Grading the Quality of Evidence of Mechanisms

University of Kent

Thesis submitted in Partial Fulfillment of the Requirements for Doctor in Philosophy

Stefan Dragulinescu

Table of contents

Acknowledgments

I am grateful for comments and discussion to my supervisors - David Corfield, Kristoffer Ahlstrom-Vij, Veli-Pekka Parkkinen, and Jon Williamson – and I thank them for their patience.

I would also like to thank Rachel Cooper, Federica Russo, and Anca Vasiliu for support and encouragement.

This thesis was generously funded by the Leverhulme Trust.

Parts of this thesis have been (or are to be) published in journals as follows

Mechanisms and Difference-Making, *Acta Analytica*, Volume 32, Issue 1, p. 29-54 (2017)
Inference to the best explanation as a theory for the quality of mechanistic evidence in medicine, *European Journal for Philosophy of Science*, Volume 7, Issue 2, pp. 353-372 (2017)
Inference to the Best Explanation and Mechanisms in Medicine, *Theoretical Medicine and Bioethics,* Volume 37, p. 211-232 (2016)
Medical Mechanisms and the Resilience of Probabilities, *Episteme*, forthcoming

Introduction to the thesis

It is commonplace that in epistemology and the philosophy of science, the nature of evidence in medicine has become in recent years one of the most researched topics. An important aspect of it is the evidence of *mechanisms*, and a group of philosophers of science have been urging for some time that mechanisms be included in the evaluation of causal medical claims, at a higher level than the one currently afforded by the Evidence Based Medicine (EBM) protocols (Clarke, Gillies, Illari, Russo, Williamson 2014), following the general lines of the Russo-Williamson thesis (RWT). Since the present thesis builds upon (and is highly indebted to) the research made by the abovementioned proponents of RWT, the best way to start this introduction is to present very briefly the content and purpose of RWT.

As laid down in Russo and Williamson (2007), RWT states that both evidence of *mechanisms* (taken to come from laboratory, microstructural research) and evidence of *difference-making* (taken to come from population level studies) are necessary in order to establish medical causal claims.[1] On the one hand, evidence of mechanisms is taken to have the role of eliminating spurious correlations. On the other hand, evidence of difference making coming from population-studies should establish the direction of causation and the net effect (which might not be clear just by using evidence of mechanisms) (Russo and Williamson 2007, p. 157).

For instance, to establish Helicobacter Pylori as a cause of gastric and duodenal ulcer, one needs both laboratory morpho-pathological assessments providing mechanistic evidence of the effects of this germ on the gastric and duodenal cells (which rules out that the association between Helicobacter Pylori is spurious or accidental), and population controlled studies providing evidence of difference making which establishes the direction of causation (from the respective infection to ulcer and not vice-versa) and/or the net effect (thus counting in alternative mechanisms and factors, e.g. preventative, stimulating, neutralising, possibly unknown, which might influence how strongly or decisively Helicobacter Pylori acts upon the gastric and duodenal cells).

A crucial reason for requiring this double evidence in RWT is that, on an ontic level, mechanisms are taken to be associated (only) with the so-called 'production' type of causation (Williamson, 2006, Williamson, 2011, Wilde and Williamson, 2016). Accordingly, one needs to appeal to population studies because it is only the latter that could provide evidence of difference-

[1] "the health sciences make causal claims on the basis of evidence both of physical mechanisms, and of probabilistic dependencies. Consequently, an analysis of causality solely in terms of physical mechanisms or solely in terms of probabilistic relationships, does not do justice to the causal claims of these sciences" (Russo and Williamson, 2007, p. 157).

making. Roughly speaking, production causation is causation underlined by identifiable processes holding between the cause and effect, whereas difference-making causation needs some counterfactually defined dependency between the cause and the effect (where an informal, but insightful illustration of different types of dependency is offered by the famous Mill methods of causation).

One alternative way in which RWT can then be defined is by saying that establishing causal claims in medicine requires evidence of both *production* and difference-making (where evidence of production should come from laboratory, microstructural studies, whereas evidence of difference-making should come from population-studies). This alternative formulation has the advantage of distinguishing the evidence *of what* is required by RWT (i.e. production and difference-making) and the evidence *from what* (i.e. from what sources should RWT draw out its evidence, namely laboratory, microstructural studies for the evidence of production and population studies for evidence of difference-making). In fact, as Phyllis Illari has nicely shown (Illari, 2011), distinguishing the evidence *of what* and evidence *from what,* allows one to disambiguate a certain aspect of RWT.

Illari maintains the original assumptions that mechanisms are concerned with production only, and that evidence of both production and difference making is required for medical causal claims. However, Illari argues that - as far as the sources of evidence are concerned (the *from what* part) - we could have evidence of difference making coming from laboratory, microstructural research. Analogously, we could have evidence of production (or of 'mechanisms') coming from population studies (Illari, 2011, § 2.2).

Illari's disambiguation enlarges the sphere of RWT and usefully articulates how mechanistic evidence is to contribute to the confirmation of medical hypotheses and causal theories. And there have been some further, fruitful developments for the RWT framework. Proponents of RWT (Clarke *et al.* 2014) have set up a project, named EBM+,[2] in order to deal with the epistemology of mechanisms and back up the abovementioned challenge they address to EBM, that evidence of mechanisms should be considered on an equal footing evidence of difference-making (from population studies).

One central conceptual development brought about by EBM+ has been their research on the *quality* of mechanistic evidence. The central idea behind looking into the *quality* of evidence is quite simple. The idea namely is that prior to pursuing *confirmation* studies, one needs a sort of hierarchy that would provide 'rules of thumb' differentiating between 'poor' and 'high quality' evidence, as well as degrees in-between. Organizing in this way the available mechanistic evidence does not

[2] Where the "+" means "mechanisms + trials", but also refers to the power of intersecting species of evidence. The analogy guiding the EBM+ group is that of steel-reinforced concrete, where the two materials—one good under compression, the other good under tension—mutually support each other. For an introduction, seeembplus.org.

neglect the truism that any evidence is fallible. However, it is a useful, preliminary step to take before proceeding to the confirmation stage. Moreover, obviously, one needs to get this preliminary step right. The protocols of EBM abound in various hierarchies of medical evidence. Yet, as I said, according to Clarke *et al.*, EBM fails to properly take into account the evidence of mechanisms, alongside evidence from population studies, as RWT demands. And, if one is to effectively challenge the hierarchies of EBM, proposing to integrate quality, good mechanistic evidence is a *sine qua non* condition.

Now, one slightly different angle from which one can understand the idea of the *quality* of evidence is by way of appealing to the weight/balance distinction, a traditional distinction in the philosophy of evidence (Joyce 2005, Kelly 2008, Kelly 2014, McCain and Poston 2014). This distinction can best be explained by way of an example. Suppose we have a chance set up in which the initial results have been strongly in favour of a certain outcome. That means the respective outcome has a strong *balance*. But the respective balance might well be accompanied by a small weight, because it might be that the chance set up is biased in various ways. That is why repeating the experiment, checking up or changing its methodology, scrutinizing its results, or making a different team do the same experiment, would have the consequence of increasing the *weight* of the evidence (even if the same outcome was obtained, and accordingly the *balance* of evidence remained apparently the same).[3]

To choose an example closer to our theme, a population level correlation might have a strong *balance*, and yet, for various reasons, its *weight* might turn out to be quite weak. That is because the size of the population might be too small and the other potential causal factors might not have been sufficiently screened off (say, by not choosing the right subjects in a case control or observational study, or by not randomizing and double blinding accurately in an RCT). When saying then that we need good, quality, or, if you like, *weighty* evidence, we are saying we are looking for (a large volume of) evidence that is unbiased, that is obtained using the right methodology (or ideally, using different methodologies that obtain the same results) and that that delivers precise and detailed results.

The case of mechanistic evidence is no exception. Indeed, Clarke *et al.* have provided in their (2014) a protocol that takes into account important criteria for grading mechanistic evidence: the independent methods that confirm (or disconfirm) a feature, the independent research groups that confirm (or disconfirm) a feature, the proportion of features found (larger or smaller), knowing analogous mechanisms as opposed to not knowing analogous mechanisms or even worse, knowing that analogous situations do not exhibit such mechanisms, robustness, i.e. being reproducible across a wide range of conditions, as opposed to fragility of mechanisms, i.e. not being reproducible even

[3] Of course, it could be argued that one cannot have a strong balance without strong weight, but this side of the discussion does not concern us here.

in slightly varying conditions.[4]

Pluses	***Minuses***
Each independent method that confirms a feature	Each independent method that fails to confirm—or, worse, disconfirms—a feature
Each independent research group that confirms a feature	Each independent research group that fails to confirm—or, worse, disconfirms—a feature
Larger proportion of features found	Smaller proportion of features found
Analogous mechanisms known	The analogy is a weak one, or, worse, analogous situations exhibit no such mechanism
Robust, reproducible across a wide range of conditions	Fragile, not reproducible in slightly varying conditions

The list is extensive and covers a large part of our intuitions regarding the criteria that should be employed for assessing the quality (or weight) of mechanistic evidence. Nonetheless, as the authors themselves urge, more work needs to be done. Three issues are in place here. First, on a general level, one would want an epistemological theory to *justify* these admittedly intuitive criteria for grading evidence. Second, one would want to put more flesh onto the bones of these criteria (in particular on the criterion of *robustness*) and see how these criteria work in the context of the entire medical evidence, i.e. when taking into account *also* the evidence of population studies. Third, it would be desirable to set out a plausible way in which the (quality) mechanistic evidence – hierarchized using these criteria at a pre-confirmation stage – could make a contribution at the confirmation stage itself.

The present thesis us an attempt to address these three issues. The thesis falls accordingly into three main parts (which are nevertheless interconnected since the framework and the results of each are carried over and enriched in the next).

i) The first part of the thesis addresses the first of the abovementioned issues. Thus, in chapters 1 and 2, I seek to provide the epistemic justification for the Clarke *et al.* criteria, by employing the framework of the Inference to the Best Explanation (IBE). Interpreted in causal terms, IBE says that we discover causes starting from their effects, simply because causes offer the best explanation for the existence of effects. It is an inferential theory founded by Gilbert Harman in 1965 and developed successively by Peter Lipton in his (1999) and (2004), having nowadays amongst its proponents prominent philosophers of science such as Alexander Bird (2010) and

[4]Clarke, *et al.* 2014, p. 357. Since these criteria are crucial for the present enquiry, the above list will be reproduced several times throughout this thesis, depending on the different perspective from which I am taking them on.

Stathis Psillos (2002).

Given its crucial importance for the rationale of my thesis, Inference to the Best Explanation will be described in detail in the first chapter of this thesis. Suffice to say in this Introduction that IBE is an ideal choice as an epistemic theory because, beyond its use as a theory of confirmation, it can be employed in the preliminary, pre-confirmation stages I have mentioned above, and does not depend crucially on the numerical expression (as is the case for instance with the Bayesian theory, to which otherwise it can be 'a friendly companion' – an aspect to be addressed in the subsequent chapters).

After depicting its general features and its principled use in the realm of theory confirmation in chapter 1, chapter 2 will show that IBE can provide a pattern of inference that can be employed to grade the quality of evidence and justify the Clarke *et al.* criteria of mechanistic evidence. This pattern of inference can be obtained developing and re-orienting the usage of IBE from the epistemology of *testimony,* starting from Peter Lipton's pioneering and inspiring work in this area (Lipton, 2007).

Roughly speaking, when applied to testimony as a source of evidence, IBE infers the (probable) truth/falsehood of testimonial reports, because the reported state of affairs is considered part of the causal background that determines (as an effect) the respective testimonial acts. *Mutatis mutandis*, when applied to other sources of evidence, the testimonial pattern of usage for IBE allows us to infer the probable truth of evidential reports, by taking into account both a set of commonsensical, but insightful causal principles (namely the famous set of Mill's methods, as advocated by Lipton himself) and a series of classical explanatory values (theoretical unity, simplicity, scope, and individuation), which are adopted by the large majority of IBE theorists.

Thus, as mentioned, chapter 1 describes the main outlines of the use of IBE as a theory of conformation, providing the necessary background for looking at the alternative uses of IBE, which allows one to make the transition to IBE as a theory of the *quality* of evidence. Chapter 2 begins by delineating and developing the application of IBE to testimony, and shows its direct relevance for the medical cases. It then goes on circumscribing the combination of causal principles and explanatory values to be used as a pattern of inference, which is applicable not just to testimony, but to all sources of evidence, and which ultimately, can be applied (or can do justice) to the criteria of grading evidence from Clarke *et al.,* which are thereby justified.

ii) The second part of the thesis (chapters 3-5) deals with the second of the abovementioned issues, namely that of adding more content to the backbone of the Clarke *et al.* criteria. With respect to robustness, for instance, it is certainly useful to know that a robust mechanism is reproducible in

a wide variety of conditions, as Clarke *et al* claim. However, one wonders - what is it that it is reproduced?

Some ready-made answers come easily to mind. For instance, one could say that it is the *functioning* of mechanisms that is being reproduced. But such ready-made answers call to mind other questions. What does the functioning of a mechanism consist in? We come thus to an important aspect of evaluating mechanistic evidence, namely that one needs to know or to establish what *ontically* a mechanism *is*, at least to a certain extent, in order to put more flesh to the bones of such preliminary epistemic criteria. More generally, the *epistemic* side of the discussion (*grading evidence* of mechanisms) needs to be attended by the metaphysical or *ontic* side (grading evidence of *mechanisms*).

In order to reach the largest audience, Clarke *et al.* use the broad, non-committal definition of mechanism in provided in Illari and Williamson (2012) which I mentioned above 'a mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon' (Illari and Williamson 2012, p. 120, *apud* Clarke *et al.* 2014, p.343). It is a subtle and neutral definition, which manages to capture the core of most mechanistic definitions in the literature. The purpose of adopting it would be to narrow the space for controversy over (mostly insignificant) details, and the argumentation can focus on the substantial epistemic work to be done on mechanistic evidence. However, in the second part of the thesis, I will seek to particularize this definition and make it sharper (taking of course the risk of going astray and entering into an area of controversy).

So what is a mechanism ontically, and what is the mechanistic causal relation ontically? Now, a strange feature of mechanistic accounts nowadays is that most of them (Illari and Williamson's included), while mentioning production, functioning, responsibility for events, etc. avoid the terminology of difference-making (in its counterfactual guise, as probabilistic dependency, or whatnot). Here are some well-known examples, beside Illari and Williamson's:

Illari and Williamson (2012) 'a mechanism for a phenomenon consists of entities and activities organized in such a way that they are *responsible* for the phenomenon' (Illari and Williamson 2012, p. 120, italics added)
Bechtel and Abrahamsen (2005) 'A mechanism is a structure *performing a function* in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is *responsible* for one or more phenomena.' (italics added)
Glennan (2002) 'A mechanism for a behaviour is a complex system that *produces* that behaviour by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations.' (italics added)
Machamer *et al.* (2000) [the so-called MDC account] 'Mechanisms are entities and activities organized such that they are *productive* of regular changes from start or set-up to finish or termination conditions.'(italics added)

The assumption that mechanistic causation works by production only is, as stated above, also adopted by RWT, both in the original framework of Russo and Williamson (2007), and in the revision/disambiguation made in Ilari (2011) (as well as in subsequent work by RWT proponents –

Clarke *et al.* 2014, Wilde and Williamson, 2016).

Again, as stated above, this assumption of mechanistic production by causation only goes hand in hand with a pluralistic view of evidence. This pluralistic view of evidence is inspired by the actual practice of medicine (which asks for evidence both from population studies and from laboratory, microstructural studies) and is reinforced by the position taken on mechanistic causation, with respect to evidence *of what* is said to be required by RWT from mechanisms (i.e. production). The pluralistic view of evidence is also maintained in Illari (2011) with respect to the evidence *of what* is required in RWT (i.e. evidence of mechanism/production, and evidence of difference-making) - even if, as we have seen, Illari clarifies an ambiguity concerning the *source* of evidence (the *from what* side of evidence).

However, whereas the pluralistic view of evidence is a fruitful and justified position, I think that the production-only assumption concerning mechanistic causation is wrong, for a number of reasons.

a) First, because it was triggered by two pseudo-problems from the metaphysics of causation, namely the problem of absences and the problem of pre-emption (Hall, 2004), which can be solved.

b) Second, because it forces us to separate evidence of difference making from evidence of mechanisms, which transforms RWT into a claim that we need both population studies and micro-structural, laboratory assessments in order to establish causal claims. This is quite problematic because the practice of medicine offers examples in which either population studies alone, or micro-structural evidence alone, seem sufficient to justify causal conclusion.[5]

c) Third, because it makes unavoidable ontic causal pluralism (i.e. the view that there are different types of causal relations). Indeed, if mechanistic causation is production-only causation, it is hard to imagine how the difference-making relation could be something other than a *different* type of causal relation. But ontic causal pluralism is a disaster for medical epistemology (and also for RWT), since it could not justify *why* and *how* different evidence is successfully aggregated. For instance, *why* would we necessarily need evidence of both production and difference-making when establishing causal claims, if production only could in itself constitute a full-blown causal relation?[6] One could not use here the reply that epistemically, we would need evidence of difference-making in order to differentiate processes from pseudo-processes (the former genuinely causal, the latter accidental) since, on the ontic pluralist view, difference-making just concerns a different type of

[5] This is solved by Illari (2011) but following a laborious, unnecessarily convoluted argumentation, precisely because the assumption of mechanistic causation as productive only is maintained.

[6] E.g. if Helicobacter Pylori can *produce* via a mechanism gastric ulcer, then evidence of its production should be sufficient.

causal relation.[7] Further on, how could evidence be really aggregated, if evidence point to distinct phenomena (distinct causal relations)? Different variegated evidence (as that provided in the framework of evidential pluralism) would not really reinforce the same causal claim, but refer to different causal relations and different causal claims. Hence, one central insight of RWT (drawn out of medical practice), namely that of combining evidence from population studies with laboratory evidence, seems to lose its relevance (and the rhetoric of EBM could only profit from that, since they advocate only the use of population studies - their views owing much to a tacit ontic pluralism, as I will show).

d) Fourth, because it does not just blur the reasons for why one would want to aggregate results from population studies with results from laboratory studies, but also - assuming that the problem outlined above in **c)** was somehow solved - it makes it difficult to see (or make progress on) the methodology of how they can fruitfully be aggregated and combined, given, again, the fact that we would be dealing with different causal relations and different causal claims.

e) Fifth, because it masks the role mechanisms can play in mitigating the problem of extrapolation (and, even more, it masks the problem of extrapolation itself) by requiring only evidence of *production* from laboratory studies in order to ground causal claims. It masks the problem of extrapolation because the problem of extrapolation arises due to minute differences at a micro-structural level can modify the intensity, direction and the very existence of causal relations, as reflected most conspicuously in the difference-making of these mechanistic causal relations (where the difference-making of mechanisms can define their robustness). And it masks the role mechanisms can play in mitigating the problem of extrapolation because it is by their difference-making that mechanisms can contribute to a solution.

Chapters 3-5 address and develop the above points, chapter 3 looking at points a)-c), chapter 4 looking at d), and chapter 5 looking at e). More precisely, chapter 3 proposes that mechanisms should be viewed as entailing both production *and difference-making*. The definition of mechanisms that is accordingly adopted is a modified version of Illari and Williamson's (2012) – a mechanism for a phenomenon consists of entities *joined by causal relations* that are simultaneously **productive and difference-making**, organized in such a way that the phenomenon is produced and is *dependent* upon them. I defend this construal of mechanisms against familiar, but arguably overstated counter-examples and problems, namely the problems of causation by absence and preemption, and show

[7] Vice-versa, the above argumentation could be applied to difference-making causation. Why would we necessarily need evidence of both production and difference-making when establishing causal claims, if difference-making only could in itself constitute a full-blown causal relation? One could not use here the reply that epistemically, we would need evidence of production in order to differentiate genuine dependencies from spurious correlations (the former genuinely causal, the latter accidental) since, on the ontic pluralist view, production just concerns a different type of causal relation.

that it is the best solution that can be adopted against causal pluralism.

Following this construal of mechanistic causation, RWT is to be formulated in a revised form as follows. In order to establish causal claims, one needs evidence of both production and difference-making. Evidence of production comes from laboratory studies (and, in extremely rare cases, it could also be gathered from population studies). Evidence of difference-making comes from both laboratory studies (the difference-making side of mechanisms) and from population studies. In other words, population studies and laboratory studies amount to two epistemic ways of access into the difference-making of the same causal relations. At the end of chapter 3, this revised form of RWT is defended against specific objections that have been raised by critics, most notably by Jeremy Howick and collaborators (2013).

The proposal that mechanistic causation should be viewed as entailing difference making offers at the same time the possibility to re-think the interplay between mechanisms and the population studies, and the way in which the revised RWT works. Again, the crucial notion is that of the *quality* or *weight* of evidence, and accordingly of how the quality of evidence should bear upon the way mechanisms and population studies reinforce each other's results. Thus chapter 4 shows on the one hand how evidence from population studies adds weight/quality to evidence of mechanisms, and thereby contributes to the grading of mechanistic evidence. On the other hand, it looks also at the converse aspect, showing how mechanisms could add weight/quality to population correlations and thereby contribute to the grading of population evidence. Lastly chapter 4 compares the revised RWT with the initial RWT with respect to how they handle the interplay between mechanistic evidence and evidence of population studies, and argues that the revised RWT offers insight into an additional feature of this interplay, namely how mechanistic evidence and research can individualise and define the causal factors that are taken into account by population studies.

The inferential framework of chapter 4 is, of course, that of IBE. In fact, once the construal of mechanistic causation as both productive and difference-making is adopted in chapter 3, the feasibility of using IBE in order to interpret the quality based interplay between mechanisms and population studies is even more obvious. The reason is that – as I have mentioned above in relation to the ways in which the difference making of mechanisms can be expressed – the various counterfactual expressions of the dependency between the cause and the effect parallel the informal intuitions about this same dependence, as expressed in Mill's methods. And moreover, at bottom, the interplay between mechanistic evidence and population studies evidence is the interplay between two epistemic ways of access into the difference (and production) of causal relations, such that this evidential interplay could be easily interpreted in the terms of IBE.

Differently put, whereas chapter 2 shows how the quality of mechanistic evidence can be justifiably graded using the Clarke *et al.* criteria – given an understanding of these criteria in terms of IBE and with a neutral, non-committal construal of mechanistic causation – chapter 4 moves the discussion of the quality of evidence on the level of the interplay between mechanistic evidence and evidence coming from population studies, taking on the construal of mechanisms as difference-making and interpreting the different stages of this evidential interplay as inferential moves that are justifiable on explanatory grounds.

Chapter 5 applies the reasoning and results of chapter 4 to the problem of extrapolation, extending the use of the revised RWT into this area as well, and defending this construal of extrapolation against another critique by Howick *et al,* advanced in their 2013.

