.

I acknowledge that this strategy works but in a relatively opaque way. Clinically, the infrequent occurrence of MG, the difficulty of diagnosing MG and the concomitant low benefit-cost ratio together prevent ordering further tests for MG. The exclusion of the possibility of MG in the first place is due to its low prevalence and incidence as demonstrated in empirical data. The cost-effectiveness judgement is built on the very low prior probability of the MG hypothesis, P(MG), and the low posterior probability of the MG hypothesis given a test (e.g. blood test, ice test or imaging), i.e. P(MG/test). Probabilistic information is sometimes needed and cannot be simply replaced by qualitative information in terms of practical considerations, at least in clinically evidential reasoning.

Appealing only to contexts without distinguishing the MG case from the toothache case is unsatisfactory as to how it works. In the toothache case, the unknown-virus hypothesis did not concern my dentist simply because its possibility is unprecedented, while in the MG case, the prevalence and incidence of MG have been recorded in detail and do not depend on its vague possibility. If clinicians do not know such information about MG, they will find it hard to avail themselves of the information about the vague possibility of MG in giving a diagnosis. My advice is that although background knowledge is useful, more detailed information such as probabilities, if any, should be made available to extend the reach of inferences. The difficulty discussed above shows that the inferentialist account is a qualitative one and quantitative information is sometimes useful in evidential inferences.

The MG example may not square with the purpose that the inferentialist account sets out, as this account is proposed to characterise the rationale for establishing *generic* causality, rather than *singular* causality (e.g. the MG

example), in medical and social sciences.[79] Let me first illustrate and discount an off-the-mark counter-example to the inferentialist account. The example comes from a researcher of public health, Per Lytsy (2017), to illustrate that (vague) priors, when empirically well supported, can disregard implausible hypotheses. Suppose that two new drugs, A and B, are separately compared with placebos to test which drug is more effective on disease D. After two large, separate, equal-sized and well-conducted RCTs, according to the frequentist null hypothesis significance test (NHST), a common measure to judge the effectiveness of treatments, their odds ratios are the same, at 1.76 (95% confidence interval, 1.21–2.31, p-value = 0.0036), which indicates that drug A and drug B are both effective for preventing D.[80]

The equilibrium, however, breaks down when previously known information is introduced. Drug A is a 'well-described pharmacological substance with a biologically known mode of action at a receptor level … has a dose-response relationship to some physiological system relevant to disease D … and has a bioavailability seemingly appropriate compared to other similar and effective drugs' and drug B is 'a homeopathic remedy … no studies of the pharmacological properties … no studies support the existence of a biological effect … has been deemed scientifically implausible' (Lytsy, 2017, p. 925).[81] These pieces of information concerning mechanisms are hardly integrated into the NHST framework, in which drug A and drug B cannot be distinguished in terms of their effectiveness. Bayesians, by contrast, can

[79] In fact, Reiss (2015a, 2015b), when illustrating the notions of evidence in his inferentialist account, uses an example of a murder investigation, which is also concerned with singular causation.

[80] The rationale behind NHST is likened to that of the indirect evidence. That data are unlikely under the null hypothesis (i.e. exceeding the pre-set confidence interval) leads to the rejection of the null hypothesis.

[81] The following quotation regarding descriptions about homoeopathy is excerpted from a report by the Science and Technology Committee appointed by the House of Commons (2010, p. 5): 'Homeopathy is a 200-year old system of medicine that seeks to treat patients with highly diluted substances that are administered orally. Homeopathy is based on two principles: "like-cures-like" whereby a substance that causes a symptom is used in diluted form to treat the same symptom in illness and "ultra-dilution" whereby the more dilute a substance the more potent it is (this is aided by a specific method of shaking the solutions, termed "succussion"). It is claimed that homeopathy works by stimulating the body's self-healing mechanisms'.

explain their difference in effectiveness, even given the positive outcome of the NHST, by assigning both a relatively high prior probability to $H_a$ (drug A is effective for preventing D) and an almost zero prior probability to $H_b$ (drug B is effective for preventing D) and by rendering their evidential support discernible by factoring in corresponding posterior probabilities.

It seems that the inferentialist account, without appealing to priors, cannot employ the information about the mechanisms in this way to judge whether drug A or drug B is more evidentially supported or warranted. For this account, the outcomes of both studies provide direct evidence for $H_a$ and $H_b$. Being well conducted, both studies also provide indirect evidence to eliminate some alternative hypotheses such as selection bias. At this stage, it is tempting to claim that both drugs are equally likely to be effective, as the hypotheses about the effectiveness of both drugs are supported by an equal amount of direct and indirect evidence.

Reiss could possibly make an argument that appeals to his notions of evidence, but not to priors, to accommodates the new information concerning mechanisms as follows:

P1. $H_b$ loosely implies that there exist mechanisms responsible for the effectiveness of homoeopathy.
P2. Scientifically, the mechanisms do not exist.
Con. Therefore, P2 is indirect evidence against $H_b$.

However, recall that the foregoing analysis recognises that the reason for the large effect size of anaesthetics being evidence for their effectiveness can be well understood by virtue of the inferentialist account. Notwithstanding their obscure mechanisms, general anaesthetics are clinically effective to a considerable degree (Urban and Bleckwenn, 2002). It seems to impale Reiss on the horns of a dilemma: he has to admit the effectiveness of both anaesthetics and homoeopathy or neither of them.

This dilemma is specious because the lack of mechanisms and the

implausibility of mechanisms are related but distinct in the way that the latter implies the former, but not vice versa. What prevents the effectiveness of homoeopathy from being evidentially warranted is that its mechanisms are biologically and pharmacologically implausible.[82] Unlike homoeopathy, general anaesthetics are still widely used to the extent that their mechanisms remain poorly understood, rather than being deemed scientifically implausible. Returning to the inferentialist account, the implausibility of the mechanisms of homoeopathy is indirect evidence against $H_b$, while the elusiveness of the mechanisms of anaesthetics does not constitute indirect evidence against the hypothesis that anaesthetics are effective. This distinction can be preserved in the inferentialist account as well as in the probabilistic approach.

Up to this point, the off-the-mark counter-example for the inferentialist account has been discounted. However, this account does not cohere well with much of the medical and social sciences, where probabilities can provide more evidential information about generic causation than do loose implication and elimination. Let us consider Bayes factors as a case in point. In the medical sciences and cognitive psychology, Bayes factors are used to evaluate the strength of hypotheses, even when an NHST decides not to reject a hypothesis (Kass and Raftery, 1995; Goodman, 1999). An NHST is a decision, possibly based on a p-value, to reject or not reject a hypothesis. From the Fisherian perspective, if the p-value is smaller than a conventional criterion (e.g. 0.05) and is thereby significant, the null hypothesis $H_0$ is rejected so that the p-value constitutes evidence against $H_0$. When the p-value is greater than or equal to 0.05 and is thereby non-significant, $H_0$ is not rejected; statistically, it is, however, not the case that the p-value offers evidence for $H_0$, because '[its] distribution does not change with increasing sample size when the null is true and the p-value itself is conditioned only on the truth of the null' (Johansson, 2011, p. 115). For the inferentialist account,

[82] That the mechanism by which homoeopathy is purportedly effective is implausible is determined by the biological and pharmacological knowledge of the day. The mechanism, if any, might be explicated if the future scientific knowledge grants its plausibility. Less ideally, this kind of mechanism might not be regarded as impossible, as the science advances, as in the case of general anaesthetics.

the rationale behind NHST proceeds as follows: the claim that $H_0$ is true loosely implies that p-value ≥ 0.05, and the pattern (p-value) in the data is supporting evidence (not equivalent to evidence in the statistical meaning) for the $H_0$. If the claim that p-value < 0.05 is confirmed, then it is indirect evidence for the alternative hypothesis $H_1$ by eliminating $H_0$.

Despite making sense of the p-value, the inferentialist account does not fully capture quantitative evidence use in the biomedical and social sciences. Before examining this account, let me sketch out the Bayes factor. In response to the increasingly wide recognition of the pitfalls of p-values, the American Statistical Association (ASA), recommends alternative methods,[83] including Bayes factors, that can replace or at least supplement p-values, when their assumptions are met (Wasserstein and Lazar, 2016). The Bayes factor (BF) is a quantitative measure of relative evidential strength, or evidential weight, by comparing two competing hypotheses to see how well they fit the data. Regarding hypothesis testing, the Bayes factor can be expressed as: $BF_{10} = P(Data/H_1) / P(Data/H_0)$, where $H_0$ and $H_1$ represent a null and an alternative hypothesis respectively.[84] Statisticians recommend, for example, that if $BF_{10} > 3$, it is at least positive evidence in favour of $H_1$ over $H_0$ and that if $1 \leq BF_{10} \leq 3$, the data are insensitive in the way that they do not provide much evidence to distinguish $H_1$ from $H_0$ (Kass and Raftery, 1995, p. 777). Bayes factors can convey the weight of evidence overlooked by the p-value.

The inferentialist account hardly explains the use of the Bayes factor in the scientific sphere with the notion of supporting evidence. For illustration, consider an example of an ESP experiment, in which a fair coin is tossed to see if a particular subject can predict the outcome. Given $H_0$: P(correct

[83] These alternative methods include 'methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates' (Wasserstein and Lazar, 2016, p. 11).
[84] A likelihood ratio is a special and the simplest form of the Bayes factor as likelihoods are point estimates of the numerator and denominator of the Bayes factor. While the likelihood ratio is believed to reflect the objective strength of evidence, the Bayes factor involves subjectivity as it depends on priors (Bickel, 2012).

guess) = 0.5, $H_1$: P(correct guess) ≠ 0.5 and 220 correct guesses in the 400 tosses, $BF_{10} \cong 7.5$ is obtained and the p-value is slightly greater than 0.05 (Berger and Delampady, 1987, pp. 329–330). If the alternative hypothesis $H_1$ is true, it is expected that $BF_{10} > 3$, so the $BF_{10} \cong 7.5 > 3$ in the study is direct evidence for $H_1$. However, the concept of direct evidence lacks the resources to make a further distinction between positive evidence ($3 < BF_{10} \leq 20$), strong evidence ($20 < BF_{10} \leq 150$) and very strong evidence ($BF_{10} > 150$) (Kass and Raftery, 1995, p. 777). Nor does the concept of indirect evidence capture that of the Bayes factor. Despite evidentially favouring $H_1$ over $H_0$, the $BF_{10} \cong 7.5$ in the study is not incompatible with the null hypothesis $H_0$ in light of the experiment's p-value > 0.05 if p-value < 0.05 indicates the incompatibility between the data and the null; the $BF_{10} \cong 7.5$ is not indirect evidence for $H_1$. To recapitulate briefly, direct evidence does not contribute much to distinguishing degrees of evidential strength, and in the inferentialist account, there is no means other than indirect evidence to distinguish the levels of evidential warrant. What happens in scientific reasoning is that $BF_{10} > 3$ supports $H_1$ over $H_0$ not by eliminating $H_0$. However, medical researchers and cognitive psychologists can still use Bayes factors to distinguish different levels of evidential strength. This is the facet of medical and social sciences practice that the inferentialist account has difficulty explaining.

It is arguable that advocates of the inferentialist account can embrace pluralism on evidential relations to avoid the difficulty of not availing of information provided by probabilities to distinguish the plausibility of hypotheses. Such pluralists would claim that when there are no good reasons to use probabilities, evidential inference proceeds in terms of qualitative relations such as plausibility, and they would accept that when there are good reasons to use probabilities, quantitative numbers can distinguish between the strengths of the evidence. This kind of pluralistic stance is similar to the meta-thesis of pluralism on theories of evidence I endorse in Chapter 1, in which I propose that a theory of evidence should be evaluated against the purpose it is intended to serve. Nevertheless, it is

doubtful whether Reiss will side with pluralism on evidential relations unless he is content with capturing only *many* of the evidential relations in the biomedical and social sciences, and not *all* of them.

Now on to the limitations. A major limitation of the inferentialist account is that it is silent where competing hypotheses are equally evidentially warranted but the evidential strengths can be distinguished via probabilities. If most of the competing salient hypotheses except $H_1$ and $H_2$ are eliminated, both, regardless of how much supporting evidence they obtain, are assessed equally as 'moderately warranted'. However, the information available may reveal the difference in the strength of evidence in favour of them. This is not to suggest that Reiss's qualitative account of evidence does not allow for comparatively evidential claims, such as 'a piece/body of E supports H better than H″ or 'a piece/body of E supports H better than does a piece/body of E″. If E is loosely implied by H but not by H′, then E supports H more than E does H′. If E is loosely implied by H but E′ is not, then E, rather than E′, supports H. These kinds of evidential claims are well-suited to the notion of supporting evidence.

My point here is simply that quantitative information, if of any relevance, can be useful in distinguishing different levels of support for hypotheses, when supporting evidence cannot do so. In some cases, H and H′ share direct evidence E by loosely implying it, and H and H′ are both supported by the same indirect evidence that eliminates some other alternative hypotheses. Here we cannot distinguish the levels of evidential support merely by direct evidence and indirect evidence. Circumstances appropriate for evidential reasoning by quantitative information include probabilistically fully specified settings (e.g. cards and dice). Even so, the inferentialist account remains silent about circumstances in which we are entitled to infer H from E probabilistically.

In addition to epidemiological investigations, we can also lay out different evidential strength in probabilistically fully specified settings. Let us consider the following examples. Suppose E: an ace has been drawn from a standard

deck, $H_1$: The card is the ace of spades, and $H_2$: The card is the ace of hearts, diamonds or clubs. The probabilities are wholly known in the standard deck, where $P(H_1/E) = 1/4$ and $P(H_2/E) = 3/4$. So $P(H_1/E)$ and $P(H_2/E)$ are not equally probable. It is reasonable to conclude that E favours $H_1$ over $H_2$ evidentially, even though $H_1$ and $H_2$ are, in light of the notion of direct evidence, equally plausible and each of them is not eliminated by indirect evidence.

Another example of the insufficiency of the resources for quantitative reasoning is the Monty Hall problem, noted in Chapter 2. This problem is as follows:

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host [Monty Hall] who knows what's behind the doors, opens another door, say No.3, which has a goat. He then says to you, "Do you want to pick door No.2?" Is it to your advantage to switch your choice? (vos Savant, 1990)

In the following argument, the letters $H_i$ and $E_j$ mean the following:

$H_1$: The prize is behind door 1.
$H_2$: The prize is behind door 2.
$H_3$: The prize is behind door 3.
$E_1$: The contestant randomly picks door 1.
$E_3$: The host opens door 3, which is not the location of the car.

The problem emerges because the contestant's initial choice is independent of the place of the car, $P(H_1/E_1) = P(H_2/E_1) = P(H_3/E_1) = 1/3$, but when the host opens door 3, $P(H_1/E_1\&E_3) = 1/3$, $P(H_2/E_1\&E_3) = 2/3$ and $P(H_3/E_1\&E_3) = 0$.[85] In terms of the inferentialist account, the probability of $H_3$ conditional on

[85] The direct calculation with Bayes' theorem proceeds as follows: $P(E_3/H_1\&E_1) = 1/2$ (the prize is behind door 1, so the host can randomly pick and open either door 2 or door 3); $P(E_3/H_2\&E_1) = 1$ (the prize is behind door 2 and the contestant already chooses door 1, so the host must open door 3); $P(E_3/H_3\&E_1) = 0$ (the prize is behind door 3, so it is impossible for the host to open door 3). Therefore, $P(H_2/E_1\&E_3) = P(E_3/H_2\&E_1)P(H_2\&E_1) / P(E_1\&E_3) = [P(E_3/H_2\&E_1)P(H_2\&E_1)] / [P(E_3/H_1\&E_1)P(H_1\&E_1)+ P(E3/H_2\&E_1)P(H_2\&E_1)+ P(E_3/H_3\&E_1)P(H_3\&E_1)] = 1 / (1/2+1+0) = 2/3$.