The central idea is that, in the extrapolation discussions, mechanisms have been viewed as a sort of panacea, one asking from them to solve *on their own* the problem of extrapolation. However, this all or nothing strategy imposes too much burden on the mechanistic evidence, and it is likely that no account of extrapolation could solve the problem by appealing to mechanisms only. Here is where the joint use of mechanisms and population studies advocated by RWT finds a proper application. Fortified with the construal of mechanisms as difference-making, this extension of RWT to the realm of extrapolation can, for one, explain why mechanisms alone cannot be up to the task (since their difference making cannot be fully assessed just by taking into account laboratory studies). For another, it can suggest a way out of the conundrum, indicating that the joint use of population studies assessments and laboratory studies can help assessing the difference making of the mechanisms in question.

iii) The third part of the thesis (chapters 6 and 7) addresses the third of the issues highlighted in the beginning of this Introduction, namely the eventual use of the pre-confirmation hierarchization of mechanistic evidence for *confirmation* purposes. After having circumscribed and graded the high quality mechanistic evidence at the preliminary level, how can this evidence be used for the confirmation of causal claims?

Chapters 6 and 7 suggest an answer to this question starting from the results of chapter 4. The main insight of chapter 6 is that, if indeed mechanistic evidence adds weight to the results of population studies, then the friendly companionship between IBE and Bayesianism should be traced out by looking at how precisely, for confirmation purposes, the weight of evidence of population studies is increased by mechanistic evidence. It will be suggested - along the lines of a proposal made by McCain and Poston in their (2014) - that the contribution of explanatory features to the Bayesian confirmation of medical causal claims amounts to the increase of the resilience of

probability functions corresponding to population level assessments that are backed up by mechanistic evidence. One additional source of inspiration for this resilience proposal is Williamson's account of epistemic causality and his corresponding account of objective Bayesianism.

Finally, chapter 7 complements the argumentation of chapter 6, by looking at how the explanatory values employed in IBE can be objectively justified, in order to be used for constraining priors and likelihoods. Chapter 7 will thus provide more speculative justification for the use of these values, in addition to the meta-induction argument from the success of science, put forward in chapter 1. Once again, the main source of inspiration for this additional justification will be Russo and Williamson's account of epistemic causation. Although more speculative, this additional justification of the objectivity of explanatory values is meant to be stronger than the usual meta-induction argument from the success of science, which is used very frequently in the literature. If this stronger justification works, then we should have accordingly more epistemic support in using the explanatory values for the (very controversial) move of constraining priors and likelihoods. In the area of application of this thesis this move would translate as the suggestion that mechanistic evidence could be used. Finally, some other problematic (and related) aspects of the 'friendly companionship' are also discussed chapter 7 - including the issue of whether, and how, Bayesianism and IBE are distinct methods of inference.

One complementary way in which the content of the second and third parts of the thesis can be summarised is to say that, starting the from the introduction of the construal of mechanisms as difference making and of the revised RWT in chapter 3, these third and second parts of the thesis discuss successively a number of epistemic advantages that the revised RWT is meant to bring about with respect to the evidential interplay between mechanistic evidence (i.e. evidence from laboratory studies) and population studies evidence, both at the pre-confirmation, quality grading level, and at the level of confirmation and extrapolation of causal claims. These epistemic advantages could be listed as follows

I) Evidence of population studies to eliminate confounding and make manifest the difference-making of mechanisms (chapter 3)

II) Mechanistic evidence and research could help to individualise the causal factors taken into consideration by population studies (chapter 4)

III) The mechanistic evidence could increase the weight of population studies evidence and hence could contribute to the pre-confirmation grading of its quality, in a way that justified on explanatory grounds (chapter 4)

IV) The difference-making evidence from the population studies could increase the weight of mechanistic evidence and hence could contribute to the grading of its quality, in a way that justified on explanatory grounds (chapter 4)

V) The evidence of difference-making from population studies could fortify the evidence of difference-making from laboratory in order to identify robust mechanisms, which should be better prepared to face the problem of extrapolation (chapter 5)

VI) In the context of the collaboration between IBE and Bayesianism, the increase of weight brought about by mechanistic evidence could influence the resilience of probabilities functions of hypotheses established by the Bayesian theory taking into account population studies evidence (chapter 6)

VII) In the context of the collaboration between IBE and Bayesianism, mechanistic evidence could be used employed to constrain the prior and/or likelihood probabilities established by the Bayesian theory taking into account population studies evidence (chapter 7).

Note that, throughout the thesis, I will be using two (related) case studies. One will draw various examples from the history of atherosclerosis. The other will look at various treatments for hypertension and heart failure, in particular in relation to the treatment with beta-blockers and calcium-blockers.

A word is in place here about the general approach of the thesis. As it must be clear by now from this Introduction, the general approach is that of trying to offer a global picture by looking at quite diverse epistemic consequences of the revised RWT proposed here. As usual, in writing a thesis, I had the choice of either focusing on a well-delineated aspect of grading mechanistic evidence, in order to chart and explore it to the last detail, or trying to offer a plausible global picture by gathering forays into different aspects of evidence grading and subsequent hypothesis confirmation, which my view of mechanistic evidence and causation led to. With the benefit of hindsight, I should have picked out the first option, for the simple reason that it would have been easier. I was led to the second option because some of my intuitions were going against the intuitions of proponents of the initial RWT, who had already drawn a comprehensive global picture of causal assessment. In defending my views and trying to answer quite diverse, and legitimate questions, I had to try and draw such a global picture myself, at the risk, of course, of not being sufficiently detailed, of not pursuing further enough my arguments, and of maintaining the level of suggestions where one would have perhaps expected a demonstration or more powerful arguments.

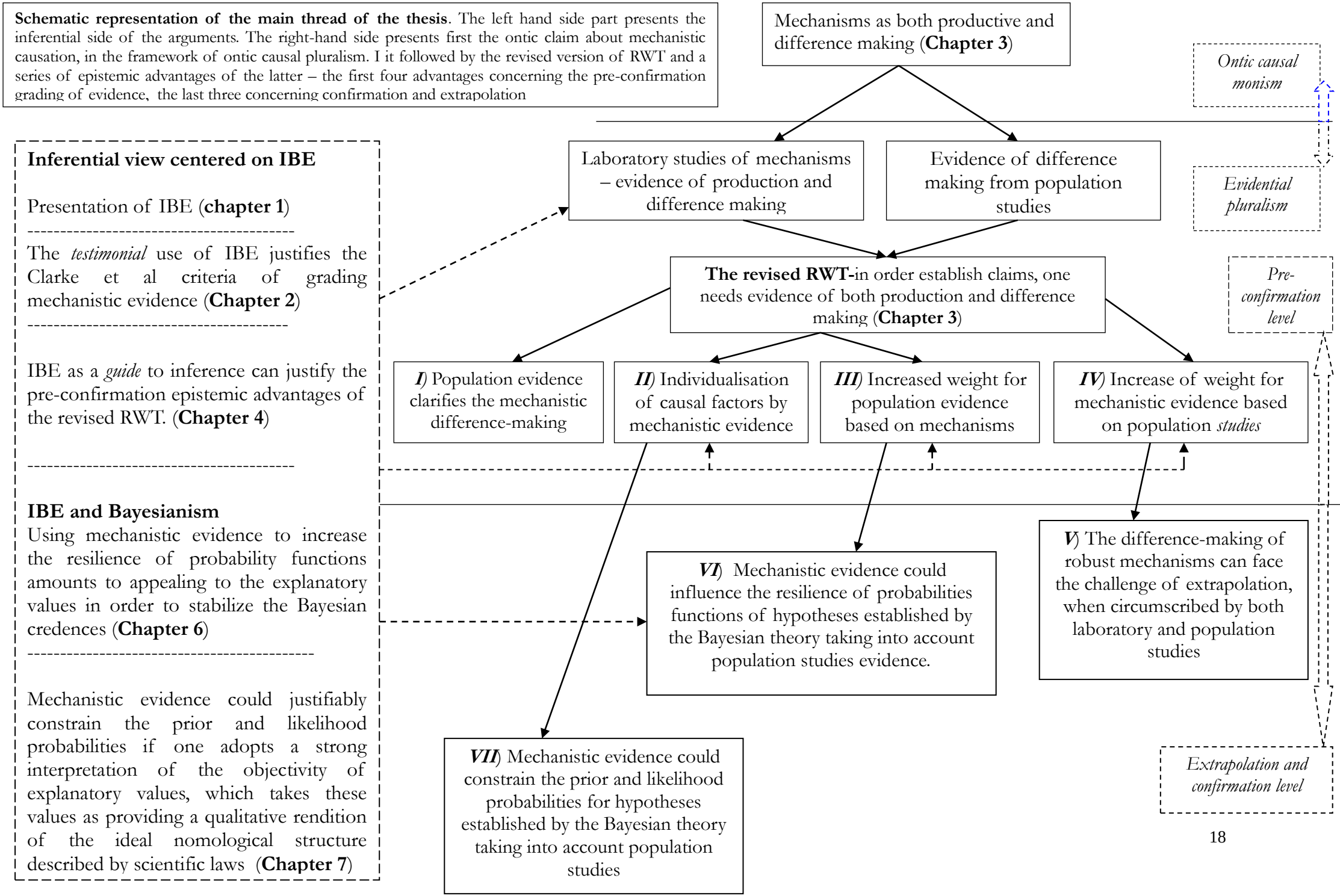
It is almost superfluous to add that the present thesis owes enormously to (and simply could not have been written without) the research done by the proponents of the initial RWT. The

criticism put forward intermittently in this thesis to some of their assumptions is only intended as a form of suggestion for potential improvement, embraces of course the caveat that the respective suggestion might be wrong, and is made, as I said, under the full awareness of the great conceptual debt owed to their research.

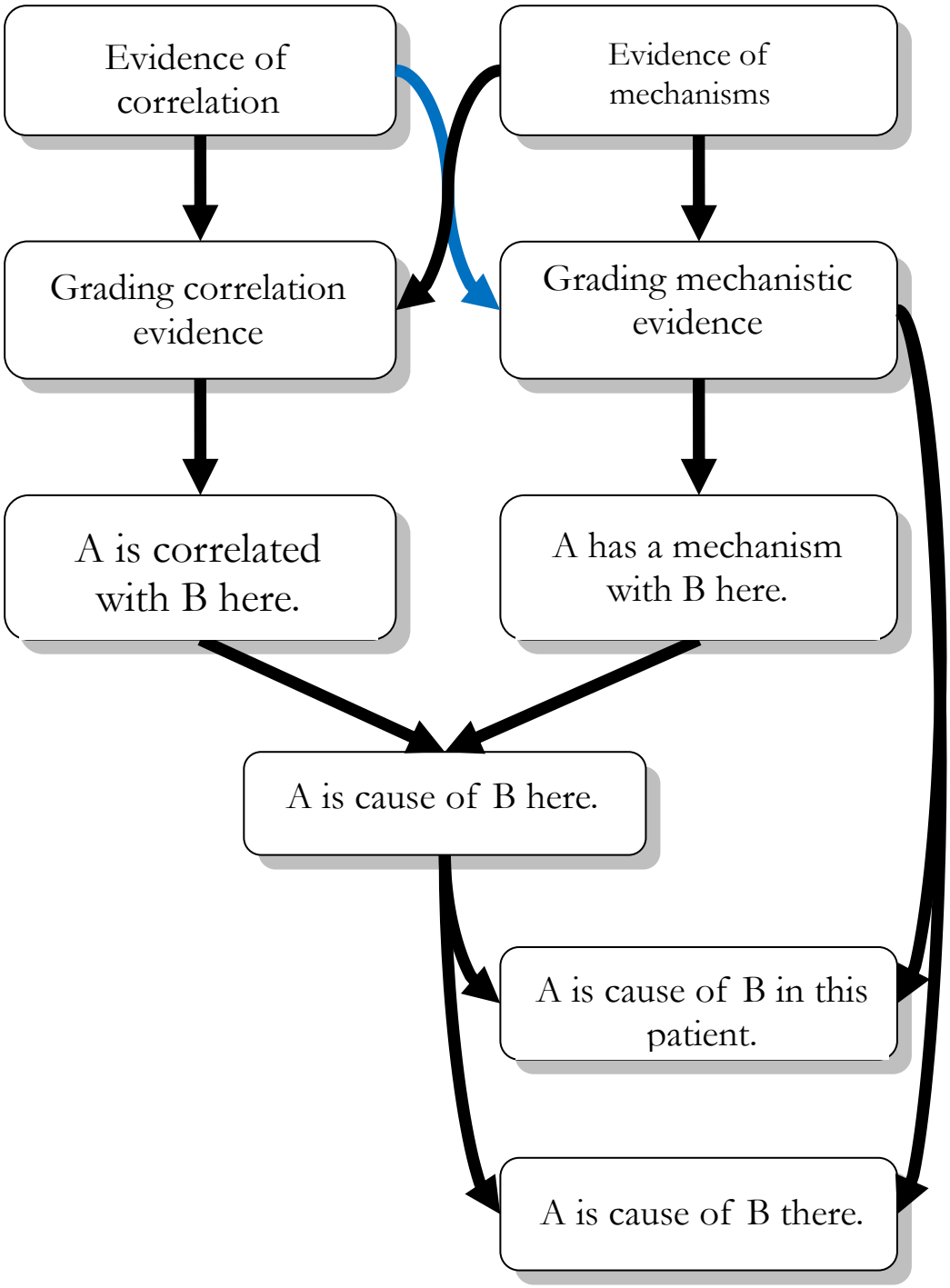
Speaking of indebtness, this final part of this Introduction has in the following two series of figures. One series presents successively the content of the subsequent chapters using as a template fig 1 from Clarke *et al* 2014 – a template used by Clarke *et al* in the framework of the initial RWT, and which I have adapted for the revised RWT. Similarly, the other series presents successively the content of each of the subsequent chapters, but using a different template, directly focused on the revised RWT and its epistemic advantages, in landscape.

The figures from the two series corresponding to each chapter will also be reproduced in the thesis before the beginning of the respective chapter. They are meant to provide a graphic, if imperfect, preview of the content of the each chapter, giving a sense of both the continuity and difference of the thesis as compared to previous work, and hopefully helping to draw the diverse aspects treated here into the global picture that was intended in discussing them.

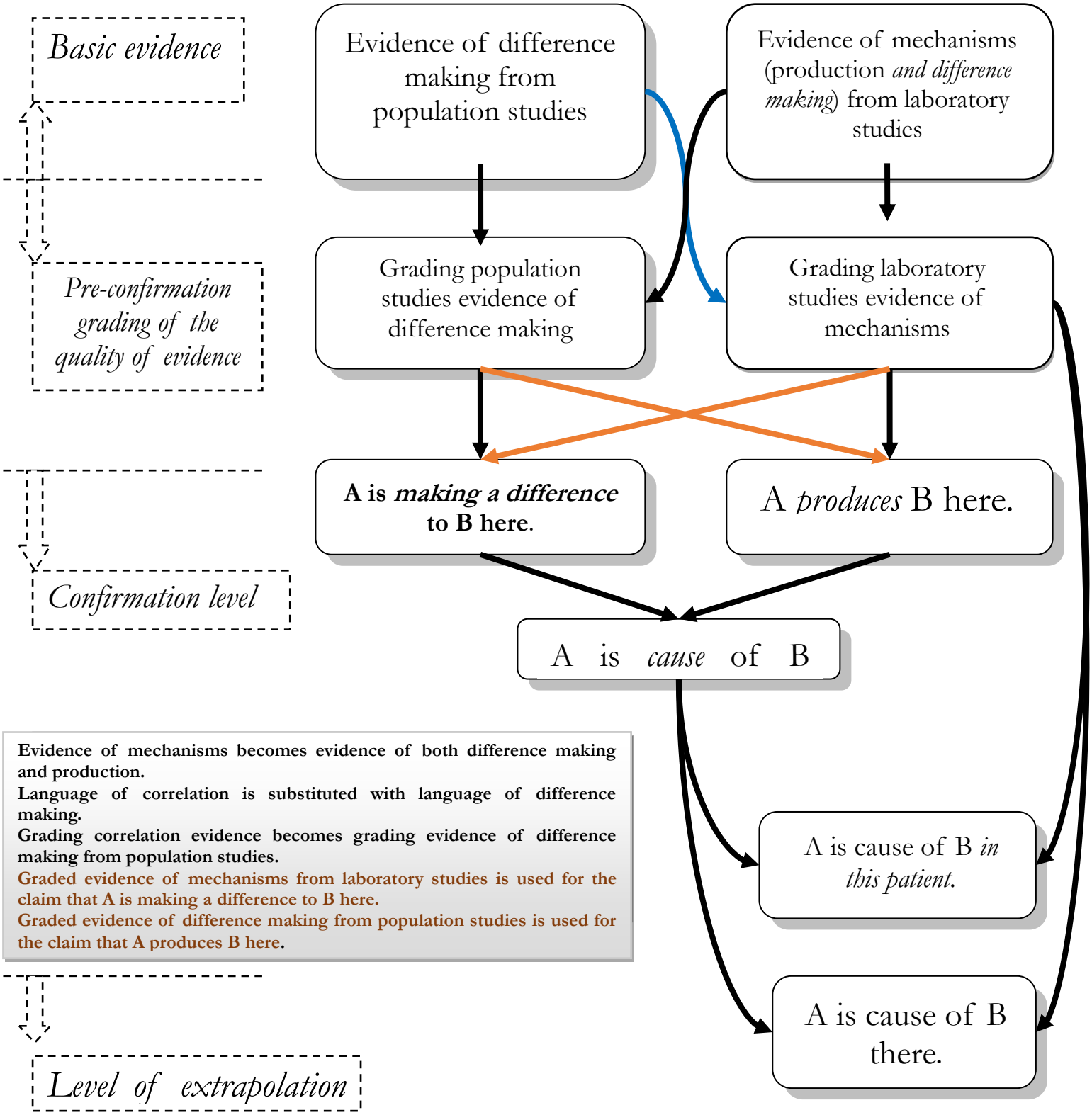
Schematic representation of the main thread of the thesis. The left hand side part presents the inferential side of the arguments. The right-hand side presents first the ontic claim about mechanistic causation, in the framework of ontic causal pluralism. I it followed by the revised version of RWT and a series of epistemic advantages of the latter – the first four advantages concerning the pre-confirmation grading of evidence, the last three concerning confirmation and extrapolation



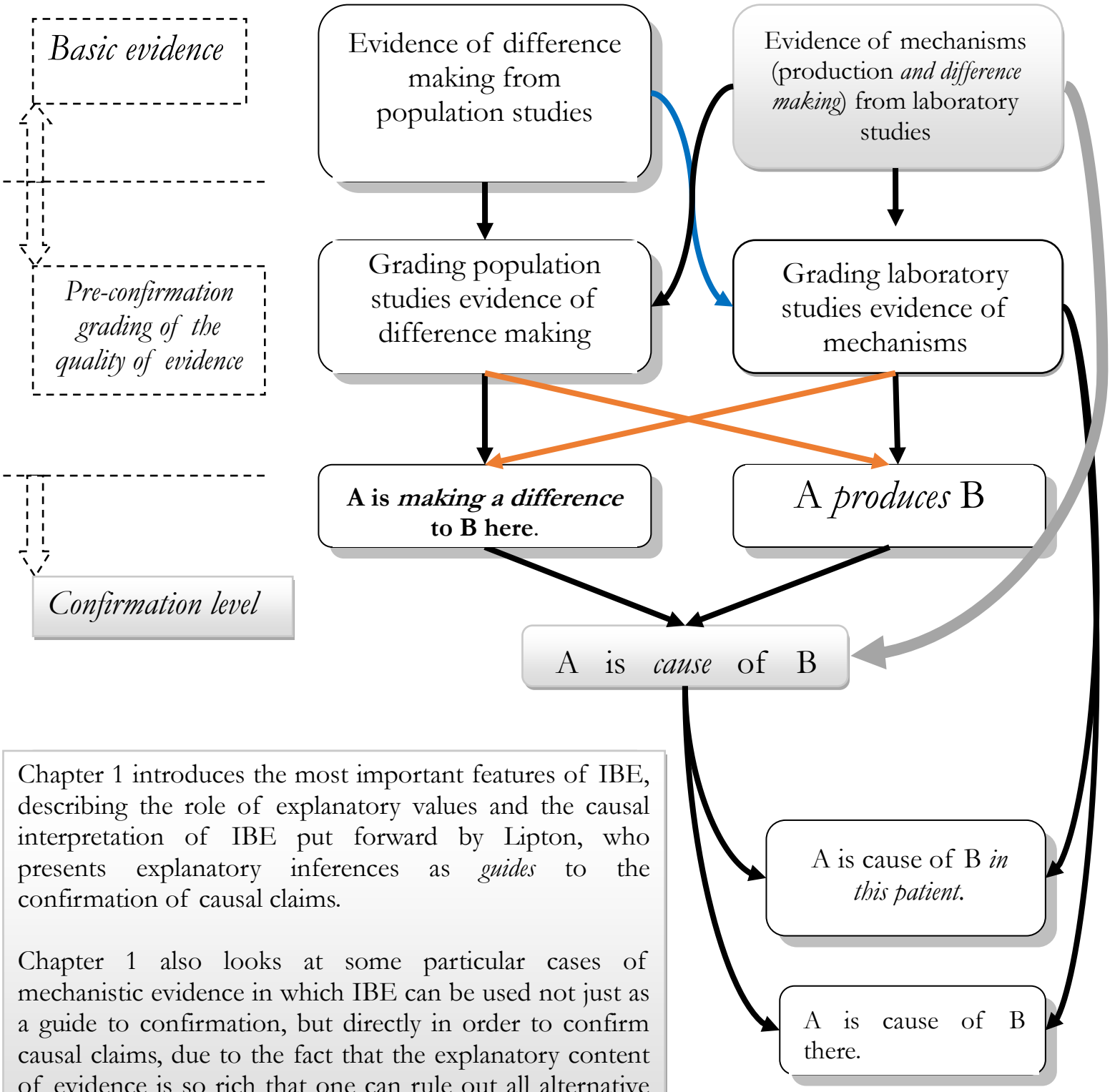
Initial figure in Clarke et al. (2014) p. 255



Overall figure for the thesis



Chapter 1



Chapter 1 introduces the most important features of IBE, describing the role of explanatory values and the causal interpretation of IBE put forward by Lipton, who presents explanatory inferences as *guides* to the confirmation of causal claims.

Chapter 1 also looks at some particular cases of mechanistic evidence in which IBE can be used not just as a guide to confirmation, but directly in order to confirm causal claims, due to the fact that the explanatory content of evidence is so rich that one can rule out all alternative explanations and pick out the right one (Inference to the Only Explanation).

Chapter 1. General lines of IBE, and its use as a theory of confirmation

Introduction

This chapter proposes to present the most important features of IBE and its main use as a theory of confirmation. The discussion in this chapter which will offer us the necessary background for making the transition in chapter 2 towards the theory of the quality of mechanistic evidence, by elaborating upon the alternative employment of IBE in testimony.

One reason why it is useful to first describe IBE as a theory of confirmation is that it makes it easier to see some of its most important advantages. These advantages are: being ampliative (i.e., amplifying, increasing knowledge), paying heed to the reliability (or weight) of evidential sources, having a close descriptive relationship with what scientists actually do, and a lack of dependence on the numerical expression.[8] This sets it apart from the Bayesian theory and makes it amenable to a wider variety of uses.

Indeed, this general presentation of IBE in its theory-confirmation use not only provides us the background for the subsequent theme of the pre-confirmation assessment of the quality of evidence, but it also has an interest *in itself*. When it comes to what theory of evidence confirmation we should choose for our causal claims in medicine (including here the causal hypotheses derived from mechanistic research) the first candidate that comes to mind is the probabilistic theory of confirmation, according to which, roughly speaking, evidence most confirms the theory whose probability it most raises, and which, in its Bayesian form, has been enormously popular for several decades. As with all theories of evidence confirmation, however, the probabilistic view has its advantages and disadvantages. With respect to mechanisms specifically, its most salient advantage in general - namely its numerical expression - does not appear as strong as usual, because in mechanistic research in medicine, the amount of data is not as comprehensive as in population studies. There is also the related problem that evidence is often expressed in qualitative, rather than quantitative terms.[9]

Saying all this is not to diminish its importance, though. Bayesianism is still an important option, the first option whenever its application is possible, and recent work suggests that its use in medical mechanisms can be very extensive and fruitful (Clarke *et al.* 2014b). But what should be done when numbers are missing or are insufficient? And is the probabilistic view the only

[8]Very recent work by Glass, Douven, Schupbach and Wenmackers aims in addition to construe IBE as a theory of confirmation*with* a numerical expression; see Glass (2012), Douven and Schupbach (2015), Douven and Schupbach (2015b), Douven and Wenmackers (2015), Douven (2016). Since this recent work is still controversial, I will not approach it here and will stick with the key features of IBE that are commonly adopted by IBE theorists.

[9] Earman 1992 has provided a full overview of the problems faced by the Bayesian approach, all the more insightful since it is provided by a Bayesian theorist.

standpoint on theory confirmation one could adopt in the intricate area of mechanism research? One response to these (slightly rhetorical) questions is simply that IBE should also be given a chance, so to speak. This type of explanatory inference is best known for its applications in the scientific realism debates, but it has also gained important proponents in the general philosophy of science as an account of scientific theory confirmation (e.g., Stathis Psillos, Alexander Bird, and Peter Lipton). To a certain extent, it has also been employed in the special sciences, medicine included,[10] although not with a focus to mechanisms.

Admittedly, IBE certainly cannot exhaust the richness of our inferential patterns in science.[11] But I will try to show that it is an insightful and interesting way of looking at the impact of evidence of mechanisms on the confirmation of causal claims in medicine, at least in certain particular cases of laboratory research, in which the qualitative side of the mechanistic evidence is prevailing (as can be seen from some examples from the history of atherosclerosis.

As for the other cases, in which not just the qualitatively rich evidence of mechanisms is in place, but also other types of evidence (in particular, the evidence from controlled studies at population level) IBE should not be viewed as a rival of the probabilistic (Bayesian) view in the realm of theory confirmation, but rather a friendly companion, whenever they might cross paths (Lipton 2004, pp. 107-117). IBE is supposed to be consistent with the probabilistic assessments, to apply, as I said, where the probabilistic results are missing or insufficient, to complement them when they are present, and to bring in some advantages of its own. But the treatment of this more complicated case of the complementarity between IBE and Bayesianism will have to wait until chapter 6.

The present chapter will present in §1 the general limes of using IBE on its own as a theory of confirmation, and will show in §2 how this confirmation use can be applied to mechanistic evidence and mechanistic hypotheses.

§ 1 General lines of IBE

IBE is an inferential method that, as its name immediately suggests, takes us from an explanandum (a phenomenon to be explained) to the truth of the explanans (the explanation)

[10] Semmelweis' discovery of the causes of puerperal fever has been a favorite example for both Lipton (2004) and Bird (2010). Bird has also provided an analysis of Bradford-Hill's criteria in terms of his own brand of IBE, in Bird (2011).

[11] In Lipton's words: 'The sensible modesty consists in making no claim that Inference to the Best Explanation is the foundation of every aspect of non-demonstrative inference.... It is glory enough to show that explanatory considerations are an important guide to inference. Consequently, there is no need to argue heroically for a perfect match between the explanatory and the inferential virtues. Similarly, in the third stage there is no need to argue that explanatory considerations are our only guide to inference, just that they are a significant guide, an important heuristic' (Lipton, 2004, p. 121).

where the latter is in the best position to account for former. As such, IBE occupies a middle ground between deduction and induction as far as the logic of inference, broadly speaking, is concerned. Analogously, it also occupies a middle ground in the more restricted area of scientific confirmation and explanation, between the hypothetico-deductive model of scientific confirmation and the covering law-model of explanation (Hempel [1948], (1970).

Let us give an easy example. On the side of the elementary logic of inference, with the famous 'Elementary, my dear Watson!', Sherlock Holmes would readily convince his companion that his reasoning was based on deduction only. Brilliant as his inferences certainly are, what his reasoning exemplifies is not deduction but IBE - inference to the best explanation. In most of the cases solved by Holmes, the possibility that someone else committed the murders is not completely ruled out. But this possibility is shown to be highly implausible, since Arthur Conan Doyle's hero would always cling on the relevant evidence and choose the hypothesis that best explains the facts (Lipton, 2004, p. 116, Lipton, 2000, p. 186).

Interestingly, on the other end of the logical spectrum of inference, even induction, when successful, can be shown to be a limit case of IBE. In our inductive practices, we generalise from the observed instances because we think that the ensuing generalisation, taken as a law or having some sort of nomic appeal, will explain the presence of such and such traits and interactions in the observed instances (Harman, 1965, pp. 90-91, Psillos 2002, p. 620). From this perspective, IBE is inter-twined with how our scientific hypotheses are generated and confirmed, and we can see this more clearly if we compare IBE with two of the grand models of confirmation and explanation from the philosophy of science.

Hempel's hypothetico-deductive model of confirmation would have it that we test our hypotheses by means of their predictions (that deductively follow from the hypotheses under test), where we are not offered any insight into how these hypotheses are devised, and, when an instance appears to disconfirm a theory, there is hardly any way to distinguish whether it is the core-theory or the auxiliary assumptions that should be dismissed. On the other hand, in the covering law model of explanation, providing an explanation is a subsequent step to the acquiring and confirming of laws, a step provided just to account for the epistemic dimension of understanding (Psillos, 2002, pp. 612-613, Lipton 2004, pp. 67, 82-83).

By contrast, in the framework of IBE, the act of explanation is already part and parcel of devising and confirming hypotheses (Lipton, 2004) and understanding is inextricably linked to the way we choose among competing theories. Such IBE practices have been shown to make up a huge part of the inferences scientists actually draw; and in medicine, recent work has shown as well its wide applications. McMullin (1992) has called IBE, in its abductive denomination "the inference

that makes science."

Importantly, IBE takes into account the *quality* of the evidence in favor of some hypothesis or another – whether, for instance, the samples used in drawing a generalisation are biased or not.[12] One way to this more clearly, is to go back to the founding article of Gilbert Harman from 1965. It was Harman who first explored the apparent contrast between simple, enumerative induction, on the one hand, and those cases of explanatory inference, typified by the Sherlock Holmes-ian conclusions ('the butler did it!'), in which one infers - from the premise that a given hypothesis would provide a 'better' explanation for the evidence than would any other hypothesis - the conclusion that the given hypothesis is true (Harman 1965, p. 89).

It was still Harman who argued that those cases of enumerative induction that are warranted are (masked) instances of IBE - where, and this is the point I wanted to insist upon - such inferences take into account the *reliability* or *quality* of the sources of evidence. Taken the enumerative induction from 'all observed As are Bs' to 'all As are Bs', and assume that we are dealing with projectable predicates. Now, one necessary condition for such an inference to have warrant is that we should have no reason to believe, for instance, that the analysed sample has been biased in the As and Bs it contains. The masked IBE that underlines cases of warranted enumerative induction is able to do so precisely because it takes into account how reliable (or, of what quality) the evidence is. When accepting on explanatory grounds that 'all As are Bs' one takes into account how reliable our samples of As and Bs are, because one thinks that this winning hypothesis ('all As are Bs') explains the evidence that 'all observed As are Bs' better than the hypotheses that, say, 'not all As are Bs' or, more directly, 'some As are Bs and the sample has been biased' (Harman, 1965 pp. 90-91). IBEs are never adequately drawn without taking into account what we know, or what we hold as true, with respect to the *quality* of the evidence. In Psillos's words, they always imply and are supported by claims about reliability or quality. This aspect of IBE is so important for the present thesis that Psillos' exposition is worth quoting here.

'The basic idea is that good inductive reasoning involves comparison of alternative potentially explanatory hypotheses. In a typical case, where the reasoning starts from the premise that 'All As in the sample are B', there are (at least) two possible ways in which the reasoning can go. The first is to withhold drawing the conclusion that 'All As are B', even if the relevant predicates are projectable, based on the claim that the observed correlation in the sample is due to the fact that the sample is biased. The second is to draw the conclusion that 'All As are B' based on the claim that that the observed correlation is due to the fact that there is a nomological connection between being A and being B such that All As are B. *This second way to reason implies (and is supported by) the claim that the observed sample is not biased. What is important in any case is that which way the reasoning should go depends on explanatory considerations. Insofar as the conclusion 'All As are B' is accepted, it is accepted on the basis it offers a better explanation of the observed frequencies of As which are B in the sample, in contrast to the (alternative potential) explanation that someone (or something) has biased the sample*. And insofar as the generalisation to the whole population is not accepted, this judgement will be based on providing reasons that the biased sample hypothesis offers a better explanation of the observed correlations in the sample. Differently put, EI [enumerative induction] is an extreme case of IBE in that a) the best explanation has the form of a nomological generalisation of the

[12] Psillos, 2002, p. 621.

data in the sample to the whole relevant population and b) the nomological generalisation is accepted, if at all, on the basis that it offers the best explanation of the observed correlations on the sample' (Psillos, 2002, pp. 620, 621, italics added).

How we come to assess an evidential source as reliable or not might well involve IBE as well, but it is very important to note that there are two *separate* issues here. One issue is how we come to confirm a certain hypothesis based on evidence, and whether one takes into account what we know of the quality of the respective evidence. A different issue is how we come to know how reliable the evidence is and how we assess the quality of evidence. An IBE theorist can always hold that the quality of evidence (no matter how it is assessed by scientists or lay people) is taken into account in the inferences for confirmation of scientific (or lay) hypotheses, without being obliged to furnish in addition any argument as to how the quality of evidence is evaluated, as such.[13]

For now, it should be noted that Harman's initial account of IBE was both too strong and too weak. Too strong because, for obvious reasons of fallibility, one should rather speak of the probable truth of the best hypothesis (Lipton, 2004); we should aim for truth but a recipe for it is impossible. Too weak because he did not say much about what criteria or explanatory values one should employ for evaluating the explanatory goodness of hypotheses. However, the explanatory goodness of hypotheses is crucial for picking the right one in the framework of IBE, much of the subsequent work has been devoted to spelling out what these criteria or explanatory values are.

The core explanatory values that are currently accepted in the IBE literature say that we should choose the theory that best fits with the relevant background knowledge (theoretical unity), that explains more evidence or the total of it (scope), makes use of fewer assumptions and theoretical entities (simplicity), and articulates a mechanism when explaining (individualisation). Here is how these values are laid down by Stathis Psillos:

Theoretical Unity: Suppose that there are two potentially explanatory hypotheses H1 and H2 but the relevant background knowledge favours H1 over H2. Unless there are specific reasons to challenge the background knowledge, H1 should be accepted as the best explanation.
Scope: Suppose that only one explanatory hypothesis H explains all data to be explained. That is, all other competing explanatory hypotheses fail to explain some of the data, although they are not refuted by them. H should be accepted as the best explanation.
Simplicity: Suppose that two composite explanatory hypotheses H1 and H2 explain all data. Suppose also that H1 uses fewer assumptions than H2. In particular, suppose that the set of hypotheses that H1 employs to explain the data is a proper subset of the hypotheses that H2 employs. Then H1 is to be preferred as a better explanation.
Individualisation: Suppose that H1 offers a more precise explanation of the phenomena than H2, in particular an explanation that articulates some causal-nomological mechanism by means of which the phenomena are explained. Then H1 is to be preferred as a better explanation.[14]

[13] Harman has also argued that how one judges various sources of evidence as to their reliability (or quality), involves, at least in part, inferences to the best explanation, where such inferences point to the truth of the evidence being furnished or provided; see Harman (1965, pp. 93-94). I will look into this aspect in chapter 2.
[14] Psillos, 2002 uses in fact different denominations for the above values (namely *consilience*, *completeness*, *parsimony* and *precision*) and adds two more, namely *importance*(doing justice to the most important parts of the evidence) and *unification* (providing a unitary explanation for the diversity of evidence. *Importance* and *unification* could be considered as sub-

There seem to be, however, two problems with the use of explanatory values in order to adjudicate which theory is more explanatory than another, namely the problem of arbitrariness and the problem of vagueness. First, there is the worry about the arbitrariness of the these values: why should nature and our theories of it be simple, theoretically unified, individuating, etc.? Despite all of the ink spilled on the subject in tackling this first problem, I can confine myself here to the remark that this is simply what science does and aims at, and why contemporary science is more successful than, say, ancient science. Anyone comparing one of Galen's therapeutic guidelines or any of the Hippocratic treatises with a contemporary medical textbook will have to acknowledge that the modern approach and the medical knowledge thereby involved are simpler and more individualised, have greater scope and greater theoretical unity, and accordingly, explain medical phenomena better than the ancient counterparts. Surely, for instance, the current classification of pulmonary diseases in terms of different pathogenic factors and morpho-pathological abnormalities does more to the systematisation of these phenomena than the classifications employed by the Galenic or Hippocratic schools in the background of their humoral theories.[15] Indeed, being actually descriptive of the advance and current practice of scientific research is one apparent feature of IBE to which I shall return.[16]

The second problem is more serious because it is easy to see that in spite of their elegant presentation provided by authors like Psillos or Lipton, how these values should be applied remains somewhat vague. In fact, they do not seem to go much beyond the suggestions for explanatory relevance that Harman gestured at in his (1965) article, declining to say more about it (Harman, 1965, p. 89).

In turn, there are two ways of escaping this problem of vagueness. One is nicely explained by Psillos - IBE is only vague when one tries to define it in an abstract way, away from the real-life situations in which it is applied (Psillos, 2007, pp. 441-447). For instance, not much can be said in general about the background knowledge that researchers possess in every particular science. This background knowledge becomes evident, however (and intimidatingly so for the outsider) when real-life examples of inference are brought forward from, say, quantum mechanics or medical microbiology. Similarly in the case of IBE and its applications, what is really important to be

---

species of (what was called above) *scope*; although important in themselves, these two values are not especially relevant for the present thesis (for which the set of core-values listed above suffices), and I will leave them aside.

[15] For a presentation of Galenic and Hippocratic views on the pulmonary pathology (and physiology), as well as the background humoral theory, see Debru (1996), Nutton (2013), and Jouanna (1992).

[16] In addition to this meta-induction argument from the success and progress of science, value epistemology (including its neo-Aristotelian tenet) has offered plausible ways of integrating these values in a normative framework (Wilkenfeld, 2014). I will offer in chapter 7 another (more speculative) justification for the use of these values in inference, based on a discussion of Lewis' approach to scientific laws.

explained, what methodologies are reliable and at what point they should be used, how to test, say, the strength of a mechanism or reach a sufficient degree of precision in its description in order to respect the value of individuation, etc., are all features that can remain vague on a general presentation but gain content when viewed from the inside of a scientific field, from the point of view of its tacit rules and practices. These rules and practices might differ in articulation from one science to another. This does not mean that they cannot in principle be spelled out, at least in part, and that one should not aim at spelling them out. An IBE theorist should try to do it, but inside a given science, by making explicit, at least in part, what is more or less tacit in a specific field.

The other way of escaping the problem of vagueness arises from Peter Lipton's work on *contrastive* explanation, which, drawing on Mill's famous causal methods, has introduced specific causation material into our explanatory reasoning. Lipton started from the insight that when seeking to explain a certain situation, we are comparing it with a foil case that resembles it as much as possible, with the difference that the explanandum does not show up. The explanation is subsequently chosen by looking at the background factors that are present (as causes) in the situation to be explained (as effect), and are absent in the foil. Naturally, hypotheses that fail to account for these background factors are eliminated. This feature of contrastivity is ingrained in our explanatory practices. As mentioned, Lipton argued that the reasoning behind such practices appeals to Mill's methods for discovering causal relations (in particular, the Method of Difference). Here is how the Methods were laid out by Mill in his *System of Logic*

a) **Direct method of Agreement**: If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon
b) **Method of Difference**: If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon.
c) **Joint Method of Agreement and Difference**: If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance; the circumstance in which alone the two sets of instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon.
d) **Method of Residue**: Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents.
e) **Method of concomitant variation:** Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation.(Mill, 2002 [1843], p. 455)

This *contrastive* type of *causal* explanation - which works by circumscribing causal factors whose presence or absence in the background do or would make a difference to the explanandum in question (in contrast to the foil situation), and thus *eliminating* spurious hypotheses - has helped enormously to put flesh on the bones of IBE because it has enriched the set of criteria based on

explanatory virtues with specifically *causal* criteria.[17]

The sceptic reader might worry why after all did Lipton appeal precisely to Mill's methods and whether the latter should be considered so significant for our inferential practices. But the worry would be unjustified. Mill's methods are basic intuitions about causal discovery and confirmation, which can be detected in the back of most contemporary accounts of causation, be they manipulative, counterfactual, probabilistic, etc., including the various forms of Humean and anti-Humean accounts.[18] Moreover, the consequence of the extremely wide application of these methods—a consequence which, cautiously, Lipton never spells out, but which is quite plausible and will be articulated further in the following chapter (starting from chapter 3) —is that the general appeal of Mill's methods (and of the explanations associated to them) comes from the fact that *difference-making* is intimately related to causal relations, at least in the medical and biological area, including here the mechanistic causation.

Speaking just in terms of plausibility, one only needs to consider the literature offering perfectly coherent accounts of causation that are universal in scope and include difference-making as a necessary condition of cause-effect relations, such as Alexander Bird's account of causal powers.[19] On the other hand, Lipton rightly notes that there is more to causation than just difference-making and Mill's methods (although the latter are an important part of it), which explains in part his reservation as to the general applicability of IBE to any sort of causal context. While explanatory considerations are relevant, they cannot represent the whole story in our inferential patterns. Moreover, given that IBE does not have a numerical, quantitative expression and is predominantly concerned with the qualitative aspects of evidence, it is not sufficiently fine-grained, in general, to really draw *confirmation* conclusions in complicated cases of hypothesis choice.

Another way to express the same idea would be to say that, in general, explanatory considerations are simply a *guide* to inferential *confirmation*. Or, in terms of a distinction Lipton also introduces, the inference to the *loveliest* explanation (i.e., the inference that pays full heed to explanatory considerations) should be a *guide* to the inference to the *likeliest* explanation (i.e., to the ultimate, warranted explanation)[20]. Apparently, Lipton was close to identifying inference to the Likeliest Explanation with a Bayesian inference that would take into account explanatory considerations (i.e. would be guided by the inference to the Loveliest Explanation) in the sense that

[17] See, for instance, Lipton (1993, pp. 39-40, 42-43; see also Bird (2007), Bird (2010), Bird (2011) and Psillos (2000), Psillos (2002), Psillos (2007). Aside from Mill's methods, Psillos and Bird also discuss the nomological side of the explanation in question, and Bird has emphasized the eliminative aspect of IBE.

[18] A comprehensive overview of Mill's methods and their significance can be found in Cartwright (1989).

[19] See Bird (2005) Bird (2007), especially his statements on the difference between the methodology of discovering causes and the metaphysics behind causation.

[20]Lipton (2004, p. 115); see also the exchange of articles and replies between Lipton and Salmon, in particular Salmon (2001) and Lipton (2001)

these explanatory considerations would have a role in the assignment of priors and likelihoods, and also in the selection of the relevant evidence (Lipton, 2004, pp.106-117).[21]Lipton's insight seems particularly relevant. The reason is that in such a joint use of IBE and of the Bayesian theory, IBE could draw on the numerical, quantitative expression afforded by Bayesian probabilities and solve the problem of making fine-grained differentiations between hypotheses, which was noted above. We will come back to this insight in chapters 6 and 7.

Up until chapter 6, I will confine myself to underlying the descriptive accuracy of IBE for scientific practice, and will mostly stick with Lipton's assessment of IBE as a *guide* to scientific inference. Moreover, up until chapter 3, I will leave aside the possible explicit articulation of mechanistic causation in terms of difference making, and will use the non-committal, general definition of mechanism provided by Phyllis Illari and Jon Williamson (2012) and also employed by Clarke *et al.*, which says that 'a mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon' (Illari and Williamson, p. 120). However, the reader should bear in mind that there is a separate argument concerning the metaphysics of causation and the definition of mechanisms *which includes difference making,* and which I will defend in chapters 3, 4 and 5; that approach to mechanisms should add further support to the conclusions of the first two chapters.

There is one last important aspect of IBE which needs to be covered in this introductory chapter. We have noted above, in connection with the guiding use of IBE, that due to its predominantly qualitative conclusions, this inferential method is not in general sufficient to really confirm hypotheses. However, there are some rare instances in which the abundant nature of evidence allows IBE to pick out the right hypotheses, because it is able, on grounds of the such explanatorily rich evidence, to rule out all alternative explanations; it can then rightly be called 'Inference to the Only Explanation' (Bird, 2009). We will look in the next section at some example of it from mechanistic research, since it illustrates at its best how Mill's methods can be used in conjunction with explanatory reasoning.

§ 2. IBE-based confirmation applied to mechanistic hypotheses

I now return to the claim that IBE is actually descriptive of the way science works, which I previously mentioned in relation to the general values of explanatory goodness (theoretical unity, scope, simplicity and individuation) and to Lipton's causal interpretation of IBE using Mill's methods.

[21] The collaboration of IBE with the Bayesian theory is a theme to which we shall return in the final two chapters of this thesis.

Indeed, much research in medicine proceeds along the lines of Mill's methods (Lipton, 2004 p. 90). Since this is a crucial part of the attractiveness of IBE for medical studies, I shall take in the following two examples from the history of atherosclerosis, a great resource for understanding how medical research unfolded in modern times, spanning as it does almost a century of investigations that incorporated all the major physio-pathological discoveries of modern medicine. Since these two examples, as well as many of my other examples from the following chapters, are drawn from these investigations, it is useful to mention briefly the milestones of this tremendous research into the causes of atherosclerosis.

The initial hypotheses taken into account at the turn of the century were that the atherosclerotic modifications of arteries would be due to protein toxicity or just amount to senescent modifications. The hypothesis regarding the pathogenic character of cholesterol -which had already been advanced in the 20s by Nikolai Anitschkow - was not taken into serious consideration until the 50s, when the full spectrum of lipoproteins was described, which was followed in the 60s by the identification of their low-density fraction (LDL) as the main carrier of cholesterol. In turn, the link between the cellular uptake of LDL and the LDL receptor was hypothesized and documented in the 70s and 80s, when the presence of foam cells inside the atherosclerotic lesions was explained in terms of macrophages taking up oxidized LDL via the famous 'scavenger' receptor.