$E_1$ becomes zero when $E_3$ is further considered on the basis that $E_3$ is indirect evidence for $H_1$ and $H_2$ by eliminating $H_3$. As for the remaining two doors, according to the inferentialist account, $H_1$ and $H_2$ are equally supported by direct evidence $E_1$ in virtue of loosely implying $E_1$, and $H_2$ is further supported by direct evidence $E_3$ because of $P(E_3/H_2\&E_1) = 1$, which is so high as to imply $E_3$ loosely. Since $P(E_3/H_2\&E_1) = 1/2$, $E_3$ is not direct evidence for $H_2$, but it is inferentially consistent with $H_2$. Most importantly, $H_1$ and $H_2$ are equally warranted because the alternative that either hypothesis eliminates ($H_3$) is the same. However, the contestant, on the basis of a new piece of evidence $E_3$, should switch to door 2 to win the car as 'by eliminating door [3] … the probability that door [2] hides the prize is 2 in 3' (Devlin, 2003). It is concluded that probabilities are informative about the strength of evidence in the way that direct evidence cannot disclose this kind of information and that indirect evidence sometimes cannot eliminate relatively-low-probability alternatives.

Another limitation for the inferentialist account is that vague probabilities can contribute to evidential information, the importance of which is understated and even depreciated in the account. Reiss (2015a) does not believe that vague probabilities can contribute to inference about causality, his position on which can be found in his footnote.

> [T]here are no physical probabilities for conditional statements such as "X leaves fingerprints on the murder weapon given X is the murderer" or "I and D are correlated given that I causes D", to assume sharp subjective probabilities is hopelessly unrealistic and misleading, and to assume vague probabilities is to give up most of the advantages of Bayesianism … Possibility and plausibility are the modalities adequate for evidential reasoning, not probability. (Reiss, 2015a, pp. 346–347)

Whether or not vague probabilities can contribute evidential information to generic causal claims is an open question, but they are of use for singular causal claims: evidential reasoning within a range of disciplines regarding the past, including historical linguistics, biology, archaeology, and history. Bayesianism is generally not applicable to these disciplines, since the priors of hypotheses, the likelihoods of evidence given hypotheses, and the

probabilities of evidence tend to be uncertain. Even where theories can provide accurate estimations of the likelihoods, the posterior probabilities are not obtained because the priors are still uncertain. Cases illustrating this type of limitation of Achinstein's, Cartwright's and Reiss's theories of evidence will be discussed at length in Chapter 5.

## 4.6 Conclusion

In closing, I would like to situate Reiss's inferentialist account on the evidential relationship map introduced in Chapter 1 based on OS (objective support), SA (subjective acceptance) and GD (guidance). I conclude that although the inferentialist account delivers an account of SA and GD, it pays little attention to OS. Supporting evidence guides scientists to employ the expectation of patterns in the data, or loose implication, to collect what they regard as relevant facts or data, providing GD. Warranting evidence emphasises the role of judgements in inferences from evidence to hypothesis at the epistemic level, rather than investigating whether evidence or a hypothesis is true or not. This shows that the inferentialist account involves SA and is silent about OS.[86]

I have argued that the inferentialist account has its own difficulties and limitations. From the tool-based view I endorse, a theory of evidence is purpose-specific. To serve the purpose this account sets out, it has to curtail the application scope that it originally attempts to reach in the biomedical and social sciences in accord with the quantitative evidence use that it cannot explain. As regards limitations, this account does not compare hypotheses by weighing their probabilities, even though they are not individually precise. This kind of evidential appraisal, however, is critical in probabilistically fully specified settings and a considerable number of disciplines.

[86] Reiss (2015b, p. 59) calls the inferentialist account of evidence a theory of inferential judgement, which indicates an epistemological approach rather than a metaphysical approach.

The inferentialist account, essentially qualitative, has several strengths. It spells out a typical structure behind scientific reasoning in which scientists compare the truth values of competing hypotheses given evidence at hand and eliminate hypotheses that are incompatible with the evidence. By virtue of the relevance-mediating vehicles of projection and elimination, it is a reform of the H-D model.

# 5
# Comparative Approach:
# The Likelihoodist Account of Evidence

## 5.1 The Nature of Evidence

Evidence mainly serves three functions, namely, explanation of the past, warrant for generic causal relations, and prediction.[87] Where an explanation for past events or some natural or social phenomenon is called for, evidence features in the support for that explanation. For instance, upon investigating a historical or legal case, a historian or lawyer collects evidence, such as overlooked documents or the testimony of an eyewitness, as parts of a jigsaw of pieces to justify their hypotheses or judgements. Evidence can also be intended to offer an empirical ground for the credibility of a hypothesis. This role of evidence is commonplace in scientific experiments and randomised controlled trials (RCTs) with which scientists and social scientists establish causal claims. When it comes to practical use, researchers and policymakers infer, or even bet on, the most likely future result on the basis of the evidence at hand.[88]

Philosophical theories of evidence have been surveyed in this thesis, but they are either silent about, or do not satisfactorily address, the role of evidence in the study of the past. Although they are concerned with evidence to establish generic causal claims, evidence to predict policy outcomes and evidence for singular events, these theories do not fully capture a crucial aspect of evidence use in historical science, i.e. the comparative nature of evidence. In disciplines concerning the past, including anthropology,

[87] This distinction is parallel to the Humean picture of induction, premised on the principle of the uniformity of nature, in which our present observations and memory traces enable us to make inferences about retrodictions, generalisations and predictions (Sober, 1988, p. 44).
[88] The three functions of evidence can be roughly expressed as follows: (a) explanation of the past: E is evidence that 'something happened'; (b) warrant for generic causal relations: E is evidence that 'something happens/can happen'; (c) prediction: E is evidence that 'something will happen'.

archaeology, geology, history and historical linguistics, hypotheses about singular causal claims are rarely confirmed in the rigorous way that scientists confirm hypotheses about generic causal claims.

Scientists normally establish generic causal claims by employing controlled experiments in which shielding conditions are right, such as the blocking or controlling of confounding factors to ensure that a variation in outcome stems solely from the adjustment of interventions or treatments. As discussed in Chapter 3, RCTs take a similar approach: assigning subjects randomly and blindly to balance the mean of net effects of confounding factors. Researchers studying historical events, by contrast, are typically unable to control conditions: instead, they appeal to constraints by theories or inductions, elimination of possibilities and comparison of plausibility.

In this chapter, I shall introduce the likelihoodist account of evidence that, I will argue, can supplement what is wanting in the historical sciences in terms of the exposition of evidential relationships. The setting concerning historiography is far from fully specified and rarely well specified. In such a setting, there are few, if any, prior probabilities or catch-all likelihoods, which Bayesians are obliged to compute or estimate, nor accessible probabilities of explanatory connection required in Achinstein's explanationist theory, nor ideal conditions flagged in Cartwright's argument theory. On the other hand, the setting contains more information about likelihoods than Reiss's inferentialist account can say. The features of the likelihoodist account match well with evidential reasoning in the historical sciences, which reflects the tool-based view about theories of evidence that I endorse from the outset of this thesis: a theory of evidence, if good enough, sheds light on specific kinds of evidential relationships and functions in specific settings in particular fields.

The tool-based view echoes Kyle Stanford's (2011) pluralistic view on evidence, explicitly expressed by the quote below:

> Philosophers of science have long sought the holy grail of *the* logical form of scientific confirmation, whether inductive generalization, the method of hypothesis, conjecture and refutation, Bayes's Theorem, or something else altogether. But I doubt that there is any such holy grail to be discovered. Scientific confirmation is a heterogeneous and many-splendored thing; let us count ourselves lucky to find it—in all its genuine diversity—wherever and whenever we can. (Stanford, 2011, p. 898, italics in original)

In Stanford's view, while standard accounts of evidence, including H-D, abductive and Bayesian accounts, may be successful in explaining some features of evidence, they fail to capture the *projective* feature of evidence, in particular regarding the historical sciences: projection from known cases to unknown cases in the past. To illustrate the latter point, Stanford considers a case of how the hypothesis of organic fossil origins ($H_o$) is well supported by projective evidence ($E_p$) such as water sorting and the mineralisation of organic remains, investigated in both the field and the laboratory. These accounts of evidence, which he terms the *consequentialist* accounts, essentially regard evidence as phenomena explained or predicted by hypotheses by exhibiting those phenomena as 'deductive or probabilistic consequences or implications of those hypotheses' (Stanford, 2011, p. 888). $H_o$ does not explain or predict $E_p$, but $E_p$ is powerful and convincing evidence for $H_o$ in scientific practice, so the consequentialist accounts cannot capture projective evidence.

Stanford's pointing out of the projective nature of evidence aside, in what regards evidence use in the historical sciences, I identify evidence in the form of comparative support. Beyond that, my aim is to formalise comparative evidential reasoning with cases of historical linguistics as a point of departure. This type of reasoning, I believe, can be extended to many other disciplines concerning the past. I will explicate the likelihoodist account as to its strengths, difficulties and limitations, and I will compare this account with the theories of evidence mentioned above. I also distinguish different types of settings in terms of the extent to which the evidential information discloses. Then I focus on the problem with prior probabilities and catch-all likelihoods in historical linguistics. Finally, I will present cases of proto-sound

reconstruction to illuminate how the likelihoodist account can capture the aspect of objective support (OS) in evidential reasoning in historical linguistics.

## 5.2 Objective Support: What Do the Present Data Say?

Before examining particular theories of evidence to see if they cohere with the practice of historical linguists, let us recall the three evidential relationships proposed in Chapter 1: objective support, subjective acceptance and guidance. These can be encapsulated, respectively, in terms of how likely a hypothesis is, given the evidence; how confident we can be of a hypothesis, given the evidence; and how evidence guides us in further investigating hypotheses. These relationships are similar to, but distinct from, those put forward by the statistician Richard Royall (1997, p. 4):

> (1) What do the present data say?
> (2) What should we believe given the present data?
> (3) What should we do given the present data?[89]

Suppose you as a physician receive a positive result of your patient's test for a certain disease D. You know the probability of a positive result if D is actually present is 0.95. You may literally report to the patient the test result with the information about the probability, which relates to (1). Regarding (2), you may diagnose the patient with D when you also know D spreads rapidly and is commonly seen during this season. However, (3) also involves values: you may give the patient a particular treatment when taking its side effects and affordability into account.

[89] The first question is concerned with likelihoodism, which will be explored below. Bayesianism answers the second question. The Neyman–Pearson theory of hypothesis testing provides general rules as to guiding behaviour in the long run, which is pertinent to the third question (Royall, 1997). As can be seen, an answer to the third question demands more than an answer to the second question; an answer to the second question demands more than an answer to the first question.

The three evidential relationships: objective support, concerned with (1); subjective acceptance, concerned with (2); and guidance, relatively unconcerned with (3), are all evidential relationships intended to characterise the concepts of evidence in the historical linguist's mind and to foreground the difference between these concepts and those in the philosophical accounts of evidence. I am not concerned here, or more generally in this thesis, with (3), as it is beyond the scope of this thesis, although it is worthy of investigation. Equipped with the evidential relationships, I will consider a canonical case of how to reconstruct Proto-Romance (spoken Latin), the proto-language of more than thirty modern Romance languages. This case serves both to illustrate how historical linguists entertain hypotheses in the light of evidence, and to set the stage for explaining why I claim that for historical linguistics, the likelihoodist account of evidence is preferable to Achinstein's explanationist theory, Cartwright's argument theory, Reiss's inferentialist account or Bayesianism. Let us start by introducing the likelihoodist account.

## 5.3 The Law of Likelihood

The essential aim of likelihoodists such as A. W. F. Edwards (1972), Richard Royall (1997) and Elliot Sober (2008) is to answer the previous question (1): what do the present data say? Specifically, (a) 'do the data speak in favour of one hypothesis over another?', or (b) 'how should we objectively assess the strength of evidence?' Question (a) is concerned with the *law of likelihood* (LL), identified by that name by Ian Hacking (1965, p. 70). I shall call it the *biconditional* law of likelihood (BLL):

> (BLL) E is evidence in favour of $H_1$ versus $H_2$ if and only if $P(E/H_1) > P(E/H_2)$.

The BLL states both sufficient and necessary conditions for evidence. The likelihood of the hypothesis H given information E is the probability of predicting or observing E under the assumption that H is true, namely

P(E/H). The inequality of the likelihoods of two competing hypotheses reflects the relative evidential import of a body of data for both hypotheses, and vice versa. From the tool-based view, the evidence meaning is not confined to the likelihood inequality, insomuch as there are other accounts of evidence that can make sense where the BLL cannot. The LL is the heart of the likelihoodist account of evidence. I stick to the following definition of evidence, similar to Edward's (1972, p. 30) characterisation of the LL,[90] for the remainder of this chapter:

(LL) E is evidence in favour of $H_1$ versus $H_2$ if $P(E/H_1) > P(E/H_2)$.[91]

The BLL implies the LL, and I will use only the *sufficient* condition for what is evidence favouring a hypothesis over another in the BLL. This does not influence the points I wish to make. Crucially, assuming appropriate settings, the LL, as the core of a relevance-mediating vehicle, can label the information at hand as evidence, without the burden of ruling that there are no other vehicles that can perform the task.

Likelihoodism holds that evidential nature is inherently comparative — given the same data, hypotheses are evaluated against one another and (nearly) never in isolation. Advocates of likelihoodist ideas believe that the LL can capture the comparative nature of evidence by faithfully reflecting whether a hypothesis has greater evidential support than another particular hypothesis. Question (b) is addressed with the likelihood ratio, $L_1(E) / L_2(E) = P(E/H_1) / P(E/H_2) = k$, which is an objective quantitative measure of the relative strength of the evidence. Conventionally, $k = 8$ represents moderately strong

[90] Edward's (1972, p. 30, italics in original) statement on the law of likelihood: 'Within the framework of a statistical model, a particular set of data *supports* one statistical hypothesis better than another if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis'.

[91] There is the *likelihood principle* (LP), which likelihoodists typically embrace. The LP states that 'two instances of statistical evidence are equivalent if and only if they generate the same likelihood function' (Royall, 2004, p. 126). One difference between the LL and the LP is that the former is concerned with 'the bearing of a single data set on two hypotheses' whilst the latter describes 'when two data sets are evidentially equivalent' (Sober, 2008, p. 35, footnote 16). As the LP will not affect the points that I intend to make, I will not pursue this matter in this chapter. For further discussion, see Birnbaum (1962), Steel (2007) and Gandenberger (2015).

evidence in favour of $H_1$ over $H_2$ and $k = 32$ represents strong evidence in favour of $H_1$ over $H_2$. These values are regarded as benchmarks for likelihood ratios suggested by 'the various possible results of one of the simplest of experiments' (Royall, 2004, p. 123). The precise value of $k$ is difficult to calculate where the likelihoods are elusive. Researchers, for example Sober (2008), may settle for the inequality of the likelihoods of two competing hypotheses, whose evidential meaning is still captured by the LL.

Nonetheless, Bayesians believe that they have more to say about evidential relationships. From Bayes' formula it follows that:

$$P(H_1/E) / P(H_2/E) = [L_1(E) / L_2(E)] \times [P(H_1) / P(H_2)].$$

The likelihood ratio is the factor by which the probability ratio is changed from $P(H_1) / P(H_2)$ to $P(H_1/E) / P(H_2/E)$. If the likelihood ratio is $k$, E brings about a $k$-fold increase in the probability ratio, which involves justification. In Royall's words (2004, p. 128), 'Bayesian statistics is primarily concerned with the question of how one's beliefs should change in response to new statistical evidence; that is, its focus is on … "What should I believe?"'. Bayesians pursue that which relates to subjective acceptance (SA).