The taking into consideration of the macrophages/monocytes, as parts of the immune system, was going hand in hand with a complementary hypothesis (detailing the mechanism of the pathogenic action of cholesterol - the 'response to injury' hypothesis. This hypothesis took the pathogenic effects of cholesterol to be augmented by the inflammation produced locally in the arteries. The hypothesis was further strengthened by a series of discoveries of inflammation-related receptors (too numerous to quote here). Suffice to say that the grand picture emerging from all this research - which is nowadays accepted but is still considered incomplete - is that the initial step of atherosclerosis consists in endothelial injury, followed by the accumulation of cholesterol in the walls of arteries and the invasion of monocytes turning into foam cells, coupled with proliferation of smooth muscle cells and local thrombus formation (Steinberg, 2007).

Now, my first, simplest example comes from the early history of atherosclerosis. As I said, around the turn of the century, one of the putative hypotheses for the atherosclerotic modifications of arteries was that they were due to protein toxicity. Indeed, initial experiments on rabbits were started in order to check this protein toxicity path. However, the diet administered during the experiments was later changed to include only the lipid component (Kritchevsky, 1995). The reason for the change in experimental diet was simply that the proteins as such *were not making any difference*

to the atherosclerotic lesions. Accordingly, the protein hypothesis was ruled out, and the explanatory grounds are easy to read. The failure of Mill's method of difference for the protein diet leads to the elimination of the protein toxicity hypothesis, as not being able to explain the atherosclerotic lesions. This is a simple, but insightful example of the use of Mill's method of difference, and of the eliminative dimension of IBE.

We can look now at a second, slightly more complicated one, which comes from the process of discovery of the LDL receptor. The discovery of the LDL receptor came about through research done into familial hypercholesterolemia (FH), and the precise modifications induced by the genetic disorder underlying it. Beginning in 1972, Joseph Goldstein and Michael Brown, two scientists trained in enzyme biochemistry, initiated a series of experiments with the working hypothesis that the high levels of cholesterol in FH might be due to a genetic disorder of HMG-CoA reductase - an enzyme with a rate-limiting effect in the synthesis of cholesterol (Steinberg, 2005). They used the cell culture technique and employed skin fibroblasts, since the use of human liver cells was very difficult (given the risks associated to liver biopsies).

Their findings showed that in normal cells, in the presence of serum, the cholesterol synthesis was low, whereas in the absence of serum, when incubated in culture medium overnight, synthesis increased almost ten-fold. The addition of LDL to the culture medium significantly reduced the synthesis. On the other hand, in FH-cells, both in the presence and in the absence of serum, the cholesterol synthesis had a high rate, *and the addition of LDL showed no inhibitory effects* (the activity of the reductase enzyme being 50 to 100-fold above normal). This seemed to lend further support to the hypothesis that feedback control by lipoproteins/cholesterol transported in the LDL-form was defective in the FH cells due to a genetic defect of the HMG-CoA reductase. However, this hypothesis was dismissed by the next experiment. Here is how Goldstein and Brown themselves describe the turning point of their research.

> The key to the receptor mechanism emerged in 1973 from studies of cells from patients with homozygous FH 8. When grown in serum containing lipoproteins, the homozygous FH cells had HMG CoA reductase activities that were 50 to 100-fold above normal. This activity did not increase significantly when the lipoproteins were removed from the serum, and there was no suppression when LDL was added back. The simplest interpretation of these results was that FH homozygotes have a defect in the gene encoding HMG CoA reductase that renders the enzyme resistant to feedback regulation by LDL-derived cholesterol. This working hypothesis was immediately disproved by our next experiment. We delivered cholesterol in ethanol instead of in LDL. When mixed with albumin containing solutions, cholesterol forms a quasi-soluble emulsion that enters cells by adsorption to the plasma membrane. When cholesterol was added in this form, the HMG CoA reductase activities of normal and FH homozygote fibroblasts were equally suppressed. Clearly, the defect in the FH homozygote cells must reside in their ability to extract cholesterol from the lipoprotein, and not in the ability of the cholesterol, once extracted by the cells, to act. But how do normal cells extract the cholesterol of LDL? The high affinity action of LDL suggested that a cell surface receptor was involved. (Brown and Goldstein 2009, p. 433)

This crucial experiment showed that the presence of cholesterol not in the LDL-transported

form *does* have an inhibiting effect on these cells, which meant in turn that the feedback response to cholesterol could be affected inside the cells and that the activity of the reductase enzyme could be down-regulated. The implication was that, when offered as part of LDL, cholesterol simply does not get into cells, indicating a missing receptor, and they managed to clone the cell in the following years, receiving the Nobel Prize in 1985.

Again, the tacit use of Mill's methods is clearly visible in their experiments. For instance, the method of *difference* is saliently in place in the experiments comparing the rates of activity of the enzyme when LDLs are present or absent in serum and the culture medium of fibroblasts (firstly, in normal cells, and secondly, in FH-cells). When cholesterol *per se* was added in the serum of FH-fibroblasts and the activity of the enzyme was finally decreased, the two scientists applied the method of *agreement* to test the hypothesis that there is a genetic defect impeding the feedback down-regulation of the enzyme in question by cholesterol. If such a genetic defect had been in place, they reasoned, it should have manifested itself across different situations, including the situation in which cholesterol is directly adsorbed into the cell, in a quasi-soluble emulsion with ethanol and albumin-containing solutions.

Furthermore, the reliability of the sources of evidence was obviously taken into account in the respective explanatory inference. To put it simply, the two scientists would not have used unreliable evidence, and their findings would not have been accepted had they based their experiments on unreliable materials or methods. They used the cell culture technique, which had been in place, with encouraging results, for almost two decades. Given the impossibility of working on human liver cells, they turned to skin fibroblasts. Among the reasons for choosing skin fibroblasts was the fact that the patients under study were suffering from a homozygous genetic disorder, and several metabolic diseases (such as galactosemia and the Lesch-Nyhan syndrome) with a similar homozygous genetic background had been elucidated by working with skin fibroblasts. Doubtless, numerous other reliability conditions were involved in the laboratory research, conditions specific for this branch of biological and medical science, and which could not be (easily) captured by some sort of algorithm or context free protocol. We need not worry here about how precisely they assessed the reliability of evidence. As mentioned in the previous section, taking into account the quality of evidence when inferring causes or drawing best explanations, on the one hand, and providing reasons why a certain source of evidence is reliable or not, on the other hand, are distinct issues, to which I will look in the next chapter.

Summing up, in the discovery of the LDL receptor, there were two candidate hypotheses to *explain* the production of cholesterol in FH patients:

(1) The high production of cholesterol (in patients with familial hypercholesterolemia (FH) is due to a gene defect (in the gene encoding HMG CoA reductase) that makes cells resistant to feedback regulation by LDL cholesterol (more precisely, that renders the HMG CoA reductase resistant to feedback regulation by LDL-derived cholesterol).
(2) The high production of cholesterol (in patients with familial hypercholesterolemia (FH) is due to the lack of a receptor (for the LDL cholesterol) that makes cells resistant to feedback regulation for the reason that LDL cholesterol cannot get into the cells.

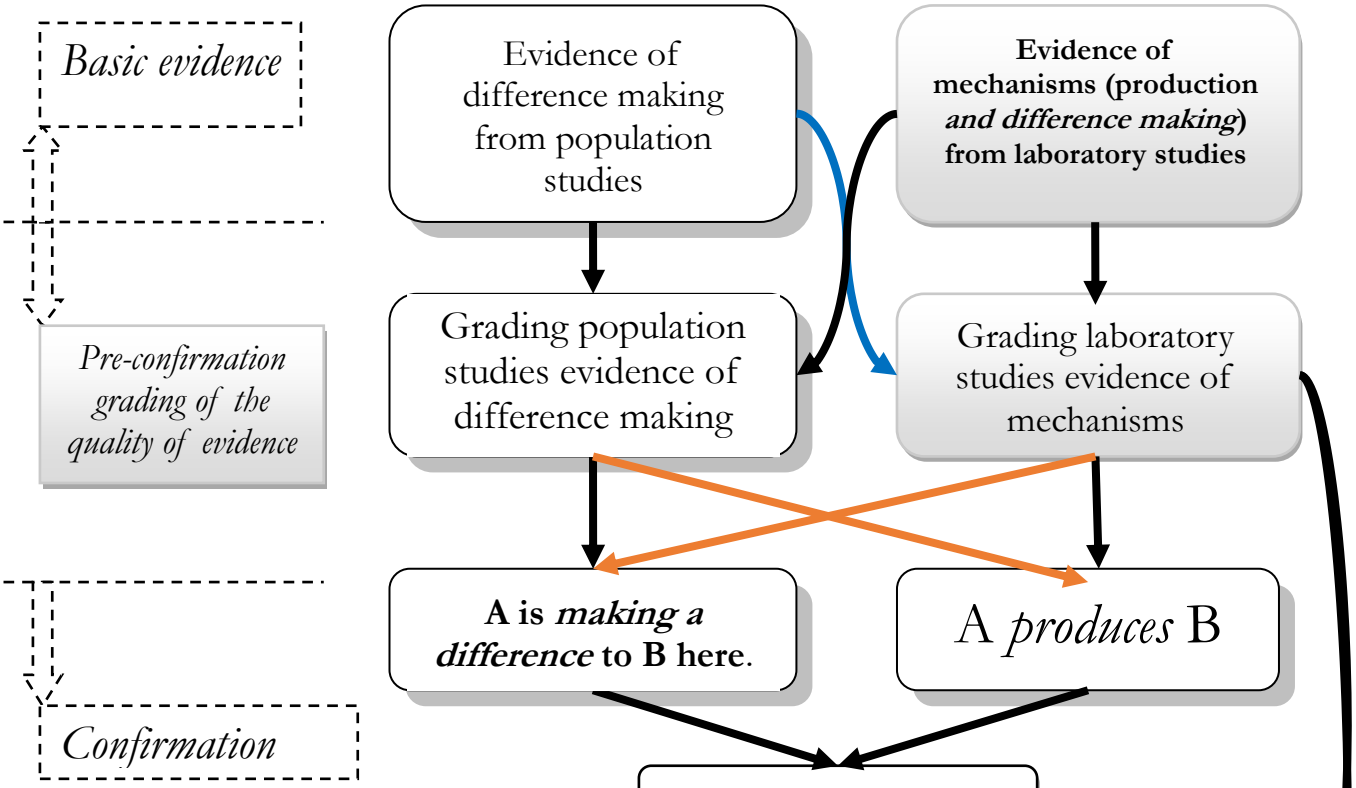
Hypothesis (1) was dismissed, as it did not pass the test of Mill's method of agreement and method of difference, and hypothesis (2) was accepted in an *explanatory* inferential process that also took into account the reliability of evidence, and which was further confirmed by the cloning of the receptor.

Such examples show convincingly how IBE proceeds to confirmation by eliminating alternative hypotheses in laboratory research, on a microstructural level. Nevertheless, there are more extensive uses of IBE to which I turn in the next chapter.

Conclusion chapter 1

In this chapter I have laid down the main features of IBE and have described its use as a theory of confirmation, providing examples from medical mechanistic research. This chapter has provided the necessary background that will allow us to make the transition in chapter 2 towards the theory of the quality or weight of evidence by elaborating upon the employment of IBE in testimony.

Chapter 2



Chapter 2 looks at the pre-confirmation level of grading the quality of mechanistic evidence.

It shows that the testimonial use of IBE provides a pattern of justification that can warrant the Clarke et al. criteria of grading the evidence of mechanisms.

Chapter 2. IBE and the quality or weight of evidence

Introduction

I have presented in the first chapter the use of IBE as a guide to confirmation for certain cases of mechanistic medical research. There are, however, more uses of IBE. More precisely, IBE can be used as an epistemological theory of testimony and, importantly, as a means of categorising and justifying the *sources* of evidence. In short, beside the use as a theory of confirmation, IBE can be employed as an epistemological theory for the *quality* or *weight* of evidence. Important traces of this use can be found in Lipton's inevitably rich and inspiring treatment of IBE, more precisely in his discussion of how IBE should be used alongside the Bayesian theory of confirmation, and also in his treatment of how it can be put to work in the epistemology of testimony (Lipton, 2004, pp. 103-126).[22]

Based on these alternative uses, I will show that IBE can *justify* the criteria of grading quality of mechanistic evidence that have been recently provided by Clarke *et al.* in their (2014) contribution to how evidence of mechanisms is to be construed alongside population studies.

As mentioned in the Introduction to this thesis, the criteria proposed by Clarke *et al.* take into account important elements such as the robustness of mechanisms, the proportion of elements found, the research groups, and the methodologies involved in mechanistic research; and they provide a perfect illustration for why we need to look into the quality of evidence. The central idea behind putting forward such criteria of such criteria is that prior to pursuing *confirmation* studies, one needs a sort of hierarchy that would provide 'rules of thumb' differentiating between 'poor' and 'quality' evidence, as well as degrees in-between. A theory of the quality of evidence should be able to provide such 'rules of thumb' even when (or especially in the circumstances in which) we do not have numerical expressions of the probabilistic dependence between the evidence and the

[22] To recall from the previous chapter, Lipton distinguishes between what he calls 'Inference to the Likeliest Explanation' and 'Inference to the Loveliest Explanation'. Roughly speaking, 'Inference to the Likeliest Explanation' is concerned with confirmation, and Lipton thought it could be identified with a Bayesian inference backed up by explanatory consideration. On the other hand, 'Inference to the Loveliest Explanation' constitutes the proper use of IBE in finding the relevant evidence, in pointing to the most fruitful and promising hypotheses, in a preliminary stage, before one reaches, or takes into account, the level of confirmation. Hence, IBE as Inference to the Loveliest Explanation could be used, in collaboration with a Bayesian approach, to approximate the prior probabilities of hypotheses to be confirmed (see Lipton, 2001, p. 22, and more recently, McCain and Poston, 2014). In a sense, my enquiry in the present chapter could be portrayed as an attempt to refine the use the Inference to the Loveliest Explanation as a theory for relevant evidence, and for grading the quality of relevant evidence. However, for the purpose of clarity of exposition, I will rather pursue the way opened by Lipton's work on testimony, since it offers a more direct access to the level of evidence and makes it easier to keep apart the concepts of confirmation and that of the pre-confirmation assessment of the quality of evidence.

hypotheses at stake, or of the probability attached to evidence itself.[23]It should be said that organizing in this way the available evidence does not neglect the truism that any evidence is fallible; but it is a necessary, preliminary step before proceeding to the confirmation stage. In Clarke *et al* words:

'In the paper up to this point, we have made the case that evidence of mechanisms can usefully supplement evidence of correlation. Of course, all evidence and all conclusions reached in medicine are fallible. Evidence of correlation is fallible; evidence of mechanisms is fallible; and conclusions drawn from that evidence are fallible. That is the nature of any science. We focus here on the important point that we can get varying quality of evidence of mechanisms, just as we can get varying quality of evidence of correlation. We have been pressing the point that this kind of variation in quality of evidence of mechanisms needs a great deal more attention—indeed, it needs just as much attention as quality of evidence of correlation. Here we make a very preliminary attempt to lay out some ways in which evidence of mechanisms may be graded. We acknowledge that much more work will need to be done in this regard." (Clarke et. al. 2014, p. 358)

Obviously, more work needs to be done, in particular with regard to the criterion of *robustness*. But first of all, by and large, one needs to make sure*to get* this preliminary, pre-confirmation step*right*. The protocols of Evidence Based Medicine (EBM) abound in various hierarchies of medical evidence, but, according to Clarke *et al.*, they fail to properly take into account the evidence of mechanisms, *alongside* evidence from population studies, as Russo and Williamson have maintained in formulating the Russo-Williamson thesis (RWT).[24] I agree with RWT and I will defend it against criticism in the next chapter, in which I will also look at the problem of combining population studies with mechanistic evidence. The preliminary question I seek to respond in the present chapter is: did Clarke *et al.* get their criteria for quality of mechanistic evidence right?[25]

I think they did, and I provide here a justification of these criteria in the framework of IBE. The main argument of this chapter will be that Lipton's framework of testimonial uses of IBE, when properly augmented and clarified, can be extended to deal with other sources of evidence beyond testimony as such, and thus provide a pattern of justificatory inference that perfectly underlies and maps the criteria of mechanistic evidence in question.

The plan for this chapter is as follows. §1 will explain the difference between the use of IBE as a theory of confirmation, and its pre-confirmation use as a means of evaluating the *quality* of evidence. §2 will look at the specific use of evaluating the quality of evidence, treating the case of *testimony,* and focusing on the model of testimonial application of IBE provided in Lipton (2007). §3

[23] It is for this reason that I think IBE is an interesting path to investigate, even if, given the tremendous popularity of Bayesianism (as a theory of confirmation) one would prima facie consider it also as a candidate for assessing the quality of evidence. We have in fact at disposal sophisticated Bayesian accounts of testimony, (e.g. Bovens and Hartmann, 2003) and I will mention in footnotes the conceptual support that my approach to IBE could gather from the Bayesian perspective

[24] See for instance, the Oxford criteria of the levels of medical evidence, in which the evidence of mechanisms is placed on the bottom of the scale. http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/

[25] Of course, one can discuss the latter question without going into the former.

will deal with IBE and the quality of the *sources of evidence,* taking into account different aspects of evidence ranging from the methodology of research, to the number of different research teams having the same research objective. It will accordingly show how IBE can be the epistemological underpinning for the criteria of mechanistic evidence provided in Clarke *et al.* 2014.

§1 Inference to the Best Explanation – confirmation versus pre-confirmation assessment of the quality of evidence

As a quick reminder from the previous chapter, IBE is a method of inference that, as its name immediately suggests, takes us from an explanandum to the truth of the explanans that is in the best position to account for the phenomenon to be explained. The explanatory power of various hypotheses is to be cashed out, on the one hand, in virtue of explanatory virtues. These are *i.* simplicity, *ii.* theoretic unity, *iii.* scope, and *iv.* individualisation - which correspond accordingly to, *i.* fewer assumptions being used, *ii.* congruence with previous theories, *iii.* doing justice to a large area of the phenomena in question, and *iv.* providing a mechanism or some fine grained description for the processes at stake (Psillos, 2002, pp. 615-616). On the other hand, as Peter Lipton has shown, one can also use certain causal criteria, such as Mill's methods or principles for adjudicating among candidate causal factors, among which of prime importance are the principles of *i.* Agreement, *ii.* Difference, and *iii.* Concomitant variation. These principle say, respectively, *i.* that in varying contexts and backgrounds with the same effect displayed, what is common to these varying contexts is indicative of a cause, *ii.* that in similar contexts with different effects displayed what differs in the context or background is indicative of the cause, and *iii.* that factors in the background that are shown to co-vary with the effects displayed are indicative of the cause.

Importantly, I have shown in the previous chapter that IBE takes into account the *quality*or *weight* of the evidence in favor of some hypothesis or another – whether, for instance, the testimony involved in a special science testimony is trustworthy, or, say, the samples used in drawing a generalisation are biased or not (Psillos, 2002, p, 621). Now, let us stick to the last feature just mentioned – *quality* or *weight* of evidence. That IBE takes into account the quality of evidence, when confirming or devising hypotheses is one thing; how exactly the *quality* of evidence is ascertained, is another thing. An IBE theorist can always hold that the quality of evidence (no matter how it is arrived at by scientists or lay people) is taken into account in the inferences for confirmation of scientific hypotheses, without being obliged to furnish in addition any statement as to how its quality is evaluated, as such. But in fact, IBE has resources to deal also with how the quality of evidence is ascertained, starting from testimony and expert knowledge, to other forms of providing

evidence.

Actually, when laying out the foundation of IBE as a general theory of inference in his seminal 1965 article, Gilbert Harman has also proposed a preliminary framework for the application of IBE in the case of testimony. Harman was contrasting testimonial explanatory inference with simple inductive reasoning, which, in a Humean manner, would infer the truth of particular instances of testimony from the correlation between past instances of testimonial acts coming from a similar type of person about a similar subject matter, provided in a certain manner, on the one hand, and actual states of affairs, on the other (which is, in fact, a (crude) form of the Humean reductionist approach to testimony, as we shall see later). According to Harman, such Humean inferences could not justify why is it that we cannot gain knowledge from testimony even if a true report is believed, in those contexts in which the truth of such a report is based on accidental features such as a misprint (for a written testimony) or a slip of the tongue (for an oral report).

> [if] in the first example we think of ourselves as using enumerative induction, then it seems in principle possible to state all the relevant evidence in statements about the correlation between (on the one hand) testimony of a certain type of person about a certain subject matter, where this testimony is given in a certain manner, and (on the other hand) the truth of that testimony. Our inference appears to be completely described by saying that we infer from the correlation between testimony and truth in the past to the correlation in the present case. But, as we have seen, this is not a satisfactory account of the inference which actually does back up our knowledge, since this account cannot explain the essential relevance of whether or not there is a slip of the tongue or a misprint." (Harman, 1965, pp. 93-94)

Harman's argument was that the epistemic warrant for testimonial knowledge in such cases cannot be offered on merely Humean grounds. One needs explanatory inferences that draw on various contextual elements (such as, for instance, the absence of a slip of the tongue or typewriter) - which are out of place in a strictly enumerative, Humean induction, but are essential in an abductive framework which seeks to find the best explanation for why a testimony was put forth in the first place. Why is it that the absence or presence of a slip of the tongue or typewriter/keyboard, is essentially important for the truth of a certain testimonial evidence? Because it is crucially involved in answering a different question - why is it that a speaker has stated such and such alleged facts? And in responding, roughly speaking 'because these are the facts, and s/he is telling the truth, and her testimony is true' we take into account, amongst other aspects, the lack of a slip of the tongue (and not, say, how the person in question was dressed), and we offer an explanation as to the very enunciation of the testimony.

Harman's indications that IBE could also be used to discuss quality of evidence as such, has opened the way to more elaborate formulations, the most important being Peter Lipton's explanatory account of testimony, to which I turn next. Indeed, Lipton has delineated a so-called 'Testimonial Inference to the Best Explanation' (henceforth TIBE) as an application of IBE to the

issue of testimony. According to Lipton, TIBE says that we infer the truth of the testimony in those cases where the truth best explains the utterance of the expression of the testimony by other means, and we reject testimony when its falseness is the best explanation for the respective utterance (whereas IBE says that we infer the truth of the hypothesis that best explains the facts under scrutiny).

"TIBE applies the general IBE scheme of inference to the particular task of assessing testimony. It is a distinctively a form of IBE, because it has it that testimonial inference is an inference to the best explanation of the relevant evidence. And it is distinctive from other forms of IBE, because it takes the central datum to be explained not to be some natural phenomenon such as red-shift of galactic light, but rather the fact that the speaker said what she did. It is an abductive inference from the fact of utterance to the fact uttered. The governing idea of TIBE is that when we are in evaluative mode, we infer that what we are told is true when its truth is part of the best explanation of the fact that the speaker said it..... the best criterion for the assessment of testimony is whether or not the production of the testimony is best explained by an account which implies its truth [...] TIBE also accounts for the diversity of factors that enter into testimonial inferences,. including facts about speakers, the content of what they say and the manner and context in which they say it [....] the evidence on which we base our decision whether to believe what we are told is highly diverse." (Lipton, 2007, pp. 243-245)

We have seen in §1 of the previous chapter that IBE adopts a middle way among various theories of inference, confirmation and explanation. Analogously, in the context of the larger epistemological discussions on testimony in the literature, TIBE adopts a nuanced, middle way view between two great families of approaches to testimony, usually taken to originate from Hume and Reid, namely the reductionist and the non-reductionist approaches.[26] Just to give a brief overview, according to the former, testimony is to be reduced to other types of evidence (perceptive, memory-derived) and is to be justified by inferential processes that account for the truth of testimony (or for the testimonial beliefs being justified or amounting to testimonial knowledge) on grounds that are dependent on the content of the testimonial report as such and the potential reliability of the speaker (e.g. sincerity and competence of the testifier, previous similar testimonies that were shown to be true, various types of track-records, etc.), in line, more or less, with Hume's injunction that our epistemic warrant for testimony is derived from 'our observation of the veracity of human testimony, and of the usual conformity of facts to the reports of witnesses'.[27] On the other hand, non-reductionist approaches advocate a default direct acceptance of the testimony (justified by various pragmatic, social and ethical principles) provided there are no defeaters, i.e.

[26] Gelfert 2010, Lipton 2007. From a different perspective than Lipton's, Lackey (2006) and Lackey (2008) also advocates a middle way between the reductionist and non-reductionist approaches; as the title of her (2006) article colorfully says 'It Takes Two to Tango'. Graham (2006) accepts that reductionism and non-reductionism are not incompatible positions, but does not go on the way of seeing a fruitful collaboration between assumptions belonging to both parties, arguing instead that justification of testimonial beliefs can be overdetermined.

[27] Hume 1977 [1748], 74. There are important differences within the camp of reductionism (e.g., general vs. local reductionism), and the positions of proponents of this view should of course not be assimilated to mere pastiches of Hume.

provided one has no reasons to doubt the credibility of the author of testimony.[28]

And here is the middle way adopted by TIBE. Like the non-reductionist approach, TIBE accepts that our default stance might well be direct acceptance. But it is also underlined that in trigger cases, where doubt shows up, or doubt should exist (as for instance in scientific contexts) an explanatory inferential process takes place or should take place. This inferential process takes into account the content of testimony, but also the manner and context of the testimonial report,[29] the reliability and track record of the speaker,[30] etc. In other words, in such trigger cases, like the reductionist approach, TIBE places an accent on our inferential practices, and seeks to bring into focus the content of testimony. In Lipton's words 'These sorts of account [non-reductionist ones] make credibility remarkably independent of the content of the testimony. What counts is who says it, not what is said.[...] In any event, it is clear that the decision whether to believe someone depends not just on who they are but also on what they say and how what they say fits with what the audience already accepts. The central question about testimony is not just whom to trust, but *what* to believe.' (Lipton, 1998, p. 14, italics added). Explanatory inferences to the best explanation are required to account precisely for this content, *but also* to take into account the other aspects related to testimony (the track record of the previous such assertions and the reliability of the speaker), albeit in a secondary way.

Truth then, should figure in the explanation of the testimonial utterance or expression in those cases in which testimony is accepted, believed, or counts as testimonial knowledge. What sort of truth figures in the respective explanation can also vary. It can be the first-order truth of the belief expressed in the testimony of the speaker (It is true that it is raining outside) or the second order truth referring to the veridical nature of the report (it is true that the speaker believes that it is raining outside). One can move from the second-order truth to the first-order one by using also TIBE (it is true that it is raining outside because it is the best explanation why the speaker believes that it is raining outside).[31]

§2. TIBE applied to medical testimony

[28] See for instance Adler, 2012, Audi, 1997, Perrine 2014.

[29]"What is to be explained is often not just that the speaker said what she said, but that *she* said it, and that she said it in the way that she did, for example in a way of exaggerated earnestness." Lipton 2007, p. 245.

[30] "There will also be evidence that has nothing to do with either the speaker or her present utterance. Thus the fact that she has been so reliable on these matters in the past encourages me to trust her this time." Lipton 2007, p. 245.

[31]Harman 1965, p. 89. One of Lackey's most important contributions is that testimonial inference could well go direct to (what I have called) first order truth, without going through (what I have called) second order truth; this is, I think, consistent with an explanatory framework; see Lackey (2008). However, I am inclined to doubt that in the scientific contexts with which this chapter is concerned, one could really dispense with the second-order truth as to the author of testimony believing the facts s/he is reporting.

Now, as I mentioned, Lipton's scheme can be readily applied to medical cases, the reports on mechanistic claims included. Here is one medical example in which the main dimensions of TIBE are put to work - the clinical report of the usefulness of beta-blockers for improving cardiac failure from Waagstein *et al.* 1975, which was very important in the prestigious history of beta-blockers, and played a role in the revolutionary idea that they can be used in patients with low cardiac index. Waagstein *et al.* reported that beta-blockers (alprenolol and proctolol) administered to a group of seven patients with cardiac failure (the basic diagnosis being congestive cardiomyopathy)[32] in order to control the cardiac rhythm, *also* improved their condition (better physical working capacity, reduced heart size, better ventricular function; Waagstein, *et al.* 1975). It is not a purely mechanistic report but is embedded in mechanistic evidence - as should be clear by looking at all the investigations pointed out in the report, as well as the relation to animal experimentation, described below.

What form would TIBE take in this case? In the most schematic form, it would be an inference from Waagstein *et al*'s report to the truth of the evidential claim that the administration of beta blockers was followed by the improvement the cardiac index of the respective patients, where this being true is the best explanation for the fact that the report has been put forward. But it is important to see beyond this schematic form to the details, which add flesh to the bones of the inference.

TIBE would take into account the content of the report, in a primary way, but also the manner and context of the report, and the track records of the authors. As to the content of the report as such, its plausibility derives from congruence or coherence with previous knowledge current at the time - previous knowledge to which Waagstein's group itself had contributed. In 1974, Waagstein's group had showed that chronic administration of noradrenaline could produce a cardiomyopathic condition in rats; in 1971 the same group had successfully introduced intravenous treatment with beta-blockers in the acute phase of transmural myocardial infarction, a state also associated with high sympathetic stimulation (cf. Waagstein 2002). Hence, there were reasons of minimal plausibility around the date of the report based on previous knowledge, even if the full justification for the revolutionary treatment for cardiac insufficiency with beta-blockers was still in the making, and the general physiopathological understanding of cardiac insufficiency did not lent support to this treatment at the time. That is to say, even if one had doubted the direct causal relation between the beta blockers and the improved cardiac index, one could have conceded that the improved cardiac index could have arrived anyway, as an indirect result of the control of cardiac

[32] Cardiomiopathy being a condition in which the heart's capacity to contract is reduced, due to damage in the myocardium produced by toxic, metabolic, or infectious agents. It could also be idiopathic. The most common form of cardiomopathy is dilated cardiomiopathy, in which the cavities of the heart become enlarged.

rhythm.

Parenthetically, we would not want our theory of testimony to rule out conceding the truth of a report which bears witness to unusual findings, especially in a scientific context, in which such unusual findings may trigger important new discoveries or pathways of research. One would want however that these unusual findings have a minimal coherence with the background knowledge of the science in question. In our example, as I said above, one could concede the truth of Waagstein's report, even if it concerned a new path of research in the domain of cardiac failure, since the findings of Waagstein were correlations that could have meant either that this new paths of research is promising, or that the correlations were accidental from the point of view of the direct efficacy of beta-blockers.

The manner and context of the report, not related to the testimonial content as such, but still important, would also contribute to TIBE. The report was published in a high-profile journal (*British Heart Journal*), using the methodology and manner of description of scientific clinical reports in medicine. Detailed description of each case, with description of the investigations performed before and after the treatment (phonocardiogram, carotid pulse curve, apex cardiogram, and echocardiogram). Finally the judgment on reliability framed in terms of TIBE would also take into account the track-record of the authors of report; they were well-respected clinical practitioners, and it is worth adding that Waagstein's group had been working in cardiology research for a number of years, with recognized results.

In the framework of TIBE, one would thus infer the truth of the 1975 report by Waagstein *et al.*, based on both dependent, directly testimonial content and independent elements, taking into account the track record and, overall, being open to the diversity of evidence that can be used for the truth of evidence, as it were. We have thus a quite straightforward application of TIBE to medical reports. However, two observations are in place here.

For one, it should be said that the results of applying TIBE to medical-oriented testimonies should not be overvalued, in the sense of charging the content of such medical reports with too much conceptual load; for another, neither should such results be considered simplistic translations of current talk in the epistemology of testimony for medical context. As for the first observation, I mean to say that there are two ways in which the application of TIBE to medical reports can be conceived, depending on how content-loaded the report is considered to be. If, as I have assumed above, the report is taken to contain strictly the observation of *correlations* (without the added causal conclusion that such and such drugs *brought about* such and such effects), then TIBE on its own suffices to infer the truth of the report. If, on the other hand, the report is taken necessarily to include the *causal* dimension, then in order to infer the truth of the report one would need first

TIBE to infer the truth of the observed correlation, and then standard IBE to infer the efficacy, i.e. the causal dimension embedded in the report (and of course, the report in question would not be sufficient on its own). The latter would move us however from the realm of IBE or TIBE as applied to the quality of evidence, to the realm of IBE as a confirmation theory, which, as I mentioned in the introduction to this paper, are two separate (though connected) realms.

As for the second observation, it should be said that the application of TIBE to medical contexts in not, unfortunately, just a general epistemological discussion with no relevance for the actual medical practice. Just as an example, in around the same period as Waagstein *et al.* were publishing their results, a scandal sparked in the same cardiology circles with respect to a series of papers co-authored by a Harvard researcher, John Darsee. The suspicion arouse from *the manner of reporting* (there were simply too many papers published in a short amount of time) and *content* flagrant discrepancies (observed when comparing the data of a multi-center study), and subsequent investigation found laboratory meddling with results (see Relman, 1983 and also Wilmshurst, 2007, who gives other examples of dishonest medical practice in publishing, and also discusses the influence of pharmaceutical companies in the entire process). For the sake of enriching the context of the examples I am using here, one can add that among the papers written by Darsee that were subsequently retracted figured two publications on cardiomyopathy (Darsee, 1983), and that Darsee was part of a research group founded by Eugene Braunwald - an (already) legendary cardiologist who had actually discovered hypertrophic cardiomyopathy, but who, on the issue of using beta-blockers in heart failure, was not on the same side as Waagstein *et. al.,* for reasons hinging on theoretical assumptions about the mechanism of heart physiology and heart failure.

I will come back to the more theoretical bone of contention between Waagstein and Braunwald in the next section, because, as I believe, IBE can play a role in the evaluation of the quality of evidence that goes beyond the testimonial aspects discussed above. More precisely, I would like to argue that there is a sense in which TIBE provides a *pattern* of criteria and values to evaluate quality of evidence *in general* - not restricted, that is to say, to testimony properly speaking, but taking into account various other sources of evidence. In other words, beyond the testimonial aspect (which of course remains an inextricable part of the act of offering evidence in science) the pattern provided by TIBE could be used to assess the quality of evidence provided from various sources in a more in-depth manner, absorbing and doing justice to more causal information, but without thereby turning TIBE (or the pattern of it) into a theory of confirmation.

However, in order to see what this pattern looks like, we will have to make more visible the roots of TIBE in the initial account of Inference to the Best Explanation provided by Peter Lipton. It is what I will do in the first part of the next section, showing then in the second part the

immediate relevance of this pattern to the set of criteria for quality of mechanistic evidence provided in Clarke *et al.* 2014.

§3. The TIBE pattern and other sources of evidence

There is indeed room to make TIBE look more similar to IBE, in the sense in which, the explanatory inferences of IBE are at the moment better defined and circumscribed than the inferences of TIBE. As I have pointed out in §1, the inferences to the best explanation appeal to certain criteria and values in order to reach what is the best among the available explanation. We are speaking here of the classical explanatory values employed by all IBE-theorists, namely simplicity, theoretic unity, scope, and individualisation, on the one hand, and the causal criteria that Lipton himself had added in order to make IBE inferences more precise, namely Mill's criteria, in particular the Method of Difference and the Method of Agreement. Lipton's basic insight was that most often, when looking for an explanation, we are following the track of inferring the cause from the manifest effect. Hence, Lipton convincingly showed, the basic causal criteria used in everyday practice in order to track the sources of causation, could also be used, in conjunction with the classical explanatory values, in order to track the best explanation.

Now, it is important to see that this causal vein is also present, or could meaningfully be envisaged, in the testimonial case. That is because one can say that in TIBE, what one does is to infer from the utterance or expression of the testimony as effect to the truth of the belief of the speaker as a cause.[33]Obviously, one need not think here of 'truth' as some sort of strange entity, postulated in order to account, out of the blue, for a causal factor mysteriously acting in the background of our acts of testimony. One just needs to think that, according to Lipton, one infers, from the testimony, an explanation within which the truth (or falsity) of the testimony plays a part, and that this whole process can be interpreted in a causal key, at least for heuristic purposes.[34]

If one accepts the interpretative lenses according to which in TIBE are inferring causes from effects through an explanatory chain, as in IBE *à la* Lipton, then the recipe of 'explanatory values+causal criteria' should do some work in making the testimonial inferences more precise and

[33] The idea that truth can be a cause might sound unusual but it has been propounded before - famously by Aristotle, who declares, when reviewing the philosophical positions of his predecessors, that, in spite of their overall errors, some had to acknowledge certain proper metaphysical statements, 'being forced by the truth - ὑπ᾽αὐτῆςτῆςἀληθείαςἀναγκαζόμενοι'; *Met* A, 984b9-10). Truth means here for Aristotle certain states of affairs that his predecessors could not ignore.

[34] A causal interpretation of testimony is provided, from a different, Bayesian perspective, in Bovens and Hartmann (2003). It should not be forgotten that Hume formulated the position that stands in the background of the reductionist approach as a particular case of his general reasoning concerning *causation* 'The reason, why we place any credit in witnesses and historians, is not derived from any connexion, which we perceive a priori, between testimony and reality, but because we are accustomed to find a conformity between them'. (Hume 1977 [1748], 75)

well circumscribed. On the one hand, some of the general explanatory virtues could be employed in order to circumscribe the conditions under which the content of testimonial reports points towards its truth. The thought here is that the more detailed, simple and coherent with previous knowledge the reports in question are, the more likely they are to be true. Lipton hinted towards using coherence in this role,[35] and the usefulness of simplicity and detail is easy to see (compare, as to simplicity i.e. fewer assumptions being adopted, 'my neighbor said that he dreamed last night that it was raining in Canterbury, and his dreams never fall astray', with 'last night it rained in Canterbury', and as to individuation, i.e. more detail being provided, compare 'yesterday at 4 pm it rained for 50 minutes in Canterbury' with 'yesterday it rained in Canterbury').

As to the role of causal talk in elucidating the explanatory inferences, it is in particular Mill's Method of Agreement - if two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause of the given phenomenon - that could make it more clear how factors other than the content of the testimonial report influence the tracking of truth. What is the explanatory inference that can be drawn, given several testimonies provided by different persons, in different circumstances, but with the same content, relating the same fact or event?[36] On grounds of the method of agreement, one would infer the that the respective content is true, i.e. that the respective fact or event was the case. Again, Lipton hinted towards the use of the method of agreement, for different testimonial sources; the method of agreement applies also, from a certain point of view, the explanatory virtue of coherence

"[…] incompatibility between what the speaker says and some of the hearer's deeply held beliefs is often a reason for rejecting the testimony. Yet, here there is no obvious explanatory link to the fact of utterance. Similar remarks apply to cases [...] of contradictions between different speakers' testimony. If our speaker is contradicted by another speaker's testimony, this provides reason not to believe, but the second speaker's testimony may bear no explanatory relation to the first speaker's testimony. But here to the defender of TIBE has some kind of reply, since she can say that negative evidence, whether from background or from contradictory testimony, will be registered by making a truth-entailing explanation less attractive and so less likely to be inferred" (Lipton, 2007, p. 251)37

This is then the properly augmented pattern of TIBE that could in general be applied to other

[35]"And clearly, the decision whether to believe what one is told will have some dependence on the prior probability one assigned to what is asserted, which can itself be based on all sorts of evidence"(Lipton, 2007, p. 245). Coherence figures also at the center of the Bayesian account developed in Bovens and Hartmann (2003), even though the authors argue that there can be no definitive *measure* of it.

[36]. The same point is made in Lipton (1998), p. 27.

[37]The explanatory values could also apply to inferences which take into account features other than the content of the testimony itself. Take the Humean track-histories, according to which the truth of a testimony is to be inferred by induction from previous correspondences between similar testimonial reports and actual states of affairs (or, when it comes to a single speaker, from previous testimonial reports, on any subject, show to correspond to facts). If Harman and Psillos are right and the successful or warranted inductions are limit cases of inferences to the best explanation guided by the explanatory values hinted at by Harman and developed by Psillos, then track-history justifications, which are compatible with TIBE, can also be shown to rest on such explanatory values.

sources of evidence, related to testimony but not directly testimonial. It is a pattern in which one chooses the explanation for the very existence of evidence - based on a combination of Mill's Methods and the explanatory values of coherence, simplicity and individualisation - inferring from the very existence of evidence (coming out of an evidential source) its reliability or non-reliability, broadly speaking, and also the quality or weight of the evidence as such in terms of its intrinsic content. More specifically speaking, the consequence of applying this pattern is that one can justifiably grade the quality of evidence overall (including both the content and the reliability of evidential sources), in proportion to the degree to which the respective explanatory values and Mill's Methods are satisfied. Finally, since grading is involved, we might want to interpret the TIBE pattern in a yet closer way in relation to Lipton-style IBE, in the sense in which Lipton, in contradistinction to other IBE theorists like Bird or Psillos, sees the IBE inferences as pointing to the probable truth of the best explanation, instead of the truth of it *tout court*. 'But Inference to the Best Explanation cannot then be understood as inference to the best actual explanation. Such a model would make us too good at inference, since it would make all our inferences true' (Lipton, 2004, p. 58)[38]

Let us now look again at the criteria for grading evidence of mechanisms provided in Clarke *et al.* (2014). These criteria take into account traits such as independent methods, different research groups, proportion of features found, knowledge of analogous mechanisms, and robustness, defined in terms of being reproducible across a wide range of conditions. Each of these traits plays a role in terms of pluses and minuses, in the evaluation of the quality of evidence for mechanisms.

Pluses	**Minuses**
Each independent method that confirms a feature	Each independent method that fails to confirm—or, worse, disconfirms—a feature
Each independent research group that confirms a feature	Each independent research group that fails to confirm—or, worse, disconfirms—a feature
Larger proportion of features found	Smaller proportion of features found
Analogous mechanisms known	The analogy is a weak one, or, worse, analogous situations exhibit no such mechanism
Robust, reproducible across a wide range of conditions	Fragile, not reproducible in slightly varying conditions

Mutatis mutandis, the pattern I have outlined above applies perfectly to the criteria laid out by Clarke *et al.*, providing an epistemological underpinning and a justification for them. The criteria in

[38]Among others, this was Lipton's way of responding to van Fraassen's charge of 'the bad lot'. Gelfert has already made the move from truth to probable truth in his own, slightly twisted version of TIBE, in which the default stance of testimonial acceptance is justified on abductive grounds, and inferences to the best explanation are used to reject testimonies that are *probably* false. 'On the one hand, the coherence and success of our testimony-based projects provides general abductive support for a default stance of testimonial acceptance; on the other hand, we are justified in rejecting specific testimonial claims whenever the best explanation of the instances of testimony we encounter entails, or makes probable, the falsity or unreliability of the testimony in question' (Gelfert, 2010, p. 386).

terms of *different* research teams coming up with the same mechanistic result, and in terms of *different* methodologies used in order to reach the same result, both appeal to the method of agreement and the quality of evidence can thereby be assessed on explanatory grounds.

On the other hand, in terms of content of evidence as such, the reports that count more mechanistic features than others will be ranked higher in a TIBE-like inference on grounds of the explanatory virtue of individuation. The evidence finding similar mechanisms in analogous situations will be ranked higher on grounds of the explanatory virtue of coherence. Finally, the evidence on mechanisms being reproducible across a wide range of conditions will be ranked higher on grounds of the explanatory value of simplicity, since the assumption of the same mechanism working across different situations is simpler than the assumption of different mechanisms functioning across different situations.

Hence TIBE or TIBE-like reasoning will justify why, say, the mechanistic evidence counting two different teams reporting two mechanistic features, using *the same* research methodology, will be graded higher than evidence of two mechanistic features provided by a *single* team, but will be graded lower than evidence coming from two different teams reporting two mechanistic features using *different* methodologies, and the latter evidence will be lower than evidence coming from two different teams reporting *three* mechanistic features using different methodologies, and so on.