A direct proof of the LL probably does not exist.[92] Instead of proving the LL directly, the argumentative strategy I employ is to postulate the LL and see how well it can explain the cases of historical linguistics. If it is successful as well as not refuted by the so-called counter-examples, and if there are no counter-examples of historical linguistics, then it is reasonable to claim that the LL is confirmed by the cases. This strategy is comparable to those adopted in most of the theories of evidence I have examined: laying out the definitions and testing them with examples. Having said that, the LL makes

[92] Forster and Sober (2004, p. 154) summarise and analyse Royall's three lines of defence of the LL and conclude that 'none is as strong as one might wish'. Perhaps it is prudent to subscribe to Fisher's (1938, p. 151) and Edward's (1972, p. 100) suggestions that likelihood should be considered a 'primitive postulate'.

intuitive sense in the same way many philosophers embark upon the definitions of evidence with their intuitions.[93]

The LL captures an intuition that if $H_1$ says the present data is more probable than does $H_2$, then the data is evidence in favour of $H_1$ *vis-à-vis* $H_2$. For example, the positive result of a medical test evidentially favours the hypothesis that the patient has a disease over the opposite hypothesis, insofar as the positive result is more probable under the former hypothesis than under the latter. To be specific, the term 'favour' indicates '*differential support*; the evidence points away from the hypothesis that says it is less probable and towards the hypothesis that says it is more probable' (Sober, 2005, p. 128, italics in original). Note that the following reading of the LL is misguided: E supports $H_1$ to a fixed degree, E supports $H_2$ to a fixed degree, and by comparing their absolute values, if $P(E/H_1) > P(E/H_2)$, E supports $H_1$ more than it does $H_2$. On the correct reading, it cannot be said that E is evidence for $H_1$ (to a specific degree) in relation to $H_1$ alone. Although specific values can be assigned to $P(E/H_1)$ and $P(E/H_2)$, it is too hasty to conclude that E supports any of them without considering the relative support for both hypotheses.

Royall (1997, pp. 67–68) illustrates, via the following scenario, that evidential meaning consists in the comparison between $H_1$ and $H_2$ in the light of the same data; evidence is embodied in this kind of three-place relation. Suppose that a person sends his valet to bring his urn and intends to test whether the urn contains 2% white balls ($H_t$). He draws a ball from the urn and finds that it is white ($E_w$). $E_w$ seems to be a piece of evidence against $H_t$. Suppose further that he keeps two urns in his urn vault: one is the original urn and another contains 0.001% white balls ($H_z$). At the moment, $E_w$ is evidence in favour of $H_t$ versus $H_z$, since $E_w$ is rare under $H_t$, but $E_w$ is even rarer under $H_z$. Moreover, in order to see that the LL is sensitive to the selection of rival hypotheses, we can expand Royall's example. Suppose, for

[93] This kind of argumentative strategy parallels John Rawls' (1971) when he tested his theory of justice with our intuitive judgments in particular moral situations.

example, that another person has two urns in his urn vault, one with 2% white balls ($H_t$) and another with 90% white balls ($H_n$). In this instance, $E_w$ (= a white ball drawn) is evidence in favour of $H_n$ versus $H_t$. $P(E_w/H_t)$ is considerably lower than $P(E_w/H_n)$. Two points are made here: firstly, E hardly becomes evidence for a hypothesis unless *a* competing hypothesis is present; secondly, whether E is evidence for the focus hypothesis depends on *the selected* competing hypothesis, however great or small the likelihood of the focus hypothesis.[94]

One might wonder if Bayesianism, as a paradigm of statistics and closely related to the likelihoodist paradigm, fares well in Royall's urn scenario mentioned above. As described in Chapter 2, no matter whether the subjective or objective Bayesian account of evidence requires a precise value for the probability of a hypothesis given the data, P(H/E) is not given in this scenario. By Bayes' theorem, which states that

$$P(H/E) = P(E/H)P(H) / P(E) = P(E/H)P(H) / P(E/H)P(H)+P(E/\sim H)P(\sim H),$$

all we need to know is the likelihood of H given E, namely P(E/H), its catch-all likelihood, namely P(E/~H), and the prior probabilities of hypotheses, namely P(H) and P(~H). The ~H is equivalent to an exhaustive disjunction of H's competing hypotheses $H_1$, $H_2$, $H_3$,…, $H_n$. Although $P(E_w/H_t)$ and $P(E_w/\sim H_t)$ are available, $P(H_t)$ and $P(\sim H_t)$ are up in the air. In obtaining both, the setting of this scenario needs to be specified more precisely. Suppose, for example, that the person has *only* two urns: one contains 2% white balls and the other contains 0.001% white balls, with each of both equally likely to be picked. It follows, by Bayes' theorem, that $P(H_t/E_w)$ = 99.95%, meaning that when he observes that a white ball is drawn, he can almost be certain that the urn under test is the 2%-white-ball urn.

[94] The second reason is why Sober (2008, p. 354) discredited the inference form known as probabilistic *modus tollens*.

There is a seeming counter-example to the LL that should be dismissed. Suppose that there are two decks that look identical on the back, and a card is drawn from one deck. Suppose further $H_s$: the deck is standard; $H_t$: the deck is composed of 52 aces of spades; and E: the card drawn is the ace of spades. $P(E/H_s) = 1/52$ and $P(E/H_t) = 1$. By the LL, E is evidence in favour of $H_t$ over $H_s$, which lacks intuitive appeal. The Bayesian theory can explain this, because the frequency of coming across a trick deck is extremely low for an ordinary setting — $P(H_t)$ is much lower than $P(H_s)$ and thus $P(H_t/E) < P(H_s/E)$. However, Royall (1997, p. 14) maintains that E is strong evidence in favour of $H_t$ over $H_s$ and that this means simply that '[E] is not strong enough to overcome the prior improbability of [$H_t$] relative to [$H_s$]'. In terms of the unsuccessful counter-example, some clarification has been made.

Leeds (2000) provides a counter-example to the LL. Suppose that an ace has been drawn from a standard deck ($E_a$). Two hypotheses are presented: $H_h$: the card is the ace of hearts; and $H_{sc}$: the card is the ace of spades or the ace of clubs. It follows that $P(E_a/H_h) = P(E_a/H_{sc}) = 1$, which, by the LL, means $E_a$ favours neither $H_h$ nor $H_{sc}$. But it would be unsound to say that, particularly when we consider the following values: $P(H_h/E_a) = 1/4$ and $P(H_{sc}/E_a) = 1/2$. From a Bayesian perspective, also highly intuitively, since $P(H_{sc}/E_a) > P(H_h/E_a)$ and, less obviously, $P(H_{sc}/E_a) > P(H_{sc})$, it is indicated that $E_a$ evidentially favours $H_{sc}$ over $H_h$.[95] In this case, the posteriors, rather than the likelihoods, capture our intuitions.

Likelihoodists could defend the LL in the same way Royall objects to the trick deck example. For from the likelihoodists' point of view, Leeds' counter-example is flawed: favouring relations of evidence are not mediated by prior

[95] Bayesians would not accept a naïve principle like this: E favours $H_1$ over $H_2$ if and only if $P(H_1/E) > P(H_2/E)$. Such a principle would incur an unfavourable situation: if $P(H_1/E) > P(H_2/E)$, we should claim that E favours $H_1$ over $H_2$, even if in a Bayesian sense, E disconfirms $H_1$, i.e. $P(H_1/E) < P(H_1)$, and E confirms $H_2$, i.e. $P(H_2/E) > P(H_2)$. To illustrate, suppose $H_1$: the card is any except for the ace of clubs, $H_2$: the card is the ace of clubs, and E: the card randomly drawn from a standard deck is a club. It follows that $P(H_1/E) = 12/13 > P(H_2/E) = 1/13$, $P(H_1/E) < P(H_1) = 51/52$, and $P(H_2/E) > P(H_2) = 1/52$. Intuitively, we are inclined to say that E is evidence for $H_2$ rather than $H_1$, which shows that the naïve principle is problematic. Fitelson (2007, p. 480) reminds us that for Bayesians, the concept of degree of probability and the concept of confirmation should be carefully distinguished.

probabilities, i.e. $P(H_h)$ and $P(H_{sc})$ here.[96] Likelihoodists may claim that because $E_a$ equally supports $H_h$ and $H_{sc}$, i.e. $P(E_a/H_h) = P(E_a/H_{sc}) = 1$, the difference in their posterior probabilities must result from that in their priors; Leeds' example does not undermine the LL.

Another attempt to put forward a counter-example to the LL is by Branden Fitelson (2007, pp. 476–477). Suppose that $E_s$: the card drawn from a standard deck is a spade; $H_s$: the card is the ace of spades; and $H_b$: the card is black. As such, it follows that $P(E_s/H_s) = 1 > P(E_s/H_b) = 1/2$ and by the LL, $E_s$ is evidence in favour of $H_s$ versus $H_b$. However, it is not intuitively appealing. To see why, it is helpful to use Fitelson's (2007, p. 477) highly intuitive principle which involves no priors and should be accepted by likelihoodists:

> (*) If E provides conclusive evidence for $H_1$, but non-conclusive evidence for $H_2$, then E is evidence in favour of $H_1$ over $H_2$, with E, $H_1$ and $H_2$ being empirical claims.

Evidently, $E_s$ guarantees the truth of $H_b$ whilst $E_s$ offers merely non-conclusive evidence for $H_s$. This is claimed to be a counter-example to the LL: $P(E_s/H_b) < P(E_s/H_s)$, but $E_s$ is evidence in favour of $H_b$ versus $H_s$.

However, Fitelson's counter-example can be blocked by imposing a constraint on the LL: $H_1$ and $H_2$ should be mutually exclusive. Insofar as $H_s$ and $H_b$ are not incompatible, the LL cannot be applied to this case. From the tool-based viewpoint, when a theory of evidence encounters difficulties in making sense of the cases that it sets out to explain, it is a natural move either to enhance its ability to explain or restrict its scope of application. Although most of the representative likelihoodists (Hacking, 1965, p. 63; Edwards, 1972, p. 30; Royall, 1997, p. 3) do not place the constraint of

[96] In fact, Leeds' example is not restricted to being represented by posteriors involving priors; instead, likelihoods and catch-all likelihoods suffice to represent his example (Fitelson, 2007). Even if this is a correct representation, I will argue in Section 5.8 that catch-alls, if any, are rare and inevitably involve priors in historical linguistics.

mutual exclusiveness on the definitions of the LL,[97] this constraint enables the LL to avoid Fitelson's counter-example.[98] An ultimate means of resolving Leeds' and Fitelson's counter-examples is to filter fully specified and well-specified evidential settings out of the scope of the LL. That is, where the prior probabilities are well supported, the LL is unhelpful. I will visit this issue in Section 5.7. However, comparing composite hypotheses without specifying the prior probabilities is not uncommon in science. This calls for other methods regarding model selection, including the Akaike information criterion (AIC) (Akaike, 1973) and the Bayesian information criterion (BIC) (Schwarz, 1978). I will ignore the discussion about these methods here, since they are not the main concern of this thesis.

## 5.4 Characteristics of Theories of Evidence

I have scrutinised Achinstein's explanationist theory, Cartwright's argument theory and Reiss's inferentialist account in the previous chapters. Now I compare them, in addition to Bayesianism, with the likelihoodist account to show the fundamental distinctions between these theories of evidence. I commence with an example discussed in Chapter 2, before presenting the relative strength of the likelihoodist account as far as evidential reasoning in historical linguistics, in particular regarding the reconstruction of proto-sounds.

Suppose that a subject undergoes a test which can detect whether one has a certain disease D. Through the procedure of *maximum likelihood estimation*, some relevant probabilities are obtained from the frequency data (Sober, 2004, p. 225). If the subject has the disease ($H_d$), the probability that the test yields a positive result ($E_t$) is 95%, namely $P(E_t/H_d) = 95\%$. If the subject does not have D ($\sim H_d$), the probability that the test yields a positive result is 1%, namely $P(E_t/\sim H_d) = 1\%$. The prevalence of D in a target population is 1/100,000, which is objectively estimated by random sampling;

[97] Sober (2008, p. 34) seemed aware of the constraint of mutual exclusiveness when he provided examples in which two hypotheses are incompatible.
[98] This constraint on the LL can be seen in Steel (2007) and Chandler (2013).

the prior probability $P(H_d) = 1/100,000$. The posterior probability $P(H_d/E_t)$ then equals approximately 0.095%. In this well-specified example, $\sim H_d$ exhausts the competing hypothesis.

Now let us discuss the theories of evidence in sequence to see how they would account for the diagnosis example. Since $P(E_t/H_d) > P(E_t/\sim H_d)$, likelihoodists would consider $E_t$ as evidence for $H_d$ versus $\sim H_d$. This is similar to what occurs in the trick deck example mentioned above. Royall (1997, p. 14) reminds us not to confuse evidence and belief; the former is concerned with 'what the data say' and the latter with 'what we should believe given the data'. Having said that, I will later argue that priors, if objectively available, are useful for evidential reasoning. Priors as objective as $P(H_d)$ are, however, rarely seen in the historical sciences. When the information about prior or posterior probabilities is unavailable, or in Hacking's phrase when 'lacking other information' (1965, p. 65), the LL fares relatively well.

A second theory is Achinstein's (2001, 2010a, 2014) explanationist theory. Recall that the explanationist theory states that E is evidence for H given B if and only if (a) P(Exp(H,E)/E&B) > 1/2; (b) E and B are true; and (c) E does not entail H, where H is a hypothesis, B is background knowledge and Exp(H,E) means that there is an explanatory connection between H and E. That is, H correctly explains the truth of E, or E correctly explains the truth of H, or some hypothesis correctly explains both the truth of E and the truth of H. As mentioned in Chapter 2, the probability of there being an explanatory connection between H and E given the truth of E and B can be decomposed into (1) the probability of there being an explanatory connection between H and E given the truth of H, E and B, and (2) the posterior probability, whose values may rest on priors and catch-all likelihoods:

$$
\begin{aligned}
& P(Exp(H,E)/E\&B) \\
&= P(Exp(H,E)/H\&E\&B) \times P(H/E\&B)^{99} \\
&= P(Exp(H,E)/H\&E\&B) \times [P(H/B)P(E/H\&B)] / [P(H/B)P(E/H\&B) + \\
& \quad P(\sim H)P(E/\sim H\&B)].
\end{aligned}
$$

In illustrating the explanationist theory, let us consider Achinstein's (2001, p. 154) own example. Suppose 70% of patients with particular symptoms (S) take medicine (M) and also get relief (R) in seven days, P(R/S&M) = 0.7. Among patients with S who take M and get relief in seven days, 60% of them get relief because of M, P(Exp(R,M)/R&S&M) = 0.6 (40% of them because of other reasons). By Achinstein's definition of correct explanation, P(Exp(R,M)/S&M) = P(Exp(R,M)/R&S&M) × P(R/S&M) = 0.6 × 0.7 = 0.42. This means that among patients with S who take M, 42% of them get relief because of M; only 42% of the time is M effective in relieving S.

Let us return to the diagnosis example, where the prior $P(H_d)$, the catch-all $P(E_t/\sim H_d)$, and the likelihood $P(E_t/H_d)$ are available, all of which make Bayes' theorem work in exchange for the posterior probability $P(H_d/E_t)$. Clearly, for the explanationist theory, $E_t$ is not evidence for $H_d$, because $P(H_d/E_t\&B)$ is too low to boost $P(Exp(H_d,E_t)/E_t\&B) > 1/2$.[100] A follow-up question is whether $E_t$ is evidence for $\sim H_d$, which amounts to whether $P(Exp(\sim H_d,E_t)/E_t\&B) > 1/2$. Although $P(\sim H_d/E_t\&B) = 99.905\%$ is exceedingly high, $P(Exp(\sim H_d,E_t)/\sim H_d\&E_t\&B)$ has to be at least greater than 1/2 in order that $P(Exp(\sim H_d,E_t)/E_t\&B) > 1/2$.[101] The problem is that the information about the probability of the explanatory strength is elusive.

To the best of my knowledge, Achinstein does not provide details about how to calculate the probability of explanatory connection. In his medical example, he simply *assumes* figures to illustrate his theory. But figures as to

---

[99] For the proof, see Chapter 2, Section 2.5.