Coming back to the history of beta-blockers, out of which was drawn the example concerning the TIBE-interpreted testimony in the previous section, one can recognize this way of hierarchizing the quality of evidence from how this treatment was evaluated. It should be said that Waagstein *et al.* showed courage in 1975, when trying on the potential benefits of beta-blockade in this pathological cardiac condition, because they were running against the physiopathological models of cardiac failure at the time, reinforced by clinical assessments seemingly agreeing with this model, and by the accepted mechanism of action of these drugs. Beta-blockers had been synthesized in 1962 (following the discovery of the beta receptor of the sympathetic nervous system in 1948) and were used initially to reduce stress in patients with angina pectoris, for arrhythmias, and, from early 1970's on, as an anti-hypertensive treatment. By blocking the beta-receptor, and accordingly the stimulating action of the sympathetic nervous system, such drugs diminished the need of oxygen in ischemic episodes, controlled tachycardia, and reduced the blood tension in arteries by diminishing heart's capacity to pump blood (its inotropic capacity). However, in virtue of precisely this mechanism, they were formally contraindicated in heart failure, since one was not supposed to use a drug decreasing the pump function when there was already insufficient inotropic activity. As I said, there were even clinical assessments that seemed to certify this counter-indication – animal experiments seemed to showed that cardiac failure worsens when the influence of the sympathetic, adrenergic

system is triggered down by adrenergic blockers, beta-blockers included (e.g. Gaffney and Braunwald, 1963, Braunwald and Chidsey, 1965, Epstein, Braunwald *et al.* 1965). In 1965, Eugene Braunwald, a legendary researcher, was resuming his findings (in collaboration with Chidsey) as follows:

'The adrenergic nervous system plays a particularly prominent role in supporting myocardial function when the latter is depressed in congestive heart failure. […]The importance of the augmented activity of the adrenergic nervous system in maintaining ventricular contractility when the function of the myocardium is depressed in congestive heart failure is shown by the effects of adrenergic blockade in patients with heart failure. In patients on a metabolic diet guanethidine frequently caused sodium and water retention, as well as intensification of heart failure (Gaffney & Braunwald 1963). Recently, we have made similar observations on the aggravation of congestive heart failure with propranolol (Epstein & Braunwald, in preparation). The adrenergic nervous system thus plays an important compensatory role in the circulatory adjustments of patients to congestive heart failure and caution is needed in the use of anti-adrenergic drugs such as reserpine, guanethidine, and propranolol in the treatment of patients with limited cardiac reserve…… . In view of the strongly positive inotropic effect exerted by the NA released from these nerves, the adrenergic nervous system may be considered to provide potential support to the failing myocardium. However, if the reduction of NA stores in some instances of heart failure is associated with a diminished release of neurotransmitter, as now appears to be the case, then this depletion of NA may be responsible for loss of the much-needed adrenergic support to the failing heart and so intensify the severity of congestive heart failure.' (Braunwald and Chidsey, 1965, 27-30, italics added)

Exactly the opposite of the line followed by Waagstein *et. al.*! Waagstein and his collaborators from the university of Göteborg discovered accidentally in late 1972 the beneficial effect of beta blockers in heart failure, when they administered alprenolol (a non-selective beta blocker) for tachycardia to a 59 year-old woman patient presenting acute pulmonary edema owing to dilated cardiomyopathy, and this dramatically improved her overall condition (*acute* cardiac failure) (Waagstein, 1975). Waagstein *et al.* hypothesized that, in *chronic* cardiac failure as well, by administering small doses of beta-blockers and thereby reducing the metabolism and energy consumption of the heart, the favorable effects would not be outweighed by the loss in the inotropic activity (Waagstein 2002, pp. 215-216, 218). The 1975 small trial on patients suffering from dilated cardiomyopathy in chronic cardiac failure stage described in the previous section was part of a series of clinical assessments, that gradually became more and more numerous and consistent, leading eventually, almost 15 years later, to a mechanistic shift of paradigm, from a hemodynamic model (viewing the heart as a pump and interpreting 'mechanically' its physiology and physiopathology) to a neuro-hormonal model (viewing the heart as lying at the intersection of the nervous and immune system and accordingly influenced by the overexpression of biologically active molecules; Mann and Bristow, 2005).

Now, 15 years is a long time, and in the decision to accept the beta-blockers treatment, a crucial role was played by some late, large-scale population trials that showed it benefits. However, if we compare the *mechanistic* evidence for the hemodynamic, mechanical model upheld by Braunwald, with the *mechanistic* evidence for the more integrated, metabolism-oriented model upheld by Waagstein, we can understand why in the initial period the proposal of Waagstein's group was

viewed with skepticism, and also why it was finally accepted (or, to put it differently, why in the later period it was decided that large scale population trials are worth doing). In the initial period, the evidence for (what I have called) Braunwald's model benefitted from the strong analogy with the mechanical interaction of a pump and its recipients, which was actually in place ever since Harvey had discovered the circulation, and the model was tacitly shared by all research groups working in cardiology, cardiac failure included. This model of mechanism also seemed to possess all (or a sizeable part of) the intermediate features, or chains of interaction – the receptors and neurotransmitters of the vagal and sympathetic nervous system, as well as the complementing effects of the renine-angiotensine-aldosterone on the volume of circulating blood, sodium retention and vasoconstriction. On the other hand, Waagstein did not have such an analogy, his group was alone in doing research on the alternative model, and very few of the intermediary features or chains were known at that initial stage. His model received impetus **i)** when other teams picked up the research (as for instance the research group of Douglas Mann from the Medical University of South Carolina, the group of Kenneth Margulies from Temple University of Philadelphia, and many others), **ii)** when more intermediary chains of the mechanism were discovered (as for instance the intracellular deficiency of cyclic AMP leading to defects in the myocardial contractile elements (Feldman *et al.* 1987), deactivation of arterial baro-receptors leading to lack of downregulation of the sympathetic effects by the circulating blood volume, arhythmogenesis due to sympathetic activation, changes in the density and function of beta-receptors as well as in the cellular proteins connecting them to the effector enzyme adenylyl cyclase, left ventricular remodeling following myocytes' loss of contractility, and the list could continue), and **iii)** when alternative methodologies came to be used (for instance, initially, Waagstein *et al.* used non-invasive methods (Swedberg, 1993), but later were employed invasive techniques such as cardiac catheterization for assessments of intraventricular pressure; there were also increasingly varied animal experiments which induced cardiac failure using pathologically relevant concentrations of the neurohormones in question; Mann and Bristow, 2005). Much later on, Braunwald would reproach himself not paying more attention to beta-blockers.

'I kick myself now for not trying beta-blockers. We were afraid. We were concerned that if we blocked the body's response to heart failure, the heart failure would get worse. In fact, we did studies that showed that beta-blockers could worsen heart failure. But the Swedes showed that if they started at a low dose and slowly increased it, beta blockers could be used safely in many patients. And the benefit could be enormous, because the high sympathetic tone that we demonstrated wasn't a protective response – it was actually part of the problem.' (Lee, 2013, pp. 162-163).

But it has to be recognized that in the initial stages the quality of evidence in favour of the neuro-hormonal model was inferior to that in favour of the hemodynamic model. And as I have argued above, this can be justified in terms of IBE, more precisely, in terms of the pattern of TIBE

underlying the Clarke *et al.* criteria. That is to say, the reason why it took several years before the mechanistic evidence accumulated such as to trigger large scale population assessments was that several criteria from the Clarke *et al.* list had to be met, and they had to be met because otherwise one could not draw an explanatory inference towards the quality or weight of the evidence in question.

I end this chapter with two caveats that are important for the scope and effectiveness of the present enquiry. First of all, note that, by keeping distinct the level of assessing the *quality* of mechanistic evidence, on the one hand, and the level of *confirming* causal claims in medicine, one is not obliged to directly draw ultimate *causal conclusions* from the TIBE-based hierarchization of the quality of evidence, even if *causal information* finds its way the evidence being hierarchized. In particular, my argumentation about the use of a pattern of inference to the best explanation for assessing the quality of evidence is entirely consistent with the claim advanced by Clarke *et al.* that mechanistic evidence needs to be combined with evidence from population studies in order to establish and confirm causal claims (Clarke *et al.* 2014, pp. 351-356). As I said, the mechanistic explorations into beta-blockade I have described above were attended by population studies assessing the net effect of the various mechanisms at stake in such a complex condition as heart failure, and played, along with the discovered mechanisms, a crucial role in the decision to implement the systematic use of (small doses) of these drugs as treatment.[39]

The second proviso refers to the fact that I have adopted the neutral definition of a mechanism provided by Illari and Williamson, according to which 'a mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon' (Illari and Williamson, 2012, p. 125) - a definition that Clarke *et al.* also adopt in their paper. But, as I have adverted in the Introduction to this thesis, there is a stronger construal of mechanisms that takes them to imply not just production but also difference-making.[40] On this strong construal, one would have, on the one hand, a more precise definition of what a *robust* or a *fragile* mechanism is (in terms of the difference-making that a mechanism is supposed to enable/produce across varying contexts). On the other hand, one would have a much firmer grip on evidence of mechanisms using the Lipton-style Inference to the Best Explanation, which tracks and integrates difference-making via Mill's principles.

[39] The mechanism of cardiac failure and its treatment are still subject to debate. In a recent interview, Braunwald was confessing that after several decades of research into the mechanisms of cardiac failure, the issues that were confronting him in the beginning of his career are still on the table for the medical community (Landau, 2012). One can say that, on the level of *confirmation*, one final resolution has not yet been reached. However, these decades of research have brought about *quality evidence* about numerous sub-mechanisms of cardiac failure, detailing the neuro-hormonal consequences of the adrenergic influences in cardiac failure. Incidentally, this touches upon a point I have been repeated throughout this chapter - that aside from the studies of confirmation, one needs a theory that does justice to the quality of evidence.
[40] Woodward (2002), Woodward (2011).

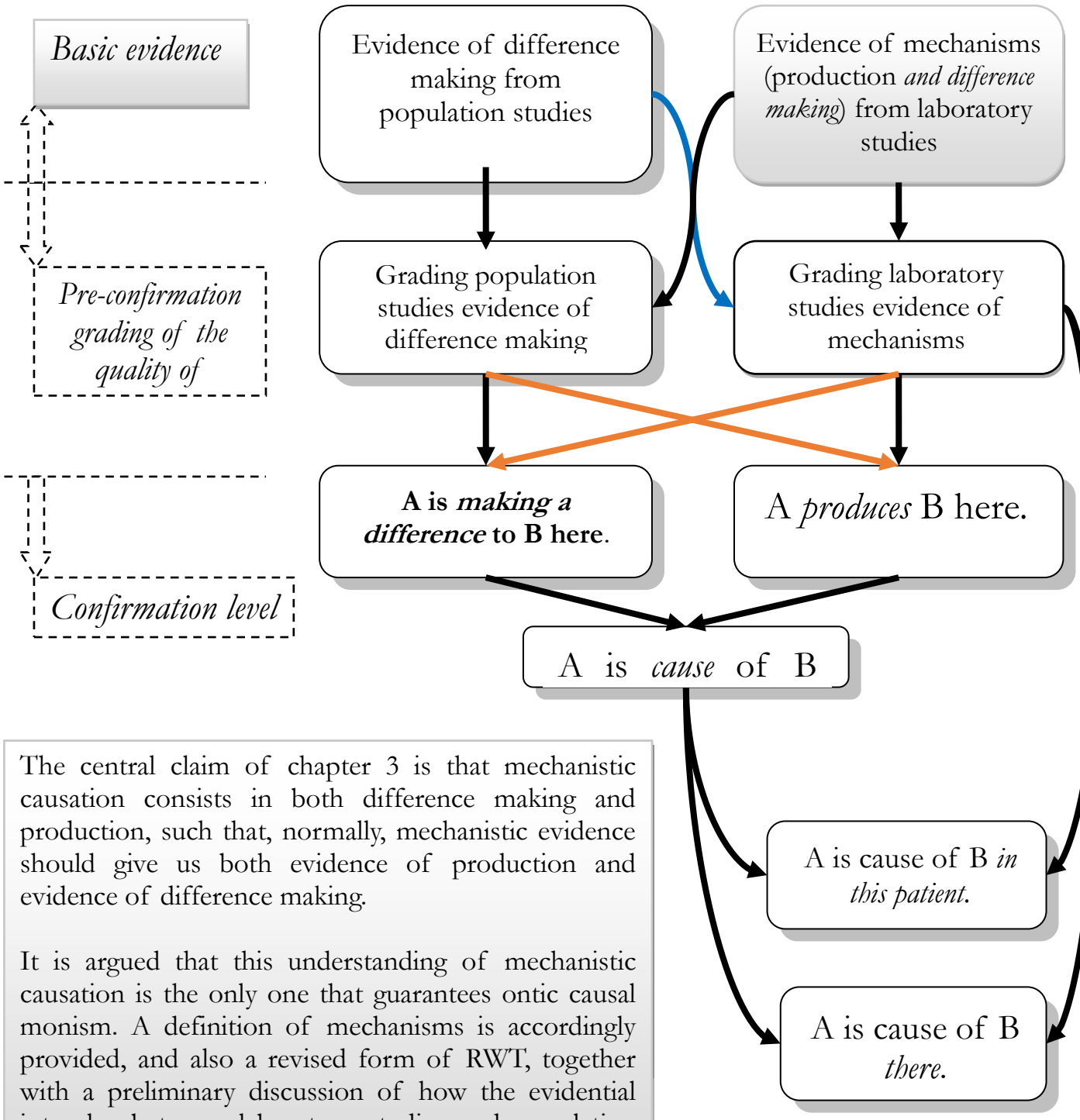
I will explore this alternative view of mechanisms in the chapters 3, 4 and 5. Suffice to say here in closing the present chapter that - as the most recent population results show - the evidence of the mechanism involved in the beneficial effect of beta-blockade in cardiac failure point towards a *fragile* mechanism.[41] The difference making produced by the administration of beta-blockers needs to be very carefully weighted. One needs to take into account the administered dose, the complex pathological background that is specific for a wide range of different cardiac diseases ending in cardiac failure, and the multitude of sub-mechanisms interacting between them, when the adrenergic stimulation of the myocardial tissue is interfered with. Such a fine-grained assessment of difference-making can only result from the careful interplay of population studies and individualized research into sub-mechanisms. But this is indeed the subject of chapters 3, 4 and 5 below.

Conclusion chapter 2

I have explored the uses of Inference to the Best Explanation as a theory for evaluating the quality of medical evidence, taking the evidence of mechanisms as a case study. I have shown that the testimonial side of IBE fits clinical reports in medicine, and that, when suitably augmented and clarified, IBE can be extended to provide a pattern that justifies explanatory inferences for other types of evidence, allowing the grading of the *quality* of evidence. I have finally shown that such an explanatory pattern justifies the set of criteria of evaluating mechanistic evidence provided in Clarke *et al.* 2014.

[41] See for instance Hernandez *et. al.* 2009, who points out a potential lack of effectiveness for elderly patients.

Chapter 3



The central claim of chapter 3 is that mechanistic causation consists in both difference making and production, such that, normally, mechanistic evidence should give us both evidence of production and evidence of difference making.

It is argued that this understanding of mechanistic causation is the only one that guarantees ontic causal monism. A definition of mechanisms is accordingly provided, and also a revised form of RWT, together with a preliminary discussion of how the evidential interplay between laboratory studies and population studies is to take place.

Chapter 3 Mechanisms and difference-making

Introduction

The present chapter marks the transition to the second part of the thesis, in which the ontic definitions of mechanisms are examined. Thus, it will first be useful to briefly take stock of the work done so far in the previous chapters.

The previous two chapters have looked into how IBE can be used in relation to the mechanistic evidence strictly speaking, i.e. the evidence obtained in laboratories, in microstructural research (leaving aside the evidence offered by population studies). More precisely, chapter 1 looked at the uses of IBE as a theory of confirmation for certain particular cases of mechanistic evidence encountered in medical practice. These are cases which offer us enough epistemic warrant, such that, in the framework of IBE, we could eliminate as inadequate rival hypotheses and be guided towards the correct one. On the other hand, chapter 2 enquired into the uses of IBE at the pre-confirmation stage in which the mechanistic evidence is given a preliminary, quality classification in terms of the (explanatorily justified) criteria of Clarke *et al.*

However, this strict focus on mechanisms adopted in the first two chapters, although obviously useful for an enquiry into grading mechanistic evidence, is admittedly a narrow one from the global perspective of establishing causal claims, and accordingly has to be enlarged. Most frequently, causal claims in medicine are assessed not only by looking at the mechanistic evidence *as such*, but also by considering the evidence brought about by population studies. This is, of course, common knowledge. The general public, medical practitioners and philosophers of science are all aware of the desiderata that the program of *Evidence-Based Medicine* (EBM) has been putting forward for more than two decades. This program aims at a rigorous assessment of medical hypotheses strictly based on evidence. However, the evidence it takes into account is mainly *population level* evidence (in particular, the 'gold standard' of randomized clinical trials). This entails, of course, that evidence of mechanisms is considered to a much less degree. If it figures at all in the protocols of evaluation of EBM, it occupies a very low place, below various types of population level evidence.

As stated in the Introduction to this thesis, this EBM position has been criticized by Federica Russo and Jon Williamson (2007), who have argued that, in order to establish causal claims, one needs to take into account *both* evidence of mechanisms and evidence of difference-making coming from population studies. This position has come to be known as the Russo-Williamson Thesis (RWT) and it has gained numerous proponents (and opponents, it should be said). In fact,

the Clarke *et al.* criteria of mechanistic evidence which we have looked at in the previous chapter issue from research done by a larger group of philosophers, who have been exploring the implications of RWT.

I am myself a proponent of RWT, and I think this thesis can fruitfully be used, in the framework of IBE, in order to extend the latter's area of application well into the evidence of population studies that should complement and fortify evidence of mechanisms (and vice-versa). However, I will argue, RWT has been inadequately associated with an ontic definition of mechanisms that is very thin, in the sense of focusing on production causation and setting aside the difference-making of mechanism components and mechanistic output. This definition is not particularly useful for the very purpose RWT was put forward (namely to conceptualise properly the interplay between mechanistic evidence and population studies). . As I shall explain later, for one, it makes RWT vulnerable to criticism inspired by medical practice (Howick, 2011). For another, it misses important features of the interplay between evidence of population studies and evidence of mechanisms, like the fact that they can reciprocally increase their quality or weight, or that mechanistic evidence can individualise and calibrate the assessments on population level. In addition, it does not offer sufficient resources to counter a widespread (if often tacit) ontically pluralistic view on causation - i.e. the view that there are distinct types of causal relations, which is at the root of the EBM program of favouring evidence of population studies. Indeed, Russo and Williamson's have justly sensed the challenges posed by ontic causal pluralism in their original (2007) contribution, but their attempt to reject it, as I will show, is unsuccessful.

However, the adoption of an ontically thin definition of mechanisms is not just a problem of RWT in its current form. It is a rather general problem of the philosophical literature, because most often mechanistic causation is understood in terms of production only (Illari and Russo, 2014, Craver and Tabery, 2015). Accordingly, the difference making aspect of causation is left aside – due, among others, to the exaggerated importance accorded to the problems of pre-emption and of causation by absence (Williamson, 2011) which treat cases in which, purportedly, we could have production causation without difference-making, and difference-making without production, respectively.

The present chapter will thus make the case that, together with production, difference making should be viewed as part of mechanistic causation, following and developing a suggestion put forward in Joffe (2013). This move will open up the way for rejecting the criticism typified by Howick (2011)'s contribution (which I will do at the end of the present chapter), to fruitfully apply IBE not just to mechanistic evidence as such, but also to the interplay between population studies evidence and mechanistic evidence with respect to their quality or weight (chapter 4), and to show

how this interplay should have a bearing on the problem of extrapolation (chapter 5).

The plan of the present chapter is as follows. §1 will lay down the reasons for preferring the monistic view that mechanistic causation is ontically both productive *and* difference-making, and will indicate how RWT is to be adjusted according to this view. It will then discuss how this monistic view rejects the opposing view of causal pluralism, as well as why Russo and Williamson's solution for rejecting the former and embracing the latter fails. §2 will deal with the problem of pre-emption (and also of absences), appealing most importantly to the work of Michael Strevens. §3 will look at how the revised form of RWT stands, after having solved the problem of pre-emption. §4will defend RWT (as strengthened by the difference making construal of mechanisms) against the criticism expressed in Howick (2011).

§1 Mechanisms, production and difference making

A strange feature of mechanistic accounts nowadays is that most of them (including Illari and Williamson's, which is currently associated to RWT), while mentioning production, functioning, responsibility for events, etc. avoid the terminology of difference-making (in its counterfactual guise, or as probabilistic dependency). Here are some well-known examples:

Illari and Williamson (2012) 'a mechanism for a phenomenon consists of entities and activities organized in such a way that they are *responsible* for the phenomenon' (Illari and Williamson 2012, p. 120, italics added)
Bechtel and Abrahamsen (2005) 'A mechanism is a structure *performing a function* in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is *responsible* for one or more phenomena.' (italics added)
Glennan (2002) 'A mechanism for a behaviour is a complex system that *produces* that behaviour by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations.' (italics added)
Machamer *et al.* (2000) [the so-called MDC account] 'Mechanisms are entities and activities organized such that they are *productive* of regular changes from start or set-up to finish or termination conditions.'(italics added)

That is to say, mechanisms are usually associated with so-called production causation, as opposed to difference-making causation. Briefly stated, production causation is typified by processes: A causes B by producing B via such and such a process. The details of what counts as a process may differ, but in the case of biological mechanisms, a process should involve at least the participation of multiple entities with a complex, specific spatio-temporal arrangement (Glennan 1996, 2005; Bechtel and Abrahamsen 2005; Machamer, Darden, and Craver 2000).

On the other hand, difference making causation is typified by the dependency between causal factors, which, in a simple counterfactual expression, means that A causes B because in the absence of A, B would not have occurred (Hall 2004; Psillos 2004). Again, the details of what counts as dependency between two causal factors may differ. It can be understood in terms of interventions on causes making a difference to effects (Woodward, 2011), as probabilistic dependency between

causal factors (Illari, 2011), or as a more complicated counterfactual dependency, for instance of the backtracking type (Broadbent, 2007; Strevens, 2013), which can be complemented by a difference-making rationale appealing to the famous INUS conditions (Mackie, 1973, Strevens, 2004, 2007), as I discuss below.

The basic intuition behind the various expressions of difference making is that causation implies a notion of 'modal force', even on a Humean interpretation of it, in order to be distinguished from merely accidental relations between factors (see Bird, 2007). That is why, among the probabilistic, interventionist and counterfactual approaches to difference making, it is arguably the counterfactual approach which is the fundamental one, and which also underlies the interventionist and probabilistic accounts. For instance, most interventionist accounts will appeal to some clause of the approximate form 'were one to intervene on x, then y would behave so and so'. Probabilistic accounts, in turn, could be easily read as counterfactual dependencies, where either the antecedent or the consequent of the respective counterfactuals is actualised.[42] In turn, putting to work Mackie's INUS conditions to track difference making is very close to appealing to backtracking counterfactuals, which infer from the hypothesized absence of the *actual* effect the hypothesized absence of the *actual* cause, or so I will argue.

One other heuristically useful way to classify the different approaches to difference-making is to say that they express the dependency between cause and effect as either the *sufficiency* of the cause for the effect, or the *necessity* of the cause to the effect. The sufficiency is at stake in the interventionist accounts wiggling the causes to show variation in the effects, Woodward 2011, or in the probabilistic accounts showing the increase in probability of effects conditional on causes (as expressed in the Bayesian framework by likelihoods, which has the probability of evidence upon hypothesis 1 as a limit case). The *necessity* of causes for effects is obviously present in the classical counterfactual analysis à la Lewis, 1986 (which, informally put follows the rationale - when cause are away, so are the effects). In turn, the backtracking counterfactuals based on Mackie's INUS conditions express a combination of necessity and sufficiency. The INUS conditions, as it well known, stand for necessary parts of sufficient conditions for effects, and the corresponding backtracking counterfactuals proceed by modus tollens to infer the absence of a certain INUS condition from the hypothesized absence of the *actual* effect.

Importantly, all these various ways of expressing difference making can be seen as more sophisticated expressions of basic causal intuitions of experimental research, reflected in Mill's

[42] When probabilities are understood either as propensities, or as credences (such that the corresponding probabilistic statements are epistemically imperfect expressions of deterministic laws). If counterfactuals could be said to attend deterministic laws (which generally express sufficiency, all Fs are Gs, where we have a counterfactual whose antecedent is actualized, if a was an F, it would also be a G) then, as a limit case, the same could be said about the probabilistic accounts reflecting our partial ignorance, or expressing probabilistic laws as such.

methods, which were discussed in the previous two chapters. However, as I said, in spite of the wealth of approaches to difference-making and their intuitive appeal, the current definitions of mechanisms choose to leave aside difference-making and focus on production. Moreover, this approach to mechanistic causation is inevitably attended by a pluralistic approach to evidence, such that the *evidence* of production is set in contrast to *difference-making* evidence. The main cited reason for this separation between evidence of production and evidence of difference, as well as for associating mechanisms to production only causation, is the existence of the problems of pre-emption and of causation by absence. Simply put, pre-emption situations are cases of (mechanistic) production in which the counterfactual approach cannot isolate the causes at stake because other causes would have produced the effects anyway. On the other hand, causation by absence applies to situations that are cases of counterfactual causation in which one cannot identify any process linking the putative cause with the putative effect.

Importantly, this pluralistic view of *evidence* (on the epistemic side of discussion) which is also embraced by proponents of RWT (Illari, 2011), has gone hand in hand, tacitly or explicitly, with a pluralistic approach of *causation*, both on the ontic and conceptual side of discussion. In its *conceptual* guise, this pluralistic approach says that we possess different concepts of *causation*. In its *ontic* (or metaphysical) formulation, the more drastic one, conceptual pluralism says that the multiplicity of concepts of causation reflects the multiplicity of causal relations as such (de Vreese, 2006, Godfrey-Smith, 2008, Longworth, 2006). Crucially, the production only view of mechanistic causation implies ontic causal pluralism. Indeed, if mechanistic causation is production-only causation, it is hard to imagine how the difference-making relation could be something other than a *different* type of causal relation.

It is worth noting that *ontic* pluralism is *not* implied by *conceptual* pluralism. To use Frege's famous example, we might have different *concepts* for a certain species of vertebrates ('has a heart' and 'has a kidney'); and one could add that accordingly, we might have different types of *evidence* we take into account to identify the respective species (or identify an individual member of it). But the main point of Frege's example is that these different concepts (or different types of evidence, as I have added) pick out the same species (ontically speaking). Transposed to the case of causation, the suggestion is that conceptual pluralism (and evidential pluralism) about causation is consistent with ontic monism on causation. Our different ways of hunting down cause-effect relations (when looking for different types of evidence) and of naming or conceptualising different aspects of these relations need not imply that our (potentially) different concepts and/or our different types of evidence pick out *different* types of causal relations on an ontic level.[43]

[43] Compare Russo and Williamson differentiation "Conceptual pluralists normally hold that each concept of cause picks

It should be said that the pluralistic view of evidence is a fruitful epistemic stand point, and its adoption is actually one of the strengths of RWT.[44]In turn, conceptual pluralism, although not an advisable position, for reasons of clarity of exposition, does not seem to pose insurmountable difficulties for the way we deal with causation matters, both in theory and in practice (see Cartwright, 2007).

Ontic pluralism, on the other hand, is, to put it crudely, a disaster for the epistemology of causation in medicine (and for RWT as well). The main reason is that, from an ontic pluralist standpoint on causation, it could not justify why and how different evidence is successfully aggregated. For instance, why would we necessarily need evidence of both production and difference-making when establishing causal claims, if production only could in itself constitute a full-blown causal relation? (e.g. if Helicobacter Pylori can produce via a mechanism gastric ulcer, then evidence of its production should be sufficient.) One could not use here the reply that epistemically, we would need evidence of difference-making in order to differentiate processes from pseudo-processes (the former genuinely causal, the latter accidental) since, on the ontic pluralist view, difference-making just concerns a different type of causal relation. Vice-versa, the above argumentation could be applied to difference-making causation. Why would we necessarily need evidence of both production and difference-making when establishing causal claims, if difference-making only could in itself constitute a full-blown causal relation? One could not use here the reply that epistemically, we would need evidence of production in order to differentiate genuine dependencies from spurious correlations (the former genuinely causal, the latter accidental) since, on the ontic pluralist view, production just concerns a different type of causal relation. Further on, how could different types of evidence be really aggregated, if they point to distinct phenomena (distinct causal relations)? Hence, one central insight of RWT (drawn out of medical practice), namely that of combining evidence from population studies with laboratory evidence, seems to lose its relevance.

Moreover, ontic pluralism is also tacitly involved in the views advocated by EBM. I mean to say that the reason why medical specialists of EBM have seemed to neglect evidence of

---

out a different causal relation—i.e., *conceptual pluralism normally presupposes ontological pluralism.* However, *it is possible to be an ontological pluralist without being a conceptual pluralist* by maintaining that there are two causal relations ambiguously picked out by a single concept of cause. Both varieties of pluralism are, we argue, implausible." (Russo and Williamson, 2007, p. 165, italics added). However, there is one possibility left out of Russo and Williamson's categorization, namely that one could be a conceptual pluralist and an ontic (or ontological) monist – as outlined above in the main text. And this means that conceptual pluralism needs not be rejected at all price, as Russo and Williamson maintain (whereas ontic pluralism has to be rejected). In return, the second of the two possibilities mentioned by Russo and Williamson (conceptual monism and ontological pluralism) is what ends up characterising their own position. The way an omniscient being draws out an acyclic Bayes graph (as stated by their concept of causation) characterises both mechanistic causation and difference-making causation. More on this to follow in due course.

[44] Illari (2011) has done much to show how different methodologies and different types of evidence can be crucial to discovering and confirming causal relations.

mechanisms lies arguably into a tacit adherence to a pluralistic view of causation. On this ontically pluralistic view, the population level studies are responsible for difference-making *causation*, whereas mechanisms as responsible for production *causation* only. Accordingly, this ontic pluralism separates mechanisms from difference-making, and it makes it hard to see how evidence coming from laboratory research could really influence the assessment and interpretation of the results obtained at the level of population studies, which is precisely what influences the reasoning of EBM theorists (as we shall further see in the last section of this chapter, when looking at Howick's criticism of RWT).[45]

One should attempt to reject thus ontic causal pluralism and advocate causal monism (in a way that would unify both production and difference-making). Indeed, Russo and Williamson have clearly sensed the dangers and challenges posed by ontic causal pluralism (Russo and Williamson 2007, pp. 165–167; Williamson 2011, pp. 435–437). Here is a relevant passage from their (2007) argumentation:

"But there is a second problem that besets pluralism, namely, that it inherits the difficulties of monistic accounts. This problem affects the pluralist who notes that there are two types of evidence for causal claims—mechanistic and probabilistic—*and concludes that there are two types of causal claim, mechanistic and probabilistic.* This is clearly a fallacious inference, and, worse, opens the pluralist up to the objections of section 5. *Suppose that the pluralist advocates two notions of cause, a mechanistic, cause1, and a probabilistic, cause2. Take any particular causal claim, e.g., 'smoking causes cancer', that the pluralist cashes out in terms of one or other of these notions but not both (there must be some such claim, for otherwise she is not a pluralist but rather takes causality to be one thing that has two aspects or components).* Now the evidence for this claim is multi-faceted, consisting of observed dependencies and mechanistic/theoretical considerations. But the pluralist's analysis of this claim will be single-faceted, say 'smoking is a cause1 of cancer'. But then the pluralist opens herself up to the epistemological problems of monism. If this particular use of 'cause' is mechanistic, cause1, then how can it be that, even when the mechanism is established and uncontroversial, further probabilistic evidence is cited in support of the causal claim? *However, if the use is probabilistic, cause2, why are mechanisms invoked as evidence, even when there is ample probabilistic evidence? The pluralist can't explain the variety of evidence for the claim: if pluralism is right, it should be possible that the evidence just be mechanistic, or just be probabilistic. So, while the pluralist may say that different uses of the word 'cause' can refer to different relations, some particular use must refer to a single relation*" (Russo and Williamson, 2007, pp. 166-167, italics added).

However, Russo and Williamson's solution fails because they only manage to reject conceptual pluralism (which, as we have seen above, needs not be rejected at all costs). Russo and Williamson argue that their epistemic concept of causation can accommodate both a) evidential pluralism and b) a type of ontic monism which incorporates the evidence of both production and difference-making. This ontic monism is taken to differ from the simple-minded monism, which says either that all causation is production, or that all causation is difference-making. On the epistemic causation view,[46] causal relations are the causal beliefs that an agent with access to *total* evidence should adopt (Russo and Williamson, 2007, p.167). Moreover, the causal beliefs in question should

[45] I am suggesting that the EBM proponents are tacitly ontic pluralists because it is the only explanation I can find for why they have not been paying more attention to the evidence of mechanisms. Alternative explanations are that they are pray to a positivistic, theory free ethos.

[46] Which is, parenthetically, a heuristically insightful view of causation to which I will abundantly appeal in the final two chapters of this thesis.

be represented by a directed acyclic graph whose nodes are the variables of interest and whose arrows correspond to direct causal connections, where this graph is constrained by evidence (and should otherwise be as non-committal as possible as to what causes what; Williamson 2006, pp. 75-82). Since the evidence in question is both evidence of difference making and evidence of mechanisms (i.e. in Russo and Williamson's view, evidence of production) it seems that evidential pluralism is thereby reconciled with ontic monism; all that we seem to be left with, on the ontic side, are the beliefs of the agent with access to the total evidence.

However, the solution in terms of the epistemic account of causation does not provide an ontically monistic reunification, but only a conceptual one, since the concepts of difference-making and production are further subsumed under the concept of causation derived from the inferential practices of the omniscient being. Evidential pluralism remains in place, which is a good thing. We also get conceptual monism, which is nice. But ontic causal pluralism also remains in place, which is a bad thing. And ontic causal pluralism remains in place because it continues to be implied by the view of *mechanistic causation* as productive only (taken together with the evidential pluralism view). Irrespective of whether one is a Humean (as Russo and Williamson are) or not, the view of mechanistic causation as productive only implies ontic pluralism about causation, as I stated earlier. If mechanistic causation is production-only causation, it is hard to imagine how the difference-making relation could be something other than a *different* type of causal relation.

It wort noting that at certain points in their (2007) argumentation, Russo and Williamson are ambiguous on ontic vs. conceptual pluralism in that it is not clear what they reject. But there are just two possibilities a) either they think they only need to reject conceptual pluralism, which does not help much, because a_1) conceptual pluralism is not to be rejected at all price, and a_2) ontic pluralism, which needs to be rejected at all price is still in place, or b) they assume that by rejecting conceptual pluralism then ontic pluralism is also rejected, which is false, because b_1) the conceptual pluralism they advocate is consistent with ontic pluralism, since b_2) ontic causal pluralism is still ontic causal pluralism whether one is a Humean or an anti-Humean (given an understanding of mechanistic causation as production only, together with the acceptance of evidential pluralism).

One additional ambiguity is that, while defining mechanisms in terms of production only, Russo and Williamson do not state whether mechanistic *production* is to be taken as *causation* or not. On the one hand, they argue that the problems of pre-emption and absences are problems of *causation*, which cannot be solved by an account in terms of difference making only, or production only, respectively (and this is the main theoretical reason for putting forward the epistemic account of causation, as a monistic solution to the pluralism that putatively emerges from the two problems) This entails that they accept the cases of production only as cases of *causation* (otherwise, we would

not have to worry about the problem of pre-emption). But this entails in turn that ontic pluralism remains in place even after one adopts the epistemic account of causation, for the reasons I stated above).

On the other hand, if their definition of mechanisms in terms of production only is not a definition of mechanistic *causation* (as Jon Williamson has let me know, in a personal communication occasioned by our supervision sessions) then, on causally monistic grounds, they should have drawn the conclusion that mechanistic *causation* includes both production and difference-making, which they never did.

Because indeed, the real solution to the issue of ontic causal pluralism is to join ontically production and difference-making. The thought was nicely expressed in Joffe (2013), following his analysis of RWT and its original framework of epistemic causation: '[….] which is the most important aspect of a causal relation? Its existence as a mechanism, seen perhaps as a fundamental power [i.e. production] or *the counterfactual difference that it makes*? The answer appears to depend on whether one's focus is primarily on the static question of what exists, or on the dynamic conception of how things change. But 'focus' is part of epistemology; *in reality, both aspects are inescapably present, because they are the two sides of a single coin*.' (Joffe, 2013, p. 188, italics added)

Joffe argues in other words that epistemology aside (i.e. the pluralistic view of evidence aside) difference making and production should ontically be seen as two sides of a single coin, picking out different aspects of the same causal relation. When it comes to mechanisms, this directly translates as being the solution to view difference-making as part of mechanistic causation, or as issuing from the productive activity of mechanisms. That is to say, on this view, whenever we have production, we should also have difference making (just like, in Frege's example, a creature with a heart will also have kidneys), or to use a slogan, all production is also difference making. Expressed in mechanistic terms, this means that whenever A is a cause of B mechanistically, then we have both production and difference making at work between A and B.

On such a construal of mechanistic causation, the entities internal to mechanisms make a difference upon each other, and upon the final output of the mechanisms in question, while mechanisms in turn make a difference upon the organisms they are part of. Thus, production and difference-making are indeed like "two sides of a single coin", to use again Joffe's insightful expression, in the sense in which wherever production is identified as a causal process, we should assume that such a process is also attended by difference-making and dependency (be it expressed in the actual world, via interventions or probabilistic dependencies, or, in possible worlds, via simple or backtracking counterfactuals, as we will shortly see).

Accordingly, the definition of mechanisms I propose is a modified version of Illari and

Williamson's (2012) – a mechanism for a phenomenon consists of entities *joined causal relations that are simultaneously productive and difference-making*, organized in such a way that the phenomenon is produced and is *dependent* upon them. Following this construal of mechanistic causation, RWT is to be formulated as follows. In order to establish causal claims, one needs evidence of both production and difference-making. Evidence of production comes from laboratory studies (and, ideal cases, it could also be inferred from population studies). Evidence of difference-making comes from both laboratory studies (the difference-making side of mechanisms) and from population studies. In other words, population studies and laboratory studies amount to two epistemic ways of access into the difference-making of the same causal relations.

We shall come back shortly to this important aspect of the ways of access into the difference making in §3, after discussing the problem of pre-emption. But before moving on to the pre-emption discussion, it should be said that, in spite of the quasi-consensus in the literature on mechanisms as being productive only, Joffe (2013) is not the only exception to have drawn attention to difference-making in relation to mechanisms.

One other exception is Woodward's interventionist account of mechanistic causation (Woodward, 2003) which has been insightfully applied to the biological realm in Waters (2007). This account, however, is dependent on the assumption of modularity (an assumption which might not be adequate for all mechanistic systems). Moreover, it does not seek to unite production with difference making but only to provide a methodology of tracking causation in terms of difference making (which thereby leaves out an important dimension of mechanistic phenomena); finally, it does not discuss the problems of pre-emption and of causation by absence.

Another exception is provided by anti-Humean authors such as Nancy Cartwright or Alexander Bird, who view causation as a universal, monistic phenomenon resulting from the manifestation of causal powers (or capacities). Cartwright has even coined a name for her capacities driven mechanisms, called 'nomological machines'. Since capacities or powers are modal properties whose manifestation entails a counterfactual expression (Bird, 2005), the difference making should attend every manifestation of powers, and this anti-Humean metaphysics is consistent with the simultaneous acceptance of process causation. However, neither Bird, nor Cartwright discusses the cases of pre-emption, and their arguments are dependent upon a particular metaphysics, i.e. the anti-Humean one.

Finally, we have a group of philosophers (e.g. Psillos, 2004, Strevens, 2013) whose views are close or complementary to the views of Bird and Cartwright. These philosophers do not adhere to anti-Humean metaphysics however, but focus more generally on the existence of scientific laws and their (counterfactually expressed or not) modal characteristics, which differentiate such laws from

accidental generalisations. In this framework, causation is viewed as difference-making since it is linked to (or derived from) the corresponding causal laws and their features of non-accidentalness. Strevens' contribution is particularly useful since it explicitly seeks to unify production and difference-making, and it fruitfully discusses the problem of pre-emption. Let us also have a look at this problem.[47]

§2 Pre-emption

The problem of pre-emption, although it has generated huge discussions, is easy to espouse in its basic lines. Imagine two billiard balls hitting one another. On a classical counterfactual analysis, this means that the first ball (or the player hitting it, depending on how one chooses to individualise the causal factors) is making a difference to the (moving of) the second ball, which is the effect. The counterfactual conditional in question would say that, had not the first ball hit the second, the second would not have moved. But now imagine a third ball being hit by a second player, just seconds after the first player hit his ball.[48] The ensuing trajectories are such that, had not the first player hit the first ball, this third ball would have hit the second one.

Once this third ball comes into the picture, everything changes, or everything seems to change. The reason is that the counterfactual analysis does not seem to work anymore in order to single out the first ball as making a difference to the second ball. Hence, it seems to follow that it cannot be called anymore a cause, on a counterfactual approach. Therefore, in order to maintain its status as a cause, one will have to follow the *process* linking the first and the second ball, and on this ground call the first ball a *production* cause, but not a difference-making cause any more. More generally, one will have to distinguish two types of causal relations - the production and the difference-making one, and claim that mechanisms are only concerned with production causation.

Now, first of all, it is worth noting a basic oddity of this inference from the existence of pre-emption to the existence of different types of causal relations (or to causal pluralism). The causal relation holding between the first ball and the second has changed once the third ball made in into the scene, even if, in the actual world, this third ball has no spatio-temporal contact or interaction with either the first or the second ball (it interacts with the second in the counterfactual scenario in which the first ball misses the mark or is not itself hit by the player). Out of a difference-making type of causal relation, the existence of the third ball has turned the interaction between the first and second ball into a production-only causal relation. How come?

[47] In spite of the various problems or shortcomings associated to the above difference-making accounts, the respective authors provide valuable contributions, and I will be using some of their insights in the next section.
[48] I am laying down directly a case of a so-called 'late pre-emption' scenario, which is the most difficult scenario of pre-emption. The solution I will draw will also work, *mutatis mutandis*, for case of early pre-emption.

Let me give a related example from the philosophy of causation to show how strange this is. Metaphysicians of causation are sometimes inclined to view singular causal interactions as not being ontologically primitive, since their status as causes seems to depend on higher causal laws (conferring them capacities or causal powers; Lewis, 1986, Bird, 2007). It is a rather thorny proposal, particularly since it is so difficult to imagine what ontic status scientific laws have. Nevertheless, such an argument is at least intelligible, since it is grounded on the ontological dependency between a higher and a lower level of factors/entities involved in causation (namely the particular causal factors interacting and the higher level law). But how could the presence of an entity with the same ontic level as those involved in the initial causal relation, and with no interaction or contact with them, change the status of the relation in question from one type of causal interaction (difference-making) to another (production)?

The sensible answer is that it could not change it. It is the same causal relation, with or without the presence of the third ball. Accordingly, if the difference making has characterized this relation before the third ball made it into the scene, one has to admit that it will characterize it also when the third ball is present. How to make this difference-making visible counterfactually? There is one easy solution at hand.[49] In the counterfactual scenario that should prove the difference-making of the first ball (by imagining that it has not hit the second ball), one should also remove the third ball. More generally, in cases of pre-emption in which several different causes target the same effect, in order to circumscribe the difference-making of a certain causal factor, one will have to remove all other causes from the scene.

One has to insist here that it is a quite straightforward solution, which reflects scientific practice. When this counterfactual scenario is made actual in the actual world by intervention, it is simply called experimentation. That is because, that experimentation entails the selection of a certain number of causes, in a specific, isolated context, and the elimination of any other factor (interference or whatnot) that might get in the way (along the lines of Mill's principles). Of course, objections could be raised but they are easy to answer with plausible replies, and we should not

[49] I am using here the basic intuition followed by proponents of difference-making causation when confronting the problem of pre-emption (e.g. Yablo 2002, Glynn 2013) namely that difference-making could still be in place if 'all thing are unchanged'. Because these proponents want difference-making only and reject production, their scenarios are unnecessarily complicated. My way of reverting to an unchanged context is by introducing an intervention or a series of interventions in order to eliminate the additional causes (mainly pre-empting, but also over-determining or neutralising). This solution resembles Woodward's intervention based methodology, which requires interventions on causes such that they induce corresponding variations in the effects, provided that the respective interventions could not have influenced effects on alternative pathways. My scenario is more straightforward though since it requires directly the eliminations of alternative causes, and it is ultimately inspired by scientific practice, as I will discuss in the main text above. Another way to say this, by invoking Mill's Methods, is that Woodward only appeals to the method of concomitant variation in framing his scenarios, whereas, by seeking to reflect actual practice of science in experimentation, my scenario appeals to the method of difference and the method of agreement, and is consistent with the method of residue and of concomittant variation.

forget that the burden of proof is on the side of the pluralist.

i) One could claim this solution cannot work as a means to *define* causes merely in terms of difference-making, i.e. it could not work as an analysis or definition of difference-making causation (Hall, 2004). A plausible response to this is that wishing to provide a 'pure' analysis of difference making causation - in which production would not show up at all - is putting the cart before the horse. It might be, on the contrary, that on a monistic construal of causation, the 'pure' causal relation to be defined is that which lets itself be decomposed into a production feature and a difference-making feature.

ii) One could also claim that, in order to eliminate *all* other causes from such a counterfactual scenario, one needs an independent way to *select* these other *causes*, before eliminating them, and it is difficult to articulate, using the possible worlds semantics, how the elimination can be meaningfully pursued (Lewis, 1973). The plausible response here is that one can well eliminate the other causes using background knowledge, for the *pragmatic* purpose of showing the way out of the pre-emption problem, even if difficulties remain with the possible words semantics and the issue of 'defining' difference making causation.[50]

iii) The third reason is that there might be cases in which the antecedent of the counterfactual could not be made actual in the actual world for physical reasons (an experiment involving the solar system needs to remain purely counterfactual for instance) or for metaphysical reasons (the alternative causes that would all have to be eliminated, could all be linked with the cause we are interested in such that no one of them could be withdrawn without the others disappearing as well or becoming non-functional, as for instance in non-modular, holistic systems; Williamson, 2011). However, Woodword (2002, 2003, 2011) has convincingly shown the meaningfulness of non-actualisable counterfactuals for physical reasons. Insofar as we are in a metaphysical or ontic discussion, the meaningfulness of such physically impossible scenarios is just the point we need here.[51] As for the cases in which the counterfactuals could not be made actual for *metaphysical* reasons, one could well use backtracking counterfactuals in order to make evident the existence of difference-making (Broadbent 2007), or one could use a non-counterfactual approach based on Mackie's INUS conditions (Strevens, 2007, 2012a and 2013).Since Strevens' approach not only provides a solution to the metaphysically intricate class of pre-emption cases, but also offers a way to incorporate all the intuitions that I have laid out above in my eliminative counterfactual scenario,

[50]See Woodward (2001) on how one can discuss causation methodologically/pragmatically, without going into the discussion of what causation ultimately *is*. I retain from Woodward's lesson that at least in some difficult, particular cases (as the cases of pre-emption are) one could adopt a pragmatic solution to solve these particular cases, and then pursue a more general discussion as to what causation ontically is.

[51]Of course, from a practical point of view, there is still a worry associated to such scenarios which are physically unrealizable in the actual world, and I will come back to this worry towards the end of this section.

and I will briefly describe it below.