[100] Philosophical niceties aside, $P(H_d/E_t\&B) = P(H_d/E_t) = 0.095\%$ in this example.

[101] $P(Exp(\sim H_d,E_t)/\sim H_d\&E_t\&B) > 1/2$ is entailed with the following proof: given that $p = qr$, $p > 1/2$, $0 \leq q$, $r \leq 1$, we obtain $q = p/r \geq p > 1/2$. By substituting $P(Exp(\sim H_d,E_t)/E_t\&B)$ for $p$, $P(Exp(\sim H_d,E_t)/\sim H_d\&E_t\&B)$ for $q$, and $P(\sim H_d/E_t\&B)$ for $r$, we obtain $P(Exp(\sim H_d,E_t)/\sim H_d\&E_t\&B) > 1/2$.

the strength of explanatory connection are not attached to empirical data and are hard to compute. In some special cases, where the explanatory connection has only one possibility, the probability of the explanatory connection is straightforward: unity. For illustration, consider the birthday problem first created by von Mises (1939): if there are *n* people in a room, what is the probability of at least two people sharing a birthday? If there are 366 people in the room ($E_{bp}$), we can guarantee that there must be a pair having the same birthday ($H_{bp}$) on the basis that there are 365 days in a non-leap year ($B_{bp}$). For the explanationist theory, $E_{bp}$ is evidence for $H_{bp}$ given $B_{bp}$, because $P(H_{bp}/E_{bp}\&B_{bp}) = 1$ and $P(Exp(H_{bp},E_{bp})/H_{bp}\&E_{bp}\&B_{bp}) = 1$ (i.e. $E_{bp}$ fully explains $H_{bp}$ given $H_{bp}$, $E_{bp}$ and $B_{bp}$).

For $E_t$ to count as evidence for $\sim H_d$ in the diagnosis example, the explanationist theory must provide a way to compute the probability that there is an explanatory connection between $\sim H_d$ and $E_t$ given $\sim H_d$, $E_t$ and B, particularly when we cannot directly see it. What are required more for the explanationist theory than for the likelihoodist account are probabilities of explanatory connection and posterior probabilities containing priors and catch-all likelihoods.

A third theory of evidence under discussion is Cartwright's (2013) argument theory. This theory holds that E is evidence for a hypothesis H relative to an argument A if and only if E is a necessary premise in A, which is a valid and sound argument for H. For Cartwright, a valid and sound argument ties some empirical claims together, in which the claims in the form of the premises and the claim appearing as the conclusion are evidence and hypothesis respectively.

In the diagnosis example, an advocate of the argument theory may put forward an argument as follows: (1) frequency data $F_1$ shows that 95% of the subjects with disease D and 1% of the subjects without D have a positive test result; (2) if $F_1$ is obtained by the procedure of *maximum likelihood estimation*, $F_1$ can be translated into the likelihoods $P(E_t/H_d)$ and $P(E_t/\sim H_d)$; (3) frequency data $F_2$ shows that 1/100,000 of the target population have D;

(4) if $F_2$ is obtained by random sampling, $F_2$ can represent the prior probability $P(H_d)$; (5) if the posterior probability is very low, a subject with a positive result is unlikely to have D. It deductively follows from (1)–(5) that the subject with the positive result is very unlikely to have D.[102] Although this argument does not deductively entail $\sim H_d$ (the subject does not have D) and thus none of the premises is evidence for $\sim H_d$, all of them are evidence for the hypothesis that the subject having the positive result is 99.905% unlikely to have D. In order to guarantee the truth of the conclusion, it is necessary that all of the ideal conditions in the valid argument are met, which demands much more than the likelihoodist account.

Next, let's move on to Reiss's (2015a, 2015b) inferentialist account. Reiss has two main notions of evidence: *supporting evidence* and *warranting evidence*. Supporting evidence consists of two kinds of evidence: direct evidence and indirect evidence.

> **Direct evidence:** $E_d$ is direct evidence for a hypothesis H if and only if $E_d$ is a pattern in the data that one is entitled to expect, supposing that H is true.

> **Indirect evidence:** $E_i$ is indirect evidence for a hypothesis H if and only if $E_i$ is a pattern in the data that is incompatible with what one is entitled to expect, supposing that one of H's competing hypotheses H′, H″, H‴ and so on is true.

Depending upon the extent to which legitimate alternatives are eliminated, four levels of warranting evidence are distinguished: proof, strong warrant, moderate warrant and weak warrant.

---

[102] In this argument, I do not consider the so-called *fallacy of probabilistic instantiation* (Mayo, 2005), discussed in Chapter 2. Put briefly, from the frequentist perspective, $P(H_d)$ is, in fact, concerned with *some* subject randomly selected from the population, viewed as a *generic* type of event, or a random variable. There is no such thing as assigning probability to a *unique* event, i.e. *the* subject's having D. If adopting the frequentist stance, we can simply add a premise such as $P(H_d)$ = P(the subject has D) and we will have a valid argument.

| | |
|---|---|
| **Proof:** | All relevant competing hypotheses are eliminated. |
| **Strong warrant:** | All salient competing hypotheses and some non-salient are eliminated. |
| **Moderate warrant:** | Most competing hypotheses, some of which are salient, are eliminated. |
| **Weak warrant:** | Some competing hypotheses, none of which are salient, are eliminated. |

Salient competing hypotheses are those supported by additional direct evidence beyond the direct evidence shared between the hypothesis of interest and its competing hypotheses. Non-salient competing hypotheses, by contrast, are those lacking additional direct evidence.

Applying the notions of evidence to the diagnosis example, we can see that $E_t$ is direct evidence for $H_d$ by virtue of being expected, supposing $H_d$ is true, and $E_t$ is inferentially incompatible with $\sim H_d$ and thus is indirect evidence for $H_d$. $E_t$ lends a certain amount of warrant to $H_d$ by eliminating $\sim H_d$. However, the information about $P(H_d)$ supports the opposite. Suppose additional background knowledge is available: D is an infectious disease with high transmissibility and the subject often goes out in public without taking any necessary precautions. If D is highly transmissible to humans and the subject does not have D, it is likely that D has a low prevalence. The low prevalence of D is expected supposing $\sim H_d$ is true and is at best inferentially compatible, if not incompatible, with $H_d$. $P(H_d)$ is direct evidence for $\sim H_d$, not for $H_d$, but, as mentioned, $H_d$ is warranted because $E_t$ eliminates $\sim H_d$.

It does not seem reasonable to claim that $H_d$, rather than $\sim H_d$, is warranted, when the subject, though testing positive for D, is 99.905% unlikely to have D. While the argument theory and the Bayesian theory can utilise all of the quantitative information appearing in the example, the inferentialist account cannot fully digest these pieces of information. Note that $P(E_t/H_d)$ and $P(E_t/\sim H_d)$ are extreme in this example, but if the values are changed, say $P(E_t/H_d) = 60\%$ and $P(E_t/\sim H_d) = 40\%$, the inferentialist account hardly

employs the information about the likelihoods because the new information is neither expected from nor incompatible with $H_d$ or $\sim H_d$; the new information is inferentially compatible with each of the hypotheses. The inferentialist account filters out only coarse-grained information but gives up the useful quantitative information about likelihoods.

Finally, for Bayesians, either $E_t$ is evidence for $H_d$ as long as $H_d$ is more probable given $E_t$ than without $E_t$, namely $P(H_d/E_t) > P(H_d/\sim E_t)$, or $E_t$ is evidence for $\sim H_d$ as $P(\sim H_d/E_t)$ is sufficiently high, for example at least greater than half.[103] To compute the posterior probabilities requires the prior probabilities and the catch-all likelihoods. Fortunately, these probabilities are available in this example. Sober (2004, pp. 231–232) doubts it is normal for scientific tests, by claiming that 'Galileo was in no position to assign an objective prior probability to [Jupiter's having moons]' and that '[i]f Newton's theory is false, what is the probability of each of the theory's specific alternatives?'. Cases like the diagnosis example, in which the probabilities needed are all specified, are rare in science, especially in the historical sciences.

## 5.5 The Comparative Method

Having spoken of the characteristics of the theories of evidence, I shall introduce the methodology of proto-sound reconstruction before using cases of historical linguistics to reveal their limitations.

When languages are genetically related, they are descended from the *proto-language*, which is 'the once-spoken ancestral language from which daughter languages descend, and, in another sense, the language reconstructed by the comparative method that represents the ancestral language from which the compared languages descend' (Campbell & Mixco, 2007, p. 158). The descendent languages are dialects of the proto-language

[103] Here I do not consider the merged objective Bayesian account (MOB), as discussed in Chapter 2.

at some point in history prior to becoming different languages, distinguishable by innovative traits. To illustrate the reconstruction of a proto-language, for simplicity's sake, only four of the most widely spoken Romance languages (namely, Spanish, Portuguese, French and Italian) have been selected. Given that these four languages belong to the Romance language family, the aim now is to reconstruct Proto-Romance, from which the Romance languages branched off.

Proto-Romance can be, and should be, approximately equated with 'the spoken language at the time when Latin began to diversify and split up into its descendant branches, essentially the same as Vulgar Latin at the time' (Campbell, 2013, p.108). Historically, Vulgar Latin, not equivalent to Classical Latin, was a language which was 'used by the illiterate majority of the Latin-speaking population' (Coleman, 1993, p. 2) and which split to create the Romance languages. By contrast, Proto-Romance is a hypothetical vernacular proto-language of the Romance languages and may correspond to an actually spoken dialect of Vulgar Latin. Proto-Romance 'has no precise location in time, being situated only in relative time, at some point to the earliest divergence required by the diachrony of reconstruction' (Coleman, 1993, p. 2). Ideally, the reconstructed proto-language, derivable from its daughter languages, coincides with what must have been an actual language, i.e. the once-spoken proto-language. In practical terms, Proto-Romance may not be perfectly reconstructed. It is nevertheless intended to match (a dialect of) Vulgar Latin.[104]

When historical linguists follow genealogical inference, three main types of questions are distinguished by Christopher Hitchcock (1998, pp. 430–431):

Q1: Do language A and language B share a common ancestral language?
Q2: Given that languages A, B and C share a common ancestral

[104] For a detailed account of the reconstruction of Proto-Romance and the relationship between Proto-Romance, Vulgar Latin and Classical Latin, see Hall (1950).

language, do A and B share an ancestral language that is not shared with C?

Q3: Given that A and B share a common ancestral language, what features does the ancestral language have?

Characteristic of the historical sciences is genealogical inference. This type of inference has come to the forefront of evolutionary biology, text criticism and historical linguistics, which are all 'stemmatic or comparative disciplines, whose task is largely the creation and manipulation of genealogical trees' (Lass, 1997, p. 113). Genealogical inference purports to determine the family tree: specifically, which family members are genetically most closely related to one other.

The *comparative method* specifically addresses Q3. In circumstances where languages have no written form, the comparative method is regarded as the gold standard in historical linguistics for reconstructing ancestral languages preceding the advent of writing (Donohue, Denham and Oppenheimer, 2012; Dunn, 2015; Hale, 2015; Weiss, 2015).[105] This method 'usually begins with phonology, with an attempt to reconstruct the sound system; this leads in turn to reconstruction of the vocabulary and grammar of the proto-language' (Campbell, 2013, p. 107). In short, the comparative method is primarily to address a question: if the observed languages are genetically related, what would their proto-language be like?

The comparative method is not, however, responsible for generating a hypothesis that the related languages originate from a single proto-language, or for confirming such a pre-existing hypothesis beforehand. That is, prior to the application of the comparative method, the genetic relatedness of the languages (e.g. Spanish, Portuguese, French and Italian) should be hypothesised and confirmed. Nevertheless, the comparative method can

[105] The comparative method remains the gold standard, insomuch as '[t]he persistent superstition that it does not work for unwritten languages or for certain families has been refuted again and again — famously for Algonquian in the classic work of Bloomfield (1925, 1946), continued by Hass (1958), Goddard (1979, 1990), Garrett (2004), Berman (2006), and others' (Kiparsky, 2015, p. 65).

later be used to examine the initially generated hypothesis by checking whether the sound–meaning pairings between the given languages are regular (Weiss, 2015, p. 128). If regularity of correspondences emerges, this is considered evidence to confirm a relatedness hypothesis that is previously generated because of the geographic proximity, shared history and textual records or otherwise of the languages. Contrarily, lack of regular correspondences is considered evidence to disconfirm the relatedness hypothesis.

Before making explicit the operation of the comparative method, I shall introduce linguistic notation. A hyphen after a sound indicates that the sound is in an initial position, as *k*-; a hyphen on either side indicates a medial sound, as -*k*-; a hyphen before the sound indicates a final sound, as -*k*. A statement of the form '*x* > *y*' or '*y* < *x*' is read as 'sound *x* changes into sound *y*', as the case of *k* > *ʃ*. An asterisk indicates reconstructed forms, as *k.

Upon reconstructing a proto-language, historical linguists take three basic steps, usually back and forth for revisions (Campbell, 2013, p. 111–128):

*Step 1*. Assemble cognate lists.
*Step 2*. Demonstrate sound correspondences.
*Step 3*. Reconstruct proto-sounds.
*Step 4*. Check if the reconstructed proto-sounds are plausible in phonology and typology.
*Step 5*. Reconstruct morphemes.

*Step 1*: Historical linguists generally begin by compiling a list of cognates from *basic vocabulary*, including common body parts, close kinship terms, low numbers, and frequently experienced aspects of the world (Campbell & Mixco, 2007, p. 25). The words with a (nearly) identical meaning are supposed to be cognates of a common ancestral word, i.e. the descendant varieties of a proto-word. Whether putative cognates are genuine cognates is not demonstrable until they show systematic correspondences (Campbell,

2013, p. 111). Putative cognates, when they are placed together, constitute a (potential) *cognate set*.

*Step 2*: Historical linguists seek systematic sound correspondences in the cognate sets. The sounds in the same cognate set are to identify their corresponding positions. An example makes this plain. Some provisional cognate sets are as follows (Campbell, 2013, p. 110):

| | *Italian* | *Spanish* | *Portuguese* | *French* | *(Latin)* | *English gloss* |
|---|---|---|---|---|---|---|
| 1 | capra /kapra/ | cabra /kabra/ | cabra /kabra/ | chèvre /ʃɛvr(ə)/ | capra | 'goat' |
| 2 | caro /karo/ | caro /karo/ | caro /karu/ | cher /ʃɛr/ | caru | 'dear' |

The words for the same meaning, 'goat', are placed in the same row. In cognate set 1, it can be seen that the first sound correspondence is established as: *k*- : *k*- : *k*- : *k*-. The *k* is called a *reflex* of the original sound of the proto-language. The second sound correspondence is -*a*- : -*a*- : -*a*- : -*ɛ*-. For the words of the same meaning, their sounds are paired according to the positions of the sounds. The more frequently the same correspondence appears in different cognate sets, the more confidence historical linguists have in the genuineness of the cognates. For example, there is the same sound correspondence *k*- : *k*- : *k*- : *k*- in cognate set 2, so the cognates here are more likely to be genuine because they appear in two sets rather than in only one set. The procedure is repeated until all the sound correspondences in the languages being compared are identified.

*Step 3*: Once correspondence sets (but not necessarily all) are established, general principles are proposed to determine which proto-sound is plausible relative to others. If sounds in each correspondence set appeared unanimously, it would be straightforward to reconstruct corresponding proto-

sounds according to the related sounds in agreement. For instance, a proto-sound **t* is naturally reconstructed in view of the sound correspondence (*t*- ; *t*- : *t*-) in the Finno-Ugric languages of Finnish, Hungarian and Udmurt (Campbell, 2013, p. 131). However, the linguistic reality is complex. The purity of correspondence sets can be affected by conditioned sound changes. For example, correspondence sets may be partially overlapping: if two correspondence sets (*k*- : *k*-) and (*k*- : *ʃ*-) are designated, general principles and relevant linguistic information are needed to inform us how to reconstruct the proto-sound(s).

*Step 4*: Given the reconstructed proto-sounds, historical linguists check if they accord with phonology and typology. Among most of the observed languages, the inventories of sounds are phonologically symmetrical with congruent patterns and have a typological tendency for particular sounds (Campbell, 2013, pp. 124–126). If the reconstructed sound does not agree with the general patterns, it will be rendered implausible unless there is strong evidence in support of it.

*Step 5*: Historical linguists use proto-sounds to reconstruct words in the proto-language. For example, **kapra* ‘goat’ is reconstructed for Proto-Romance on the basis of the sounds provided by its daughter languages.

From Step 1 to Step 5, through comparing the cognate sets of its daughter languages, the proto-language is derivable.

### 5.6 Limitations of the Theories of Evidence

Below I simplify and sketch the reconstruction inference described by the historical linguist Lyle Campbell (2013, pp. 113–118). Let us stay with the Proto-Romance family of languages, and consider only the words for ‘goat’, for which the counterpart has been reconstructed as ‘capra’. For ease of discussion, I repeat the cognate set below:

| *Italian* | *Spanish* | *Portuguese* | *French* |
|---|---|---|---|
| capra<br>/kapra/ | cabra<br>/kabra/ | cabra<br>/kabra/ | chèvre<br>/ʃɛvr(ə)/ |

The present task is to reconstruct the proto-sound for the correspondence between the first sounds. We have the following information:

$E_1$: The correspondence set for the first sounds is *k*- : *k*- : *k*- : *ʃ*-.

Suppose the hypotheses entertained are as follows:

$H_1$: The proto-sound is **k.*
$H_2$: The proto-sound is **ʃ*.

Below I introduce several general principles for the reconstruction of proto-sounds. The first principle is *directionality*, which says that 'some sound changes which recur in independent languages typically go in one direction (A > B) but usually are not (sometimes are never) found in the other direction (B > A)' (Campbell, 2013, p. 113). This principle can also be called *process naturalness* (Lass, 1997, p. 137). For example, the data show that the direction of sound change is generally from *k* to *ʃ*, and the converse direction generally never happens. In other words, the former direction is more 'natural' than the latter.

The second general principle, *majority wins*, is that everything else being equal, the sound shared between the majority of sister languages is the original sound in their proto-language. In the Proto-Romance case, the correspondence between the first sounds is mostly *k* except French, which is *ʃ*.

The third general principle, *economy*, is that everything else being equal, among alternative sounds, the sound that supposes fewer steps of change is the proto-sound. In the case discussed here, if we were to postulate that the

first sound in the Proto-Romance is *ʃ, the sound would undergo three independent changes from *ʃ to *k*, each for Italian, Spanish and Portuguese. This is more than the one change from **k* to *ʃ* by simply postulating **k* for the Proto-Romance. The second and third principles can be combined and called *simplicity*. That is, the single-source hypothesis is more parsimonious (in that it deploys fewer sound changes) and therefore more evidentially preferable than the multiple-convergence hypothesis, with the exception of 'natural' traits that are widely shared between languages (e.g. prenasal nasalization) (Lass, 1997, p. 137).

Now let us examine the theories of evidence in sequence to see how well they fit in with the foregoing case, in which historical linguists generally regard the correspondence set *k*- : *k*- : *k*- : *ʃ*- ($E_1$) as evidence for the proto-sound **k* in Proto-Romance ($H_1$).

The first theory is Achinstein's (2001, 2010a, 2014) explanationist theory. Suppose that there is a historical linguist who knows that $E_1$ and some relevant background knowledge B (e.g. the general principles above) are true. In order to determine whether $E_1$ is evidence for $H_1$, what the linguist requires, according to the explanationist theory, depends solely upon whether $P(Exp(H_1,E_1)/E_1\&B) > 1/2$. However, how to compute the probability of explanatory connection, $P(Exp(H_1,E_1)/H_1\&E_1\&B)$, is puzzling. So is the posterior probability, $P(H_1/E_1\&B)$, since direct and indirect ways of obtaining this value are lacking. Unlike fully specified settings (e.g. a standard deck of cards), there is no direct way to know the posterior probability in historical linguistics, or even in other historical sciences. Nor is there any indirect way. The prior probability, $P(H_1)$, is either undesirably subjective if subjective Bayesianism is adopted, or dubiously objective if objective Bayesianism is adopted, and the catch-all likelihood $P(E_1/\sim H_1\&B)$ must depend upon the full identification of all possible competing hypotheses, all of which are implausible in the historical sciences. The distinction between fully/well-specified and poorly specified settings is made in Section 5.7, and the problems about priors and catch-all likelihoods are explored in Section 5.8. I shall conclude that Achinstein's explanationist theory of evidence does not

make sense either of the case of reconstructing proto-sounds in Proto-Romance, or of evidential reasoning in historical linguistics.

The second theory of evidence to be examined here is Cartwright's (2013) argument theory. A valid argument in favour of $H_1$ can be constructed as follows:[106]

P1. $E_1$: The first sound correspondence is Italian *k*, Spanish *k*, Portuguese *k*, French *ʃ*.
P2. If the principles of directionality, majority wins and economy all converge on the same sound, and there were no non-inheritance factors (e.g. diffusion and chance) that brought about the sound correspondence, the reconstructed sound is the proto-sound.
P3. The principles of directionality, majority wins and economy all converge on *k.
P4. There were no non-inheritance factors.
Con. $H_1$: The sound in Proto-Romance is **k*.

Ideally, if P1–P4 are met, $H_1$ is deductively entailed; according to the argument theory, P1–P4 are evidence for $H_1$. However, as opposed to scientific experiments and RCTs, in which confounders are well controlled and ideal conditions may be confirmed, historical linguists have meagre resources to ensure P4. For example, sounds may resemble one another owing to language contact, or languages may develop the same or similar traits independently and coincidentally. Although historical linguists can claim that inheritance explains $E_1$ better than diffusion or chance, it is difficult for them to exclude these possibilities completely. The argument theory is not well suited to the practice of historical linguistics, particularly regarding the proto-sound reconstruction.

[106] For simplicity's sake, I do not take into account partially overlapping correspondence sets and phonological and typological considerations.

Turn now to Reiss's (2015a, 2015b) inferentialist account. On the supposition of the truth of $H_1$, $E_1$ is, according to the principles of majority wins and economy, a pattern in the data that is expected, because in the sound correspondence, *k* outnumbers *ʃ* and there are fewer changes if **k* > *ʃ*. Thus, $E_1$ is direct evidence for $H_1$. Although $H_2$ is inferentially compatible with $E_1$, it is eliminated by indirect evidence describing the implausibility of **ʃ* > *k*. Thus, the inferentialist account fares well in this case.

However, the inferentialist account turns problematic when we consider another case, where historical linguists aim to reconstruct a sound in Proto-Indo-European. The cognate sets are as follows:

| | *English* | *Latin* | *Old Irish* |
|---|---|---|---|
| 1 | father | pater | athir |
| 2 | fish | piscis | ĩasc |

$E_2$: The correspondence between the first sounds is *f* : *p* : *Ø* (zero).

There are three hypotheses for historical linguists to entertain:

$H_3$: The sound in Proto-Indo-European is **p*.
$H_4$: The sound in Proto-Indo-European is **f*.
$H_5$: The sound in Proto-Indo-European is **Ø*.

According to Lass (1997, p. 137), *p* becomes *f* or *f* becomes *p*, both of which are possible, though the former is more common than the latter, and *Ø* is virtually unseen. This shows directionality, so $H_5$ can be ruled out. However, neither $H_3$ nor $H_4$ is judged warranted merely by the single, impoverished correspondence set $E_2$. Stalemate is broken when extra information is added: 'Choice of a larger outgroup shows that all other Indo-European subgroups except Armenian (Greek, Slavonic, Indo-Iranian, Albanian, etc.) have /p/. If /f/ were original, there would be multiple convergences; if /p/ is original, then only Germanic (and Armenian in part) innovate' (Lass 1997, p.

137).[107] On the basis of the information, historical linguists apply the principles of majority wins and economy and conclude that the postulated proto-sound *$p$ is more evidentially favourable than *$f$, which indicates that $E_2$ is direct evidence for $H_3$ and does not count as direct evidence for $H_4$. Nevertheless, none of the pieces of information above are inferentially incompatible with $H_3$ or $H_4$. Even though there is more direct evidence for $H_3$ than $H_4$, both should be equally warranted unless one of them is eliminated, because more direct evidence just tells us more about what would be relevant to the hypothesis. Thus, according to the inferentialist account, there is no further indirect evidence and warranting evidence that can evidentially distinguish $H_3$ from $H_4$. Since historical linguists have, in effect, reconstructed the proto-sound as *$p$ ($H_3$), the inferentialist account does not fully capture the aspects of evidential reasoning in historical linguistics.

## 5.7 Settings Matter

Before delving into the issues of prior probabilities and catch-all likelihoods and examining whether the Bayesian theory fares well in the practice of sound reconstruction, I shall bring into focus the types of settings in terms of information availability. Sober (2008, p. 26) recognises proper priors that come from fully specified or well-specified evidential settings, on the one hand, but deems other kinds of priors either subjective or ill-defined, on the other hand. Fully specified evidential settings amount to complete-probability-model ones, such as coins, dice or a deck of cards. Well-specified evidential settings, by contrast, refer to empirically defensible settings, including medical cases, where disease prevalence, test sensitivity and test specificity are well supported by frequency data,[108] and genetics, where the Mendelian theory functions as probability assignments to hypotheses on solid empirical ground (Sober, 2008, p. 26). This is why Sober (2005)

[107] This is another way to symbolise phonemes. For example, /p/ is identical to *p*.
[108] In probabilistic terms, disease prevalence can be defined as P(the proportion of a population having the disease), sensitivity as P(a positive test / the patient has the disease), and specificity as P(a negative test / the patient does not have the disease). If these figures are available and correct, we can know the probability that the patient has the disease given that his test outcome is positive or negative.

suggests that:

> The Law of Likelihood should be restricted to cases in which the probabilities of hypotheses are not under consideration (perhaps because they are not known or are not even "well-defined") and one is limited to information about the probability of the observations given different hypotheses. (Sober, 2005, p. 128)

Despite being well suited for poorly specified settings, which are commonplace in scientific practice, especially for the historical sciences, the LL fails to make full use of information in fully specified and well-specified settings. This is one limitation of the LL. Another limitation, as mentioned above, is the inability to compare composite hypotheses in terms of the data.

Apart from fully specified or well-specified priors, the credibility of priors of other kinds is dubious. Recall that in Chapter 2, I have examined the subjective and objective Bayesian theories of evidence and have used counter-examples to present their difficulties; from a tool-based point of view, neither does its job properly. They need to either restrict their scope of application or improve their functions to overcome those counter-examples. Here I want to focus on their limitations, such as a hammer hardly replaces a screwdriver. In the simplified definitions (leaving out background knowledge B), the Bayesian theories states that (1) E is evidence for H if and only if P(H/E) > P(H) and/or (2) E is evidence for H if and only if P(H/E) attains a certain high level. Clearly, priors enter into both definitions.

## 5.8 The Problems with Priors and Catch-Alls

Priors, in the subjective Bayesian sense, can undermine the objectivity of science. Subjectivists maintain that priors are subject to agents' beliefs and are relatively free of rational constraints as long as they obey the axioms of probability. From the subjectivists' viewpoint, although initial probability assignments to parameters or hypotheses would vary substantially from individual to individual, their probabilities, as data are increasingly gathered overriding the influence of the different priors, tend to converge to a common

posterior probability. This is correct, but in a limited way. Consider astrophysics as a case in point.[109] For parameter estimation, the posterior probabilities of a particular parameter may eventually converge, even if scientists initially hold different priors representing a blending of their subjective beliefs and knowledge. But data are sometimes not sufficient to override priors in many different kinds of cases, e.g. 'for small sample sizes, or for problems where the dimensionality of the hypothesis space is larger than the number of observations (for example, in image reconstruction)' (Trotta, 2008, pp. 77–78). If priors are assigned based largely on one's beliefs and knowledge, this reflects a turn away from objectivity that researchers, including historical linguists, set out to discover.[110]

Contemporary objective Bayesians place more constraints on priors than do subjectivists.[111] For example, Jon Williamson (2010) puts forth three norms:

> **Probability:** degrees of belief should be probabilities,
> **Calibration:** they should be calibrated with evidence,
> **Equivocation:** they should otherwise equivocate between basic outcomes. (Williamson, 2010, p. iii)

[109] Many contemporary astrophysicists embrace the Bayesian view of probability, for frequentist methods are usually based on a full specification of the probability distributions, which may not accurately describe the problem under consideration, and frequentists draw on these distributions to describe possible, and unobserved, data, which leads to the *stopping rule* problem (Trotta, 2008, pp. 74–75). In short, for physicists, 'Bayesians address the question everyone is interested in by using assumptions no-one believes, while frequentists use impeccable logic to deal with an issue of no interest to anyone' (Lyons, 2007, p. 363). I do not consider these issues here, because they are not my main points. For further details on the stopping rule problem, see Royall (2004, pp. 126–127) and Howson and Urbach (2006, pp. 248–250).

[110] Sometimes subjective priors can be elicited by a route that can be taken as legitimate by scientists and objective Bayesians. For example, a claim that there is a 60% chance of rain tomorrow is elicited from expert opinions (Williamson, 2010, pp. 17–18). However, experts who are the stakeholders of the company may intentionally 'cook' extreme numbers. The priors elicited from these biased expert opinions can influence the conclusion of the investigation and further lead to severe environmental consequences (Dennis, 1996, p. 1099).

[111] 'Contemporary' objective Bayesianism stands in contrast to Carnapian objective Bayesianism, which, as noted in Chapter 2, interprets probability as a logical relation between propositions and links the logical probability to rational degrees of belief. An important distinction between the two standpoints is that '[t]he logical interpretation typically focuses on equivocation at the expense of calibration, while the [contemporary] objective Bayesian interpretation takes the Equivocation norm to be subsidiary to the Calibration norm — one should only equivocate to the extent that calibration with evidence does not fully determine which degrees of belief to adopt' (Williamson, 2010, p. 22).

The probability norm, also shown in subjective Bayesianism, is that beliefs should obey the axioms of probability. The calibration norm states that one's degrees of belief (i.e. probabilities) should allow for empirical information; this is pertinent to what I mentioned earlier about admissible prior probability assignment in response to fully specified or well-specified evidential settings.[112] Wherever relevant empirical information is present, we should try to use priors representing beliefs or hypotheses as far as possible. The equivocation norm requires that the (remaining) possibilities that cannot be calibrated with empirical information should be assigned equal degrees of belief (i.e. probabilities). Let us consider Darrell Rowbottom's (2015, pp. 62–64) example to illustrate these norms in turn. Imagine there is a tetrahedral (four-sided) die. Given the sample space $\Omega = \{1, 2, 3, 4\}$, we believe $P(\Omega) = 1$, $P(\emptyset) = 0$ and so forth, which meet the probability axioms. From the previous data showing that the frequency of '1' occurrence is 40%, we believe $P(1) = 0.4$; this is calibration. We should equivocate after calibrating. By the equivocation norm, we assign an equal probability to the other possibilities and believe $P(2) = P(3) = P(4) = 0.2$.

Priors, in the objective Bayesian sense, can be undecidable. Equivocating the probabilities of parameters, events, models, hypotheses or anything on which can be conferred probabilities hinges upon the way that their possibilities are carved up. Different ways of carving up possibly yield inconsistent results; this is the so-called Bertrand paradox. To illustrate how the possibilities divided according to the equivocation norm can lapse into contradiction, let us consider van Fraassen's (1989, pp. 302–304) version of the paradox: the *cube factory*.[113] Suppose that a factory manufactures cubes with side length between 0 and 1 cm in a perfectly random pattern, which can be described as a uniform distribution. A cube is then randomly chosen.

[112] There is a more sophisticated stance towards subjective Bayesianism, called empirically based subjective Bayesianism, which 'takes the Probability norm together with a further Calibration norm as necessary and sufficient for rationality at a particular time' (Williamson, 2010, p. 15).
[113] Bertrand paradoxes also include, among others, the book paradox, the wine/water paradox and the chord paradox. For details, see Gillies (2000, pp. 37–42).

By the equivocation norm, it follows that P(the side length of the cube is 0 to 0.5 cm) = P(the side length of the cube is 0.5 to 1 cm) = 1/2. However, if an area-based method of carving up events is adopted, this implies that P(the face area of the cube is 0 to 0.25 $cm^2$) = 1/4, meaning that two different probabilities can be assigned to the same event. The implementation of the equivocation norm fails to arbitrate between the two inconsistent results in similar cases: it is sometimes silent about how to count possibilities of infinite and continuous events before assigning equal probabilities to them.[114]

The equivocation of priors can also be suspect for historical linguists, who typically expect finite and discrete events. To illustrate this, imagine that a historical linguist, upon reconstructing proto-sounds, attempts to utilise the objective Bayesian approach to help calculate and determine whether a particular sound, say **k*, in a proto-language is sufficiently probable given the sound correspondence of its daughter languages, *k*- : *k*- : *k*- : *ʃ*- ($E_1$). By Bayes' theorem, it follows that $P(*k/E_1) = P(*k)P(E_1/*k) / P(E_1)$. Let us assume that the linguist successfully estimates $P(E_1)$ and $P(E_1/*k)$ in some way. Now the question is: how can the linguist estimate the prior P(**k*)? Suppose the frequency is a good estimate of P(**k*). The frequency of **k* depends upon the kinds of proto-sounds. Even with acquiring this information and knowing how many kinds of possible proto-sounds existed, the linguist still needs to know how likely each proto-sound is (this is similar to the previous tetrahedral dice example).

I would reiterate that calibration should be applied prior to equivocation. Lay people would be allowed to assign equal probabilities to each of the proto-sounds. However, the linguist, with knowledge of phonology and phonetics, believes that the plausibility for each proto-sound is likely to vary, but fails to estimate P(**k*), which is a precise value. Still, even if the linguist is advised to equivocate the possible proto-sounds, this would be tantamount to the likelihoodist account when the two hypotheses are compared. Suppose another alternative proto-sound is **ʃ*. It is obtained that $P(*k/E_1) / P(*ʃ/E_1) =$

[114] Williamson (2010, §9.1) acknowledges that some paradoxes are insoluble.

P(*$k$)P($E_1$/*$k$) / P(*$ʃ$)P($E_1$/*$ʃ$), which equals P($E_1$/*$k$) / P($E_1$/*$ʃ$) on the basis that P(*$k$) = P(*$ʃ$) = 1 / $n$, where $n$ stands for the number of all alternatives. When one compares the posterior probabilities in this situation, the priors are equal and can be cancelled off, and what is left is the likelihood ratio.

There are no priors as such in historical linguistics: if any existed, they would not be considered scientifically objective. Historical linguists do not, and cannot, confer an objective and precise probability upon hypotheses as to the past: how can they know or objectively determine the prior probability of the common origin hypothesis or of the diffusion hypothesis for a specific language group? Sometimes they may be able to assign vague priors to hypotheses instead. For example, the probability of the hypothesis that the proto-sound is velar trill is almost zero, as the sound is phonetically impossible to pronounce, while the probability of the hypothesis that the proto-sound is palatal trill is low because the sound is possible for humans to make but linguists have never found them actually being used in any of the world's languages. Equipped with the vague priors, we can obtain the inequality of priors. Note that historical linguists would not take any inequalities of priors into account. As Campbell (2013, p. 126) puts it, 'Certain inventories of sounds are found with frequency among the world's languages while some are not found at all and others only very rarely. When we check our postulated reconstructions for the sounds of a proto-language, we must make sure that we are not proposing a set of sounds which is never or only very rarely found in human languages'. The information about vague priors and inequalities of priors is, after all, of little use in historical linguistics because of the problem with catch-alls.

Perhaps the problem of priors can be circumvented by using only catch-alls and likelihoods. This is Fitelson's (2007) idea, when investigating the common ground between likelihoodism and Bayesianism. He puts forth the *weak law of likelihood*:

> (WLL) E is evidence in favour of $H_1$ versus $H_2$ if $P(E/H_1) > P(E/H_2)$ and $P(E/\sim H_1) \leq P(E/\sim H_2)$.

The LL entails the WLL. The proof is as follows:

1. If $P(E/H_1) > P(E/H_2)$, E is evidence in favour of $H_1$ versus $H_2$. [the LL]
2. $P(E/H_1) > P(E/H_2)$ and $P(E/\sim H_1) \leq P(E/\sim H_2)$. [ACP: Assumption for Conditional Proof]
3. $P(E/H_1) > P(E/H_2)$. [from 2]
4. E is evidence in favour of $H_1$ versus $H_2$. [from 1&3]
5. If $P(E/H_1) > P(E/H_2)$ and $P(E/\sim H_1) \leq P(E/\sim H_2)$, E is evidence in favour of $H_1$ versus $H_2$. [from 2–4 CP: Conditional Proof]

The WLL is logically weaker than the LL, which is why it is called a *weak* version of the LL.[115] Since likelihoodists believe in the LL, there is no reason for them not to accept the WLL.[116]

Nevertheless, likelihoodists can maintain that the WLL can be applied where catch-alls are available, while arguing that in myriad cases where catch-alls are elusive, the LL can serve to determine the inequality of evidential strength, but the WLL is not applicable because of the problem with catch-alls. Now I shall argue that catch-all likelihoods are out of intellectual reach in historical linguistics for the following two reasons.

Firstly, alternative hypotheses may not be collectively exhaustive. There are a great variety of events where their catch-all hypotheses cannot be identified. A catch-all hypothesis, P(~H), is the negation of the hypothesis of interest and can break down into a complete collection of its alternative hypotheses which are mutually exclusive.

[115] In Fitelson's (2007) paper, when he mentioned the LL, he in fact referred to the BLL: E is evidence that favours $H_1$ over $H_2$ if and only if $P(E/H_1) > P(E/H_2)$.
[116] Fitelson (2007) uses the WLL to illustrate that the LL cannot reject his counter-example regarding principle (*), which I discussed in Section 5.3.

Sober (2008, pp. 28–29) nicely illustrates the exhaustion problem with a case from physics. The scientist Arthur Eddington, in an attempt to test the general theory of relativity (GTR), measured the extent to which light deflects (E) during a solar eclipse in 1919. While P(E/GTR) can be estimated, P(E/~GTR), i.e. P(E/the disjunction of hypotheses other than GTR), remains unknown insofar as the catch-all hypothesis is hard to fully unpack. In fact, Eddington tested the GTR against a specific theory, Newtonian gravity (NG): this can be expressed with the inequality of specific likelihoods: P(E/GTR) ≫ P(E/NG). Reverting to the primary concern of this chapter, a historical linguist hypothesises that one of the sounds in Proto-Romance is **k* ($H_1$). Even granting that the probability of the data given $H_1$ is available, it seems impossible to compute $P(E_1/\sim H_1)$ in practice, because the proto-sounds may be impossible to fully identify, and the plausibility of each of them is likely to be inconclusive.

Secondly, the problem of priors rearises, even if it can sometimes be ascertained that historical linguists have exhausted the possibilities of competing hypotheses, as, for example, all and only sources of language traits, namely inheritance ($H_i$), diffusion ($H_d$) and chance ($H_c$). Given that E is a set of language traits, $H_d$ and $H_c$ are mutually exclusive, and $P(E/\sim H_i)$ is equivalent to a weighted average of the individual likelihoods, namely $P(E/H_d \vee H_c)$, it follows that:

$$
\begin{aligned}
& P(E/\sim H_i) \\
&= P(E/H_d \vee H_c) \\
&= [P(E)P(H_d \vee H_c/E)] / P(H_d \vee H_c) \\
&= [P(E)P(H_d/E) + P(E)P(H_c/E)] / [P(H_d) + P(H_c)] \\
&= [P(H_d)P(E/H_d) + P(H_c)P(E/H_c)] / [P(H_d) + P(H_c)].
\end{aligned}
$$

This formula inevitably involves the priors, namely $P(H_d)$ and $P(H_c)$. As has been argued above, priors are elusive, and it is unclear as to how to avoid using the priors in this case. This argument also handles the above **k* case: even if the possibilities of *~k are identified, the problem of the prior is still

irresolvable. Thus, the WLL, invoking the catch-alls, is not helpful with evidential evaluation in historical linguistics.

On a negative note, I have argued why the explanationist theory, the argument theory, the inferentialist account, the subjective Bayesian theory and the objective Bayesian theory do not fit well with historical linguistics (cf. Tucker, 2004) and that the Bayesian approach carries the burdens of priors and catch-alls, which are not required in likelihoodism and are elusive in historical linguistics. In a more positive vein, I shall argue that the comparative nature built into the likelihoodist account fits well with the methodology of historical linguistics. For this purpose, I will attend to the cases of sound reconstruction.

## 5.9 Case Study: Strengths of the Likelihoodist Account

Now I shall demonstrate the strength of likelihoodism for evidential reasoning in historical linguistics. Before returning to the previous case that is used to examine the explanationist theory, the argument theory and the inferentialist account, I formalise and comparatively quantify why one reconstructed sound, rather than another, is favoured in the light of a 'natural' direction of sound change, majority wins and economy.

Inspired by the way that Sober (2008) employs the LL and relevant assumptions to demonstrate that the observed trait shared between two species evidentially favours the common ancestry hypothesis over the separate ancestry hypothesis, I assume that the sounds in the correspondence set under investigation concern a dichotomous variable, which has two states, *r* and *s*. Let X denote a descendant sound and *X denote a proto-sound. The events include X = *r*, X = *s*, *X = *r* and *X = *s*. Let P(X = *s* /*X = *r*) = $\theta_1$, (X = *r* /*X = *r*) = 1−$\theta_1$, P(X = *r* /*X = *s*) = $\theta_2$ and P(X = *s* /*X = *s*) = 1−$\theta_2$. We have four assumptions, which I will explain later:

(1) $\theta_1 < 1-\theta_2$ and $\theta_2 < 1-\theta_1$. [backwards inequality]

(2) $\theta_1 > \theta_2$. [from directionality]

(3) $P(X_1 : X_2 : \ldots : X_k /{}^*X) = P(X_1/{}^*X)\, P(X_2/{}^*X) \ldots P(X_k/{}^*X)$, where k is the number of the sounds in a particular correspondence set. [sound change independence]

(4) $0 < \theta_1, \theta_2 < 1$. [non-extreme probabilities]

From (3), $P(X_1 : X_2 : \ldots : X_k /{}^*X = s) = \theta_2^n\, (1-\theta_2)^m$, where m and n are natural numbers and m + n = k. Let m < n and we have

$$\begin{aligned}
&\theta_2^n\, (1-\theta_2)^m \\
&= \theta_2^{n-m}\, \theta_2^m\, (1-\theta_2)^m \\
&= \theta_2^{n-m}\, [\theta_2\, (1-\theta_2)]^m \\
&= \theta_2^{n-m}\, (\theta_2 - \theta_2^2)^m.
\end{aligned}$$

Since from (1), (2) and (4), $0 < \theta_1+\theta_2 < 1$ and $0 < \theta_1-\theta_2$, we obtain

$$\begin{aligned}
&(\theta_1+\theta_2)\, (\theta_1-\theta_2) < (\theta_1-\theta_2) \\
\Rightarrow^{117}\ &\theta_1 - \theta_1^2 > \theta_2 - \theta_2^2.
\end{aligned}$$

Then we have

$$\theta_2^{n-m}\, (\theta_2 - \theta_2^2)^m < \theta_2^{n-m}\, (\theta_1 - \theta_1^2)^m.$$

From (1), we have

$$\begin{aligned}
&\theta_2^{n-m}\, (\theta_1 - \theta_1^2)^m < (1-\theta_1)^{n-m}\, [\theta_1\, (1-\theta_1)]^m \\
\Rightarrow\ &(1-\theta_1)^{n-m}\, \theta_1^m\, (1-\theta_1)^m < \theta_1^m\, (1-\theta_1)^n.
\end{aligned}$$

Therefore, from assumptions (1)–(4), we have

$$\theta_2^n\, (1-\theta_2)^m < \theta_1^m\, (1-\theta_1)^n.$$

[117] '⇒' stands for 'logically entails'.

Let us revert to the case of reconstructing a proto-sound in Proto-Romance on the basis of the following sound correspondence:

$E_1$: Italian *k*- : Spanish *k*- : Portuguese *k*- : French *ʃ*-.

Historical linguists entertain two hypotheses:

$H_1$: The proto-sound is **k.*
$H_2$: The proto-sound is **ʃ*.

As is known, historical linguists have reconstructed **k*. The rationale behind the reasoning can be spelled out. According to the LL, the task amounts to comparing $P(E_1/H_1)$ and $P(E_1/H_2)$: if $P(E_1/H_1) > P(E_1/H_2)$, then $E_1$ is evidence that favours $H_1$ over $H_2$. Let $P(E_1/H_1) = P(k : k : k : ʃ / *k)$ and $P(E_1/H_2) = P(k : k : k : ʃ / *ʃ)$; we obtain $P(k : k : k : ʃ / *k) = \theta_1(1-\theta_1)^3$ and $P(k : k : k : ʃ / *ʃ) = \theta_2^3 (1-\theta_2)$. Assuming *k* and *ʃ* are the only two possible states, it follows from assumptions (1)–(4) that $\theta_2^3 (1-\theta_2) < \theta_1 (1-\theta_1)^3$. By the LL, we obtain $P(E_1/H_1) > P(E_1/H_2)$. Assumptions (1)–(4), with the LL, suffice to show that $E_1$ is evidence in favour of $H_1$ versus $H_2$. The conclusion of the proof accords with the reconstruction suggested by the principles of directionality, majority wins and economy. The reconstructed sound **k*, relative to **ʃ*, is supported by evidence.

The above assumptions are not *a priori* true, but they generally hold in historical linguistics. The model entails a *backwards inequality*, stating that $P(a \longrightarrow a) > P(b \longrightarrow a)$, where *a* and *b* are two different states that a character has in a lineage, and $P(a \longrightarrow b)$ is the probability that a character changes from *a* to *b* (Sober, 2008, pp. 215–216).[118] Assumption (1) obeys a backwards inequality: given a descendant's state, the ancestor's state is most likely to be the same. Assumption (2) is based on the linguistic data,

[118] It is worth noting that there is another inequality called the *forwards inequality*, $P(a \longrightarrow a) > P(a \longrightarrow b)$. The forwards inequality is 'highly contingent', whilst the backwards inequality is 'extremely robust' (Sober, 1988, p. 223).

which indicates a relatively natural direction of sound change. Suppose that the changes of sounds that linguists endeavour to reconstruct are due to language contact. If so, assumption (3) fails to hold. However, the comparative method is an effective method that can block or lessen the possibilities, other than inheritance, that may bring about sound changes.[119]

Note that assumptions (1)–(4) suffice for the sound correspondence to be evidence that favours the reconstructed **k* over **ʃ*, but if one of these assumptions is not true, this does not necessarily mean that $E_1$ is not evidence in favour of $H_1$ versus $H_2$. It may be the case that $E_1$ still supports **k* more than **ʃ*, but some additional or a different set of well-supported assumptions are called for. Whether $E_1$ is evidence in favour of $H_1$ versus $H_2$ depends upon empirical matters of fact.

It becomes more complicated and unsure in this case if a sound can change into more than one sound, i.e. multiple states. For example, **k* can remain as *k* or change to *ʃ*, *č*, and so forth; $P(k / *k) + P(ʃ / *k) \leq 1$, as described in assumption (5). Let $P(ʃ / *k) = \theta_1$, $P(k / *k) = \theta_{1*}$, $P(k / *ʃ) = \theta_2$ and $P(ʃ / *ʃ) = \theta_{2*}$. We add two additional assumptions:

(5) $\theta_1 + \theta_{1*} \leq 1$, $\theta_2 + \theta_{2*} \leq 1$. [multiple possible sound changes]
(6) $\theta_{1*} > 1/2$. [tendency towards non-change]

From (3), $P(X_1 : X_2 : \ldots : X_k / *X = s) = \theta_2^n \theta_{2*}^m$, where m and n are natural numbers and m + n = k. Let m < n and we have

$$\theta_2^n \theta_{2*}^m = \theta_2^{n-m} \theta_2^m \theta_{2*}^m.$$

From the *inequality of arithmetic and geometric means* (the AM–GM inequality): $\sqrt[k]{a_1 a_2 \ldots a_k} \leq \frac{a_1 + a_2 + \ldots + a_k}{k}$, we have

[119] Similarities emerging from sheer chance can be dismissed by checking whether there are systematic sound correspondences. For example, the *m-* : *m-* correspondence postulated by the match between *mes* in Kaqchikel and *mess* in English does not repeat itself and the *m* sounds coincide by chance. (Campbell, 2013, p. 112).

$$\theta_2^{n-m}\theta_2^m\theta_{2*}^m \leq \theta_2^{n-m}(\theta_2+\theta_{2*})^{2m}/\ 2^{2m}.$$

From (2), (5) and (6), we obtain

$$\theta_2^{n-m}(\theta_2+\theta_{2*})^{2m}/\ 2^{2m} \leq \theta_2^{n-m}(\theta_2+1-\theta_2)^{2m}/\ 2^{2m} = \theta_2^{n-m}/\ 2^{2m}$$
$$\Rightarrow \theta_2^{n-m}/\ 2^{2m} < \theta_2^{n-m}\theta_{1*}^{2m}$$
$$\Rightarrow \theta_2^{n-m}\theta_{1*}^{2m} < \theta_1^{n-m}\theta_{1*}^{2m}$$
$$\Rightarrow \theta_1^{n-m}\theta_{1*}^{2m} < \theta_1^m\,\theta_{1*}^n.$$

Therefore, from assumptions (2)–(6), we have

$$\theta_2^n\,\theta_{2*}^m < \theta_1^m\,\theta_{1*}^n.$$

We know $P(E_1/H_1) = P(k : k : k : ʃ / *k) = \theta_1\theta_{1*}^3$ and $P(E_1/H_2) = P(k : k : k : ʃ / *ʃ) = \theta_2^3\,\theta_{2*}$. By the LL, we obtain $P(E_1/H_1) > P(E_1/H_2)$, which means that $E_1$ is evidence that favours $H_1$ over $H_2$. Assumption (6) can be supported by the fact that 'voiceless stop consonants have, by and large, been preserved intact in word-initial position throughout the Romance territory' (Hall, 1964, p. 551).

However, the three general principles sometimes conflict, and this under-determination of sound reconstruction can be formalised and comparatively quantified with LL. Suppose the correspondence set described in $E_1$ is, in fact, the following set:

$E_1$*: *ʃ*- : *ʃ*- : *ʃ*- : *k*-.

As opposed to **k* suggested by the principle of directionality given $E_1$*, the principles of majority wins and economy suggest that the proto-sound should be reconstructed as **ʃ* in the light of $E_1$*. A verdict cannot be reached with the LL either. The likelihoods of $H_1$ and $H_2$ are: $P(E_1*/H_2) = P(ʃ : ʃ : ʃ : k / *ʃ) = \theta_2(1-\theta_2)^3$ and $P(E_1*/H_1) = P(ʃ : ʃ : ʃ : k / *k) = \theta_1^3(1-\theta_1)$, respectively. But it,

using only assumptions (1)–(4), cannot prove whether $H_1$ or $H_2$ is favoured by $E_1$*.

Finally, let us return to the other case, where $E_2$: The correspondence set is *f*- : *p*- : Ø-, $H_3$: The sound in Proto-Indo-European is **p*, and $H_4$: The sound in Proto-Indo-European is **f*. When the correspondence set is expanded according to the information about all other Indo-European subgroups except Armenian having *p*, the same reasoning invoking the LL above can be applied to the case on the basis that *p* outnumbers *f* in the corresponding positions in the languages and that the directionality indicates P(*f* / **p*) > P(*p* / **f*). Here, assumption (6) can be supported by another fact that '[other aspects of Proto-Indo-European], for example the voiceless stops, seem phonetically secure' (Baldi, 2002, p. 18). According to the likelihoodist account, $E_2$ is evidence in favour of $H_3$ versus $H_4$.

I have argued that the likelihoodist account can shed light on the evidential reasoning in the proto-sound reconstruction. This account has made explicit how the principles of directionality, majority wins and economy collaborate to turn the data into evidence for hypotheses of proto-sounds and why they fail to do so when they disagree. By the LL, as the core of a relevance-mediating vehicle, the linguistic data are therefore labelled evidence.

## 5.10 Conclusion

Reconstruction of the past is therefore, for the most part, being done piecemeal. It is worth quoting Sober's (1988) view at length here:

> We must not ask whether the past is knowable, but whether this or that specific aspect of the past is knowable. …it would be folly to try to produce an *a priori* argument that shows that evolutionary history must always be recoverable. Whether this is true depends on contingent properties of the evolutionary process. The folly would be greater still to try to mount some general philosophical argument to the effect that the past as a whole must be knowable. The history of stars, of living things, and of human languages, to mention just three examples, will be retrievable only if empirical facts specific to the

> processes governing each are favorable. This is no global question for the armchair philosopher to answer by pondering evil demons or other epistemological fantasies. Rather, the pertinent questions are local in scope, which the astronomer, the evolutionist, and the linguist can each address by considering the discriminatory power of available data and process theories. (Sober 1988, p. 5)

The tool-based view of theories of evidence echoes the quote. For the historical sciences, there are not many logically strong assumptions of a process model on which we can rely. The likelihoodist account of evidence does not appeal to the impractical quantitative explanation strength, or the highly strict argument, or the narrow-scoped elimination strategy. Rather, this account compares hypotheses regarding the past in weighing their likelihoods, and explicates sound reconstruction and potentially other activities of evidential reasoning in the historical sciences.

I am not claiming that prior probabilities are not informative about evidential strength. Instead, priors can sometimes provide additional information about the extent to which evidence supports a hypothesis, but only if they are grounded on scientific theories or empirical data. For subjective Bayesians, priors are estimated and vary depending on individuals, which does not cohere with the objectivity we seek. Objective Bayesians, by contrast, can follow Williamson (2010) and appeal to the equivocation norm that assigns the same probabilities to competing hypotheses independently of us. But this assignment is neither true *a priori* nor secured by scientific theories or empirical data. Priors, if any, are scarce; lack of reliance on priors is another strength of likelihoodism.

The likelihoodist account of evidence has its own limitations. It is not well suited to composite hypotheses and well-specified and fully specified settings. This account specifically attends to objective support (OS) but is silent about other evidential relationships including subjective acceptance (SA) and guidance (GD). Perhaps as Cartwright (2014) points out, when evidence is not sufficient to guarantee a hypothesis, or evidence is meagre, as commonly seen in the historical sciences, we have to plump for the hypothesis. This is an epistemic reality we face.

# 6
# Conclusion

Throughout this thesis, I endorse a tool-based view of theories of evidence. A theory of evidence is a purpose-specific and setting-sensitive tool, not a Swiss army knife. Achinstein's explanationist theory is a good tool for explaining the concept of evidence used in manifold situations in science and daily lives. Cartwright's argument theory is good at reminding us of inferential gaps. The Bayesian approach fares well where probabilistic information is sufficient. These theories are well suited to fully specified or well-specified evidential settings. However, such settings are scarce in the historical sciences, where much information is lacking; these theories do not work properly in poorly specified settings. Still, there is some information about vague likelihood comparison. Reiss's inferentialist account is a useful tool for spelling out why a variety of sources of evidence are deemed legitimate in the biomedical and social sciences, but it cannot digest this kind of quantitative information. I have argued that the likelihoodist account of evidence is a better tool for this purpose.

The likelihoodist account can explicate the comparative nature of evidential reasoning in the historical sciences. I have illustrated this point with cases of the proto-sound reconstruction in historical linguistics. Despite being an appropriate tool for labelling what the information says about hypotheses as evidence, this account is silent about whether we are justified in believing a hypothesis given the evidence. Even in such poorly specified settings, if historical linguists are fortunate in that a volume and diversity of evidence mounts up and zeros in on the same hypothesis, they may be justified in accepting the hypothesis. However, it is worrying that when the evidence is meagre or discordant, historical linguists would require a leap of faith. This is not a counsel of despair, but rather honesty about our epistemic reality and a spur to the development of better tools for methodology descriptions and evidential reasoning.

Much work remains. Genealogical inference in historical linguistics is not limited to reconstructing the features of proto-languages, but embodies many aspects of language change, such as whether the inheritance hypothesis is generally evidentially favourable over the diffusion or chance hypothesis, and what kinds of rational considerations or constraints would influence working linguists' judgements about evidence. Moreover, archaeology, text criticism, history and other fields concerning historiography involve evidential reasoning similar to likelihoodism, and whether or to what extent the law of likelihood is extendable into those domains merits further investigation. Finally, in the historical sciences, it remains an open question as to whether there is a meta-analysis that can inform researchers of the direction that the evidence leans towards or even evaluate the strengths of the amalgamated evidence.

# References


Achinstein, P. (1983) *The Nature of Explanation*. New York: Oxford University Press.

Achinstein, P. (1995) 'Are Empirical Evidence Claims A Priori ?', *British journal for the philosophy of science*, 46(4), pp. 447–473.

Achinstein, P. (2001) *The Book of Evidence*. New York: Oxford University Press.

Achinstein, P. (2005) *Scientific evidence: Philosophical theories and applications*. Baltimore: Johns Hopkins University Press.

Achinstein, P. (2010a) *Evidence, Explanation, and Realism: Essays in the Philosophy of Science*. New York: Oxford University Press.

Achinstein, P. (2010b) 'Mill's Sins or Mayo's Errors?', in Mayo, D. G. and Spanos, A. (eds) *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge: Cambridge University Press, pp. 170–188.

Achinstein, P. (2014) 'Evidence', in Curd, M. and Psillos, S. (eds) *The Routledge Companion to Philosophy of Science*. 2nd edn. New York: Routledge, pp. 381–392.

Akaike, H. (1973) 'Information Theory as an Extension of the Maximum Likelihood Principle', in Petrov, B. N. and Caski, F. (eds) *Proceeding of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–281.

Angrist, J. D. (1990a) 'Errata: Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records', *American Economic Review*, 80(5), pp. 1284–1286.

Angrist, J. D. (1990b) 'Lifetime Earnings and the Vietnmam Era Draft Lottery: Evidence from Social Security Records', *American Economic Review*, 80(3), pp. 313–336.

Angrist, J. and Evans, W. N. (1998) 'Children and Their Parent's Labior Suppy: Evidence from Exogenous Variation in Family Size', *The American Economic Review*, 88(3), pp. 450–477.

Baldi, P. (2002) *The Foundations of Latin*. Berlin: Moulon de Gruyter.

Bareinboim, E. and Pearl, J. (2013) 'A General Algorithm for Deciding Transportability of Experimental Results', *Journal of Causal Inference*, 1(1), pp. 107–134.

Bennett, J. (1988) *Events and their Names*. Oxford: Oxford University Press.

Berger, J. O. and Delampady, M. (1987) 'Testing Precise Hypotheses', *Statistical Science*, 2(3), pp. 317–335.

Bickel, D. R. (2012) 'The Strength of Statistical Evidence for Composite Hypotheses: Inference to the Best Explanation', *Statistica Sinica*, 22, pp. 1147–1198.

Bird, A. (2010a) 'Eliminative Abduction: Examples from Medicine', *Studies in History and Philosophy of Science Part A*. Elsevier Ltd, 41, pp. 345–352.

Bird, A. (2010b) 'Inductive knowledge', in Bernecker, S. and Pritchard, D. (eds) *The Routledge companion to epistemology*. New York: Routledge, pp. 271–282.

Birnbaum, A. (1962) 'On the foundations of statistical inference', *Journal of the American Statistical Association*, 57(298), pp. 269–306.

Brueckner, A. (1992) 'What an Anti-Individualist Knows A Priori', *Analysis*, 52(2), pp. 111–118.

Campbell, L. (2013) *Historical Linguistics: An Introduction*. 3rd edn. Edinburgh: Edinburgh University Press.

Campbell, L. and Mixco, M. J. (2007) *A Glossary of Historical Linguistics*. Edinburgh: Edinburgh University Press.

Carnap, R. (1937) *Logical syntax of language*. London: Kegan, Paul, Trench, Teubner & Co.

Carnap, R. (1962) *Logical Foundations of Probability*. 2nd edn. Chicago: University of Chicago Press.

Cartwright, N. (1999) *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Cartwright, N. (2007) 'Are RCTs the gold standard?', *BioSocieties*, 2, pp. 11–20.

Cartwright, N. (2010) 'What Are Randomised Controlled Trials Good For?', *Philosophical Studies*, 147(1), pp. 59–70.

Cartwright, N. (2012) 'Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps', *Philosophy of Science*, 79(5), pp. 973–989.

Cartwright, N. (2013) 'Evidence, Argument and Prediction', in Karakostas, V. and Dieks, D. (eds) *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*. Basel: Springer, pp. 3–17.

Cartwright, N. (2014) ‘A Question of Nonsense’, *Iyyun: The Jerusalem Philosophical Quarterly*, 63, pp. 102–116.

Cartwright, N. and Hardie, J. (2012) *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford University Press.

Chandler, J. (2013) ‘Contrastive confirmation: Some competing accounts’, *Synthese*, 190, pp. 129–138.

Clarke, B. *et al.* (2014) ‘Mechanisms and the evidence hierarchy’, *Topoi*, 33(2), pp. 339–360.

Coleman, R. G. (1993) ‘Vulgar Latin and Proto-Romance: Minding the Gap’, *Prudentia*, 25, pp. 1–14.

Davies, M. (1998) ‘Externalism, Architecturalism, and Epistemic Warrant’, in Wright, B. C. S. and Macdonald, C. (eds) *Knowing Our Own Minds*. Oxford: Oxford University Press, pp. 321–362.

de Finetti, B. (1937) ‘Foresight: Its logical laws, its subjective sources’, in Kyburg, H. E. and Smokler, H. E. (eds), Kyburg, H. E. (tran.) *, 1980, Studies in subjective Probability*. 2nd edn. New York: Krieger, pp. 53–118.

Dennis, B. (1996) ‘Discussion: Should Ecologists Become Bayesians?’, *Ecological Applications*, 6(4), pp. 1095–1103.

Devlin, K. (2003) ‘Devlin’s Angle: Monty Hall’, *The Mathematical Association of America*.

Donohue, M., Denham, T. and Oppenheimer, S. (2012) ‘New methodologies for historical linguistics?: Calibrating a lexicon-based methodology for diffusion vs. subgrouping’, *DiachronicaDiachronica International Journal for Historical Linguistics. Founded by E.F.K. Koerner, General Editor, 1984–2001*, 29(4), pp. 505–522.

Draper, D. and Madiganv, D. (1997) ‘The scientific value of Bayesian statistical methods’, *IEEE Intelligent Systems and Their Applications*, 12, pp. 18–21.

Dunn, M. (2015) ‘Language phylogenies’, in Bowern, C. and Evans, B. (eds) *The Routledge Handbook of Historical Linguistics*. New York: Routledge, pp. 190–211.

Dunning, T. (2012) *Natural experiments in the social sciences: A design-based approach*. New York: Cambridge University Press.

Edwards, A. W. F. (1972) *Likelihood*. Maryland: Johns Hopkins University Press, 1992.

Encyclopaedia Britannica. (2019) ‘Grigori Rasputin’, *Encyclopædia Britannica*. Encyclopædia Britannica, inc. Available at: https://www.britannica.com/biography/Grigory-Yefimovich-Rasputin (Accessed: 29 March 2019).

Falk, R. (1992) ‘A Closer Look At the Probabilities of the Notorious Three Prisoners’, *Cognition*, 43, pp. 197–223.

Fitelson, B. (2007) ‘Likelihoodism, Bayesianism, and Relational Confirmation’, *Synthese*, 156(3), pp. 473–489.

Florkowski, C. M. (2008) ‘Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests’, *The Clinical Biochemist Reviews*, 29 Suppl 1, pp. S83–S87.

Forster, M. and Sober, E. (2004) ‘Why Likelihood?’, in Taper, M. L. and Lele, S. R. (eds) *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago: The University of Chicago Press, pp. 153–190.

Galiani, S. and Schargrodsky, E. (2010) ‘Property rights for the poor: Effects of land titling’, *Journal of Public Economics*. Elsevier B.V., 94(9–10), pp. 700–729.

Gandenberger, G. (2015) ‘A New Proof of the Likelihood Principle’, *British Journal of the Philosophy of Science*, 66, pp. 475–503.

Gillies, D. (2000) *Philosophical Theories of Probability*. London: Routledge.

Glymour, C. (1980a) ‘Hypothetico-Deductivism is Hopeless’, *Philosophy of Science*, 47, pp. 322–325.

Glymour, C. (1980b) *Theory and Evidence*. New Jersey: Princeton University Press.

Goodman, S. N. (1999) ‘Toward Evidence-Based Medical Statistics. 2: The Bayes Factor’, *Annals of internal medicine*, 130(12), pp. 1005–1013.

Guyatt, G. H. *et al.* (2011) ‘GRADE Guidelines: 9. Rating Up the Quality of Evidence’, *Journal of Clinical Epidemiology*, 64(12), pp. 1311–1316.

Hacking, I. (1965) *Logic of Statistical Inference*. Cambridge: Cambridge University Press, 2016.

Hale, M. (2015) ‘The Comparative Method: Theoretical Issues’, in Bowern, C. and Bethwyn Evans (eds) *The Routledge Handbook of Historical Linguistics*. New York: Routledge, pp. 146–160.

Hall, R. A. (1950) ‘The Reconstruction of Proto-Romance’, *Language*, 26(1), pp. 6–27.

Hall, R. A. J. (1964) ‘Initial Consonants and Syntactic Doubling in West Romance’, *Language*, 40(4), pp. 551–556.

Hempel, C. G. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hempel, C. G. and Oppenheim, P. (1948) ‘Studies in the Logic of Explanation’, *Philosophy of Science*, 15(2), pp. 135–175.

Hitchcock, C. (1998) ‘The Common Cause Principle in Historical Linguistics’, *Philosophy of Science*, 65(3), pp. 425–447.

Hitchcock, C. and Sober, E. (2004) ‘Prediction versus Accommodation and the Risk of Overfitting’, *British Journal for the Philosophy of Science*, 55(1), pp. 1–34.

House of Commons Science and Technology Committee (2010) *Evidence Check 2: Homeopathy: Fourth Report of Session 2009–10*. London.

Howick, J., Glasziou, P. and Aronson, J. K. (2009) ‘The evolution of evidence hierarchies: what can Bradford Hill’s “guidelines for causation” contribute?’, *Journal of the Royal Society of Medicine*, 102(5), pp. 186–94.

Howson, C. (1997) ‘A Logic of Induction’, *Philosophy of Science*, 64(2), pp. 268–290.

Howson, C. and Urbach, P. (2006) *Scientific Reasoning: The Bayesian Approach*. 3rd edn. Chicago: Open Court.

Jaynes, E. T. (1968) ‘Prior Probabilities’, *IEEE Transactions on Systems Science and Cybernetics*, 4(3), pp. 227–241.

Jeffreys, H. (1938) ‘Maximum Likelihood, Inverse Probability, and the Method of Moments’, *Annals of Eugenics*, 8, pp. 146–151.

Johansson, T. (2011) ‘Hail the Impossible: P-values, Evidence, and Likelihood’, *Scandinavian Journal of Psychology*, 52(2), pp. 113–125.

Johnson, A. *et al.* (2007) ‘Guts & Glory H . pylori : Cause of Peptic Ulcer’, *Eukaryon*, 3, pp. 67–74.

Kass, R. E. and Raftery, A. E. (1995) ‘Bayes factors’, *Journal of the American Statistical Association*, 90(430), pp. 773–95.

Kelly, K. T. (2007) ‘A New Solution to the Puzzle of Simplicity’, *Philosophy of Science*, 74(5), pp. 561–573.

Kiparsky, P. (2015) ‘New Perspectives in Historical Linguistics’, in Bowern, C. and Evans, B. (eds) *The Routledge Handbook of Historical Linguistics*. New York: Routledge, pp. 64–102.

Lakatos, I. (1976) ‘Falsification and the Methodology of Scientific Research Programmes’, in Harding, S. G. (ed.) *Can Theories be Refuted?* Dordrecht: D. Reidel Publishing Company, pp. 205–259.

Lambert, K. (2002) *Free Logic: Selected Essays*. Cambridge: Cambridge University Press.

Lass, R. (1997) *Historical Linguistics and Language Change*. Cambridge University Press.

Last, J. M. (2001) *A Dictionary of Epidemiology*. 4th edn. Oxford: Oxford University Press.

Laudan, L. (1981) ‘A Confutation of Convergent Realism’, *Philosophy of Science*, 48(1), pp. 19–49.

Leeds, S. (2000) ‘Other Minds, Support, and Likelihoods’, *unpublished manuscript*.

Leibovici, L. (2001) ‘Effects of Remote, Retroactive Intercessory Prayer on Outcomes in Patients with Bloodstream Infection: Randomised Controlled Trial’, *British Medical Journal*, 323(7327), pp. 1450–1451.

Lipton, P. (2004) *Inference to the Best Explanation*. London: Routledge.

Lyons, L. (2007) ‘A Particle Physicists’s Perspective on Astrostatistics’, in *Statistical Challenges in Modern Astronomy IV*. Pennsylvania State University, Pennsylvania, USA, pp. 361–372.

Lytsy, P. (2017) ‘Creating Falseness—How to Establish Statistical Evidence of the Untrue’, *Journal of Evaluation in Clinical Practice*, 23, pp. 923–927.

Mackie, J. L. (1965) ‘Causes and Conditions’, *American Philosophical Quarterly*, 2(4), pp. 245–264.

Mayo, D. G. (1996) *Error and the Growth of Experimental Knowledge*. Chicago: The University of Chicago Press.

Mayo, D. G. (1997) ‘Response to Howson and Laudan’, *Philosophy of Science*, 64(2), pp. 323–333.

Mayo, D. G. (2004) ‘An Error-Statistical Philosophy of Evidence’, in Taper, M. L. and Lele, S. R. (eds) *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago: The University of Chicago Press, pp. 79–118.

Mayo, D. G. (2005) ‘Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses’, in Achinstein, P. (ed.) *Scientific Evidence: Philosophical Theories & Applications*. The Johns Hopkins University Press, pp. 95–128.

Mayo, D. G. (2010) ‘Sins of the Epistemic Probabilist Exchanges with Peter Achinstein’, in Mayo, D. G. and Spanos, A. (eds) *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge: Cambridge University Press, pp. 189–201.

Mayo, D. G. (2018) *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Mayo, D. G. and Spanos, A. (2010) *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Edited by D. G. Mayo and A. Spanos. Cambridge: Cambridge University Press.

McKinsey, M. (1991) ‘Anti-Individualism and Privileged Access’, *Analysis*, 51(1), pp. 9–16.

Mellor, D. H. (1995) *The Facts of Causation*. London: Routledge.

National Health and Medical Research Council of Australia (NHMRC). (1999) *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines*. Canberra: Commonwealth of Australia.

Norton, J. D. (2003) ‘A Material Theory of Induction’, *Philosophy of Science*, 70(4), pp. 647–670.

Norton, J. D. (2014) ‘A Material Dissolution of the Problem of Induction’, *Synthese*, 191(4), pp. 671–690.

OCEBM Levels of Evidence Working Group (2011) *The Oxford 2011 Levels of Evidence*. Oxford Centre for Evidence-Based Medicine. Available at: http://www.cebm.net/index.aspx?o=5653 (Accessed: 24 August 2018).

Parascandola, M. (2004) ‘Two Approaches to Etiology: The Debate over Smoking and Lung Cancer in the 1950s’, *Endeavour*, 28(2), pp. 81–86.

Pearl, J. (2000) *Causality: Models, Reasoning and Inference*. cambridge: Cambridge University Press.

Potts, M. *et al.* (2006) ‘Parachute Approach to Evidence Based Medicine’, *BMJ : British Medical Journal*, 333(7570), pp. 701–703.

Prasad, S. and Galetta, S. (2009) ‘Trigeminal Neuralgia: Historical Notes and Current Concepts’, *The neurologist*, 15(2), pp. 87–94.

Quine, W. V. and Ullian, J. S. (1978) *The Web of Belief*. 2nd edn. New York: McGraw-Hill, Inc.

Ramsey, F. P. (1926) 'Truth and Probability', in Braithwaite, R. B. (ed.) *Foundations of Mathematics and Other Logical Essays*. London: Routledge and Kegan Paul, 1931, pp. 156–198.

Rawls, J. (1971) *A Theory of Justice*. Cambridge: Harvard University Press.

Rehg, W. (2009) *Cogent Science in Context: The Science Wars, Argumentation Theory, and Habermas*. London: The MIT Press.

Reiss, J. (2015a) 'A Pragmatist Theory of Evidence', *Philosophy of Science*, 82(3), pp. 341–362.

Reiss, J. (2015b) *Causation, Evidence, and Inference*. New York: Routledge.

Reiss, J. (2015c) 'Two Approaches to Reasoning from Evidence or What Econometrics Can Learn from Biomedical Research', *Journal of Economic Methodology*. Routledge, 22(3), pp. 373–390.

Robling, M. *et al.* (2016) 'Effectiveness of a Nurse-Led Intensive Home-Visitation Programme for First-Time Teenage Mothers (Building Blocks): A Pragmatic Randomised Controlled Trial', *The Lancet*, 387(10014), pp. 146–155.

Rowbottom, D. P. (2013) 'Empirical Evidence Claims Are A Priori', *Synthese*, 190(14), pp. 2821–2834.

Rowbottom, D. P. (2015) *Probability*. Cambridge: Polity.

Royall, R. (1997) *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall.

Royall, R. (2004) 'The Likelihood Paradigm for Statistical Evidence', in Taper, M. L. and Lele, S. R. (eds) *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago: The University of Chicago Press, pp. 119–152.

Russo, F. (2015) 'Causation and Correlation in Medical Science: Theoretical Problems', in Schramme, T. and Edwards, S. (eds) *Handbook of the Philosophy of Medicine*. Dordrecht: Springer, pp. 1–11.

Russo, F. and Williamson, J. (2007) 'Interpreting Causality in the Health Sciences', *International Studies in the Philosophy of Science*, 21(2), pp. 157–170.

Salmon, W. C. (1966) *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Salmon, W. C. (1989) *Four Decades of Scientific Explanatioin*. Minnesota: University of Pittsburgh Press.

Schwarz, G. (1978) 'Estimating the Dimension of a Model', *The Annals of Statistics*, 6(2), pp. 461–464.

Smith, G. C. S. and Pell, J. P. (2003) 'Parachute Use to Prevent Death and Major Trauma Related to Gravitational Challenge: Systematic Review of Randomised Controlled Trials', *BMJ (Clinical research ed.)*, 327(7429), pp. 1459–61.

Sober, E. (1988) *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge: MIT press.

Sober, E. (2001) 'Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause', *British Journal for the Philosophy of Science*, 52(2), pp. 331–346.

Sober, E. (2004) 'Likelihood, Model Selection, and the Duhem–Quine Problem', *Journal of Philosophy*, 101(5), pp. 221–241.

Sober, E. (2005) 'Is Drift a Serious Alternative to Natural Selection as an Explanation of Complex Adaptive Traits?', in O'Hear, A. (ed.) *Philosophy, Biology and Life*. Cambridge: Cambridge University Press, pp. 125–154.

Sober, E. (2008) *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Spillane, J., Higham, E. and Kullmann, D. M. (2012) 'Myasthenia gravis', *BMJ*, 345, p. e8497.

Sprenger, J. (2011) 'Hypothetico-Deductive Confirmation', *Philosophy Compass*, 6(7), pp. 497–508.

Stanford, P. K. (2011) 'Damn the Consequences: Projective Evidence and the Heterogeneity of Scientific Confirmation', *Philosophy of Science*, 78(5), pp. 887–899.

Steel, D. (2007) 'Bayesian Confirmation Theory and the Likelihood Principle', *Synthese*, 156(1), pp. 53–77.

Strevens, M. (2008) *Depth: An Account of Scientific Explanation*. London: Harvard University Press.

Thagard, P. (1998) 'Explaining Disease: Correlations, Causes, and Mechanisms', *Minds and Machines*, 8, pp. 61–78.

Trotta, R. (2008) 'Bayes in the Sky: Bayesian Inference and Model Selection in Cosmology', *Contemporary Physics*, 49(2), pp. 71–104.

Tucker, A. (2004) *Our Knowledge of the Past: A Philosophy of Historiography*. Cambridge University Press.

Urban, B. W. and Bleckwenn, M. (2002) 'Concepts and Correlations Relevant to General Anaesthesia', *British Journal of Anaesthesia*, 89(1), pp. 3–16.

U.S. Preventive Services Task Force. (2017) *Procedure Manual*. Available at: https://www.uspreventiveservicestaskforce.org/Page/Name/procedure-manual (Accessed: 9 May 2018).

van Fraassen, B. C. (1989) *Laws and Symmetry*. Oxford: Clarendon Press.
Frankish, K. (2004) *Mind and Supermind*. Cambridge: Cambridge University Press.

Viswanathan, M. *et al.* (2013) *Assessing Risk of Bias and Confounding in Observational Studies of Interventions or Exposures : Further Development of the RTI Item Bank*, *Agency for Healthcare Research and Quality (US)*.

von Mises, R. (1939) 'Über Aufteilungs – und Besetzungs – Wahrsheinlichketien', in Frank, P. et al. (eds) *Selected Papers of Richard von Mises*. Providence, R.I.: American Mathematical Society, pp. 313–334.

vos Savant, M. (1990) 'Ask Marilyn', *Parade Magazine*, p. 16.

Wasserstein, R. L. and Lazar, N. A. (2016) 'The ASA's Statement on P-Values: Context, Process, and Purpose', *The American Statistician*, 70, pp. 129–133.

Weiss, M. (2015) 'The Comparative Method', in Bowern, C. and Evans, B. (eds) *The Routledge Handbook of Historical Linguistics*. New York: Routledge, pp. 127–145.

Wilde, M. and Williamson, J. (2016) 'Evidence and Epistemic Causality', in Wiedermann, W. and Eye, A. von (eds) *Statistics and Causality: Methods for Applied Empirical Research*. New Jersey: John Wiley & Sons, pp. 31–41.

Williamson, J. (2010) *In Defence of Objective Bayesianism*, *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.

Woodward, J. (1983) 'Glymour on Theory Confirmation', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 43(1), pp. 147–152.

Woodward, J. (2011) 'Data and Phenomena: A Restatement and Defense', *Synthese*, 182(1), pp. 165–179.

Worrall, J. (2007a) 'Evidence in Medicine and Evidence-Based Medicine', *Philosophy Compass*, 2(6), pp. 981–1022.

Worrall, J. (2007b) ‘Why There’s No Cause to Randomize’, *British Journal for the Philosophy of Science*, 58(3), pp. 451–488.

Wright, C. (2000) ‘Cogency and Question-Begging: Some Reflections on McKinsey’s Paradox and Putnam’s Proof’, *Philosophical Issues*, 10, pp. 140–163.