The main insight of Strevens' approach, as said, is that one can view the difference-making causes as described by Mackie's INUS conditions, i.e. as necessary parts of sets of sufficient factors for determining effects (where this sufficiency can be expressed as an entailment relation between the corresponding statements of causes and effects). To see which factors are INUS conditions and hence difference-makers, given an actual state of affairs in which the effect obtain, one proceeds to successively set aside factors from the background, and checking each time whether the effect is still in place (and, I should add, one can easily recognize here the traces of Mill's method of difference).[52] In logical terms (which, as Strevens underlines, are *not* counterfactual) given a conditional whose consequent stands for the effect and whose (complex) antecedent stands for the background state of affairs at the time the effect is produced, one proceeds by successively setting aside (or abstracting away) the clauses of various background factors, and then checking whether the entailment relation still holds. If, upon abstracting away such a clause, the entailment relation still holds, we are not dealing with an INUS condition and the corresponding factor is not a difference-maker. If the entailment does not obtain, we are dealing with an INUS condition and the corresponding factor is a difference-maker (Strevens, 2007, §4).

There are two features of this procedure of abstracting away (the *kairetic* procedure, as Strevens calls it in his later publications) that are particularly important for pre-emption cases.

a) First, it differs markedly from the procedure employing counterfactuals. In the counterfactuals case, the background state of affairs needs to take into account the entire world state before the effect occurs and to draw possible words scenarios in which, by operating 'minimal', 'surgical incisions' on the world state, the putative causes are eliminated (which should be followed by the absence of the effect in question if the eliminated factors are difference-makers). In the *kairetic* procedure, the background state of affairs needs not be the entire world state. One can just focus on a particular (candidate) set of factors deemed to be jointly sufficient for the effect and check which of these particular factors are really INUS conditions or not (Strevens, 2007, §2).

In addition, causal factors are not strictly speaking eliminated from the background state of affairs. By setting aside clauses corresponding to particular factors from the antecedent of the background state of affairs, one does not eliminate these factors (i.e. one does not logically negate their corresponding clauses), but one abstract them away, one chooses not to take them into consideration. To use Strevens' example, abstracting away gravitation from the background state of

[52] Just to recall, Mill's Method of Difference says: if an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon.

affairs does not mean negating its clause in a possible world scenario in which gravitation does not exist; it means staying in the actual world and not taking it into consideration when drawing factual conditionals corresponding to cause-effect relations to be ascertained.

b) The relation of entailment that is under scrutiny in the kairetic procedure is not purely logical entailment (as in Mackie's original, empiricist account) but it is a relation of *causal* entailment. Strevens means by this that the factors taken into consideration in the background state of affairs need to be linked by a production causal process to the effect under scrutiny. Among other reasons, Strevens makes this move in order to solve a series of problems associated to Mackie original account - in which one could count as INUS conditions factors individualized by Goodman predicates. At any rate, this move transforms the kairetic procedure into a monistic analysis of causation, in which both production and difference-making are involved Strevens, 2007, § 5.1).

Now, cases of pre-emption, involving as they do actual (i.e. pre-empting) and potential (i.e. pre-empted) causes, can be described as cases in which multiple such sufficient sets of factors act or could act to induce the same effects. And there are two (interrelated) solutions that can be advanced in the terms of the kairetic procedure. The first is quite straightforward, based on the *a)* feature outlined above. One can just focus on one set of candidate sufficient set, and assess the INUS conditions of its components. In our example, one can pick out the set of conditions involving the first player, which are sufficient to entail the effect. Suppose that while hitting the first ball, the first player whistles.[53] Does his whistling count as a difference making (ontically) and as an INUS condition (logically)? This is an event that is arguably causally related by some sort of process to the event of the second ball moving (through the emitted sound waves, say). But the clause incorporating it into the background state of affairs is not an INUS conditions and the whistling is not a difference-maker, as it can be abstracted away while the relation of entailment holds. Is the hitting of the first ball a difference-maker? It is, because its clause cannot be abstracted away without breaking the entailment.

The second solution places more weight on the *b)* feature outlined above, in the sense that the pre-empted, possible causes of pre-emption scenarios are not strictly speaking related by causal processes to the effect under scrutiny. Remember that we are not in a possible world, counterfactual scenario, but we have remained in the actual world. And in the actual world, the pre-empted causes do not actually get to produce the effect under scrutiny (Strevens, 2007, §5.4). It is only the actual cause that gets to produce the effect.

Notice that these (interrelated) solutions work perfectly also for causes that are metaphysically inseparable in a counterfactual scenario, since we need not enter into a counterfactual scenario at all.

[53] I have interchangeably used in this section the terminology and examples of causal factors and causal events. There is some debate as to which terminology is the more appropriate and the reader should choose her favourite denomination.

One will either focus on the INUS conditions that includes the actual (pre-empting) cause, and/or, in case the pre-empted, metaphysically inseparable cause is also taken into consideration, one will be able to abstract it away on grounds of it not being linked by a process to the effect under consideration.

In more recent publications (2012a, 2013), Strevens has refined his kairetic approach by considering as starting point the fundamental processes of physics (Strevens, 2012a, pp. 451-452) and conceiving of the abstracting away procedure in such a way that absences, resulting from the attribution of negative properties, could also be conceive as difference-makers attending production relations (Strevens, 2013, p. 313). Since the cases of causation we are interested in concern the level of biology/medicine, we do not have to look at the possibilities of beginning the abstractive procedure from the level of fundamental physics.

Similarly, since the present discussion of mechanisms as difference making only requires that all production is also difference-making, we need not enter into the discussion whether all difference-making is also production, and accordingly need not look into his account of causation by absence. But it is worth just noting that medicine is rife with assertions of causation by absence. Take the example of the claim that lack of C vitamin causes scurvy. This claim that an absence is a difference-maker for a disease picks up a causal relation (or a series of causal relations) for which a description in terms of *production* is *also* readily available – the process of formation of spots on the skin, and of spongy gums, the bleeding from the mucous membranes, the defective collagen fibrillogenesis, etc. In fact, the story is much broader, because the whole field of medicine can be viewed as resulting from causation by absence. In contemporary medicine each and every pathological process admits a causal rendition in terms of the dysfunctioning, or the lack of normal functioning of a certain organ, system, tissue, type of cell, etc. Surely, we would not want to say that medical causation consists solely in difference making causation by absence which is separated from the processes of production causation (as ontic causal pluralist would have it, or as it would follow if indeed the problem of causation by absence was really a problem for ontic causal monism); the entire array of physiopathological processes involving *production* stands behind any claim that dysfunctioning or lack of functioning has produced such and such harmful effects.

One last word here about the problem of pre-emption as such. We have seen that Strevens carefully distinguishes his kairetic procedure from the counterfactual approach to pre-emption, and there are clear conceptual advantages that are thereby obtained. But there is one counterfactual approach that comes close, I believe, both to his insight and to his results, namely Broadbent's (2007) approach in terms of counterfactuals. Just like Strevens, Broadbent focuses on the sufficiency of causes rather than on their necessity, as in the classical counterfactual approach. But

instead of removing factors from sets of INUS conditions in order to test the sufficiency of these sets, Broadbent chooses to pick out the counterfactual scenarios in which the effect does not show up, and to infer by backtracking the absence of the actual difference-maker, which was sufficient to produce the effect in the actual world (Broadbent, 2007, p. 170).[54] The procedure works in cases of pre-emption, because arguably, the closest possible world in which the effect is absent is one in which actual pre-empting cause is missing and the actual pre-empted cause has not yet acted (Broadbent, 2007, pp. 177-182). In our billiard example, the closest possible world in which the second ball is not moving is one in which the first ball has not been hit, and this possible world scenario corresponds to the small (perhaps infinitesimal) period in which the third ball has not yet reached the second (admitting that the third was hit at all).

That the distance between Strevens' approach and Broadbent's is not so great can be seen by looking at Mill's method of difference, where the intuitions of both approaches are present. The method of agreement says that 'If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon' (Mill, 2002 [1843], p. 455). This is a method of inferring causes that supposes minimal changes in the causal background associated to the presence or absence of the effect (as opposed to Mill's method of agreement, for instance, which requires, a great variation in the causal background), and as it happens, Mill covers in the method of difference the possible discovery of both the sufficient cause, and of a necessary part of a sufficient cause (which corresponds to the INUS conditions). Broadbent and Strevens seem to pick upon the different ends of the same stick (which are both the right ends). Broadbent focuses on and starts from the absence of the effect, in order to infer the absence of the sufficient cause. Strevens focuses on and starts from the sufficiency of (the set of candidate INUS conditions making up) cause for the emergence of the effect, abstracting away various candidate INUS conditions and checking whether the entailment of the effect still holds (which is the same as 'testing' different conjunctions of INUS conditions to see if the effect obtains). Assuming the cause effect relation can be represented as a conditional, Strevens starts his approach from *modus ponens* whereas Broadbent starts his approach from *modus tollens*.

These are then the reasons why I am inclined to see some common ground between the approaches of the two authors. But be that as it may, they both offer insightful solutions to the problem of pre-emption, which offer satisfying, reasonable and plausible responses to potential

[54] Backtracking counterfactuals are not as uncommon as one might think. As Broadbent points out, even the Inference to the Best Explanation could be seen as applying a backtracking counterfactual conditionals (Broadbent, 2007, pp. 172-173).

difficulties associated to my scenario of elimination drawn at the start of this section. Of course details to be clarified remain (an entire alternative thesis could be written on this subject) and such details would probably be arduously picked up by the proponent of ontic causal pluralism. But again, the burden of proof lies on the latter's side, since causal pluralism is such an unnatural view to uphold, at least in the field of medicine, where it would lead us into great epistemological aporias and even skepticism about causation, as pointed out in the beginning of this chapter.

The position on pre-emption I have sketched above could be applied for mechanisms as a whole (when their difference-making is pre-empted, but also neutralised, or modified, by other causes or mechanisms) and, *mutatis mutandis*, for the causal relations inside the mechanisms themselves. On this approach, we end up with production mechanistic processes that are attended by difference-making, where the diference-making in question could express the sufficiency of the cause (if C was in place, E would be in place) and/or its necessity (if C was not in place, E would not be in place), and/or its backtracking, explanatory status for the effect (if E was not in place, C had not been in place), and/or the manipulability dependency to the effect (if the value of C was wiggled, the value of E would be wiggled as well). And again, all these conditionals approximately transpose Mill's Methods into the language of counterfactuals, where Mill's Methods, as I mentioned above, are basic postulates of any experimental research. I will look in the next section at how all this bears on the revised form of RWT I propose.

§3. The revised RWT

Now, if there is one clear advantage that looking into the pre-emption brings about in our discussion of difference making and mechanisms in medicine, is that it helps us to see the necessity of RWT and to grasp a better hold of the way in which the difference-making mechanisms contribute to the revised version of RWT I have proposed.

To recall from the previous section, first, I have propose a definition of mechanisms as a modified version of Illari and Williamson's (2012) – a mechanism for a phenomenon consists of entities *causal relations that are simultaneously productive and difference-making*, organized in such a way that the phenomenon is produced and is *dependent* upon them. Second, following this construal of mechanistic causation, I have proposed a revised form of RWT. This revised RWT says that, in order to establish causal claims, one needs evidence of both production and difference-making – where evidence of production comes from laboratory studies (and, in ideal cases, could also be inferred from population studies), whereas evidence of difference-making comes from both laboratory studies (the difference-making side of mechanisms) and from population studies.

A consequence of this reformulated RWT is that population studies and laboratory studies

amount to two epistemic ways of access into the same difference-making of the same causal relations. Why would we need two different epistemic ways of access in the same difference-making of the same causal relations? Which is the same as asking – how is the evidence from population studies and the evidence from laboratory usefully aggregated and fruitfully used in conjunction for the purpose of establishing causal claims?

This is a question with multiple answers, and some of them will require further elaboration in the following chapters.[55]At least the first answer that can be straightforwardly provided now – given the overview of pre-emption situations we went through – namely that *I)*evidence of difference making from the level of population studies can usefully complement the evidence of difference making obtained from laboratory experimentation in order to ensure that pre-emption and analogous situations do not distort our assessments of the causal actions of mechanisms. By analogous situations I mean cases in which other mechanisms, parts of mechanisms or causal factors act or could act, neutralizing, over-determining or modifying the effect of the mechanism, the part of mechanism or the causal factor we are interested in.

The simple eliminative scenario I have brought forth in the beginning of this section requires for the *actual* manifestation of the difference-making of mechanisms that the alternative causes be removed. I have noted this is a pragmatic solution, and there is no need to insist on the general appeal of its underlying intuition, which is present in various inferential techniques devised to delineate causal relations (including the Bayes nets approach). But there is still a pragmatic question to be asked: how can we be sure that all the alternative causes have been eliminated? Let us grant that in ideal situations, laboratory experimentation can proceed by complete isolation of the mechanism or part of mechanism under study (in case the mechanism in question is modular). But obviously this isolated context is different from the *in vivo* context in which mechanisms act within an organism. And the appeal of population studies at this point is precisely that they could offer an assessment of difference making in which the alternative causes and influences are eliminated or rendered as less interfering as possible. By randomization, RCTs can eliminate or reduce the bias associated to interfering causal contexts, and I should not lose the occasion of pointing out that they do so by putting into practice Mill's Method of Agreement.[56]

In addition, the interplay with the population studies allows us to solve a practical problem

[55] The work to be done in the following chapters, as related to RWT, has been alluded to in the introduction to the present chapter. Since we have here a suitable place for sign-posting, after the convoluted discussion of pre-emption, I will remind the reader the main thread of the thesis and how the following chapters discuss the advantages of the interplay between laboratory studies of difference-making mechanisms and population studies. The respective advantages will be numerotated using roman letters.

[56] To recall, Mill's method of agreement says that: if two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon.

noted above in relation to physically impossible counterfactuals. These counterfactuals have antecedents which cannot be realized in the actual world. The same goes with a vengeance for the metaphysically impossible counterfactuals that seem to be necessary to solve pre-emption cases in which the pre-empted and pre-empting cases being metaphysically inseparable, as was also noted above. The fact that, as Woodward has argued, physically impossible counterfactuals are still *meaningful* saves the day as far as the strictly ontic discussion of causal pluralism is concerned. On the other hand, for the pre-emption cases in which the pre-empted and pre-empting cases being metaphysically inseparable the other hand, Strevens' kairetic procedure allows us to make the point that ontic causal monism (and its consequence under focus here, that mechanistic production should be seen as inseparable from difference making) is not endangered even by such pre-emption cases.

As Nancy Cartwright used to remark with respect to her capacities - 'even when they are not manifested, they are still there' (Cartwright, 1999) –we can similarly hold that, even when the difference-making is not manifested in the actual world, it still remains an ontic or metaphysical feature of causal relations. Take another analogous example. Laws in the special sciences are notoriously applied *ceteris paribus*, and it has been argued that such *ceteris paribus* clauses should also be ascribed to the laws of the exact sciences, physics included, in order to specify for instance the lack of interference (Cartwright, 1983).[57] Are such *ceteris paribus* clauses - which seem to impose restrictions on the scope of law-like statements and on the manifestations of laws as such - a reason to strip the respective laws of their governing characteristic of universality and being exceptionalness? There are some strong arguments to answer *no* (Kline and Matheson, 1986). Laws should not be stripped of their law-hood by ceteris paribus clauses that seem to limit their *actual* manifestations even for the case of physics. Similarly, we can add, the mechanistic production causation should not be stripped of its difference-making dimension by pre-emption cases that seem to limit the *actual* manifestation of this difference-making.

But then, this point having been agreed on the strictly metaphysical or ontic side of the discussion, we are left, as I said, with a practical or methodological question. What to do in those cases in which the manifestation of difference-making cannot be enabled by actual interventions? The worry appears most evidently for certain cases of *internal* mechanistic causation, in the case of those biological mechanism that are not modular, i.e. in which the functioning of mechanistic items depends holistically on the entire mechanisms such that sub-mechanisms cannot be detached in order to isolate their actions. The worry is *much* diminished insofar as the output of mechanisms is concerned, since it is much *less* likely that the output of alternative mechanisms pre-empts or

[57] Anti-Humean accounts of laws have clauses stipulating the presence of stimuli and the lack of antidotes in the canonical form of their rendition of scientific laws, of physics included (see Bird, 2007).

interferes with the output of alternative mechanisms, in *such* a way that laboratory experimentation (complemented if needed by the randomisation of population studies) could not allow us to evaluate the difference-making of the mechanism of interest.[58]

Nevertheless, the worry should be dealt with, and the interplay with population studies specific of RWT offers a way out because, on a practical or epistemic level, the population studies offer us an average difference-making that reflects the difference-making of the mechanism we are interested in, from the point of view of entire human organisms.[59] And of course, the problem of pre-emption aside, we are always interested not just in the difference-making of mechanisms, as it can be assessed in isolated laboratory setting, but also in how this difference-making ends up being reflected in real-life, complexly interacting human organisms.

At this point, we need to face another obvious objection. Suppose that an EBM proponent accepts the entire above argumentation about the difference-making of mechanisms. Nonetheless, such a proponent will also surely claim (in line with the obvious preference of EBM theorists for population studies) that it is the difference making of population studies that does all the job for the purpose of testing and confirming causal claims, in contrast to the difference making of mechanisms obtained from laboratory studies. Mechanisms might matter in the context of discovery, when we learn initially about causal relations, but in the context of confirmation, we are only left with population studies, which have the entire epistemic superiority vis-a-vis the laboratory studies. So why all the fuss about the difference-making of mechanisms? Relatedly, it could be pointed out that, after all, we would be interested in mechanisms, and we would be interested in the difference-making of mechanisms, insofar as this assessing this difference-making is helping us assess causal relations in medicine. Why invoke after all the fact that the difference making of population studies could help us assess the difference making of mechanisms, as I maintained above? Are we not primarily interested in assessing the medical causal claims simpliciter (it is organisms which should be cured after all)?

---

[58] Broadbent (2007) observes that most philosophical examples of complicated pre-emption involve various strange entities and situations (wizards casting spells for instance, in the trumping cases) and that such complicated cases of pre-emption are much less likely to be encountered in real life. Cases of pre-emption have not deterred Kenneth Waters, for example, one one the most established philosophers of biology, to adopt a difference making approach (derived from Woodward's) for life sciences causation; see Waters (2007).

[59] A somehow similar point has been made by the proponents of RWT who maintain that evidence of difference-making from population studies can help us in cases in which evidence of mechanisms is unclear as to the overall direction of causation at the level of entire organisms, given multiple mechanisms that might have interfering pathways. Given our previous discussion in this chapter, an obvious observation is that, on the premise of causal pluralism (a premise entailed by the definition of mechanistic causation as productive only), it is not very clear how evidence of difference making causation could be relevant for the evidence of production causation (production causation being a different causal relation in the framework of causal pluralism). In our terms, of course, the evidence of difference-making from population studies gives us the overall direction of causation and the overall difference-making for a certain series of causal interactions linking two factors A and B (say, most generally, the treatment and the result of the treatment). A more rigorous comparison between the epistemic consequences of my revised RWT and the old form of RWT will be provided in the next chapter.

Having played the role of the devil's advocate, let me return to the angelic (or prosecution) side, because this objection offers the ideal background to present the other advantages of the revised form of RWT. We have looked above in some detail at *I)* the epistemic advantage that evidence of difference making from the level of population studies can usefully complement the evidence of difference making obtained from laboratory experimentation in order to ensure that pre-emption and analogous situations do not distort our assessments of the causal actions of mechanisms. But we also have the fact that *II)* the difference-making of mechanisms matters for individualizing the causal factors assessed by population studies. This simply means that *which* causal factors are to be ascertained in relation to *which* effects is substantially influenced by our mechanistic knowledge.[60] Of course, this is only possible in the framework of ontic causal monism, in which the productive and difference making causal relations of mechanisms are the same as the causal relations reflected in the dependencies of population studies. Another way to say this is that the classical distinction between the context of discovery and the context of confirmation should be carefully applied when it comes to the testing of medical claims. A population study backed up by detailed mechanistic knowledge will be much more finely grained and targeted in its results, than a population study backed up by approximate knowledge of mechanisms and roughly individualised causal factors

An almost direct consequence of *II*) above are *III*) and *IV)* –which indicate that mechanistic evidence and population studies evidence reinforce reciprocally their quality and weight when they cohere (again, justifiably so only in the framework of the difference making of mechanisms). It follows directly from *II*) above that *III)* evidence of population studies backed up by detailed mechanistic evidence should be qualitatively superior to the evidence of population studies backed up by poor mechanistic evidence. And, on a closer look, the reverse should also be true – *IV)* the mechanistic evidence whose findings are reflected in population studies assessments should be graded higher than the mechanistic evidence whose findings are not reflected in the assessments from the level of populations.

And this leads us to *V)* - the fact that the evidence of difference making of mechanisms, when backed up by evidence of difference making from the level of populations, allows us a better grasp of the problem of extrapolation. Recall the objection just stated above. Why invoke after all the fact that the difference making of population studies could help us assess the difference making of mechanisms, as I maintained above? Are we not primarily interested in assessing the medical causal claims *simpliciter*? Yes, we are primarily interested in assessing the medical causal claims *simpliciter*. But

[60] Notice the contrast the claim of Clarke *et al* in their (2014) that the main use of mechanistic evidence is to rule out spurious causation. We can see that the difference making of mechanisms allows us to have a much more rich view of the interplay with the population studies. This will be discussed in more detail in the next chapter.

it happens that mechanisms have been invoked as the main epistemic device of solving the difficult problem of extrapolation. It was, most probably too much of a charge to put on mechanistic evidence, and it has been argued that mechanisms can only but fail in this task. However, the revised RWT can usefully be used as a framework to improve upon the role of mechanisms in extrapolation, and it can do by a joint assessment of the difference making of mechanisms, coming both from laboratory studies and population studies. Or so I will argue in chapter 5.

Indeed, even if we will see how *I)* and also *II)-V)* preliminarily at work in the next section - where I discuss Howick's criticism and show that the revised form of RWT can plausibly withstand this criticism - the *II)-V)* epistemic advantages are worth and require further development. *II), III)* and *IV)* will be discussed in more detail in the next chapter with respect to the notion of the weight or quality of evidence, in a pre-confirmation context, and *V)* will be discussed in chapter 5.

Having mapped the epistemic advantages of the revised RWT which will be discussed in the next chapter, let us look next at Howick's criticism of the initial RWT.[61]

§4 Howick's criticism

In his (2011) Jeremy Howick lays down five (interrelated) lines of criticism against the initial form of RWT which views as necessary for establishing causal claims both the evidence of difference-making from population studies, and the evidence of mechanisms (i.e. of production only) from laboratory studies. Howick's claims are that:

a) in certain cases population studies are sufficient for prevention or treatment purposes, even if causal claims have not yet been established (and hence we need not appeal in addition to mechanistic evidence). (Howick, 2011, p. 930)

b) in certain cases population studies are sufficient for confirmation of causal claims (and hence we need not appeal in addition to mechanistic evidence). (Howick, 2011, pp. 931-932)

c) population studies cannot establish the existence of a mechanism (and hence in cases outlined in a) and b) above, one cannot argue that mechanistic evidence was still somehow taken in consideration when using population studies results). (Howick, 2011, pp. 933-934)

d) in certain (rare) cases, evidence of mechanisms from laboratory studies can be sufficient for establishing causal claims (and hence one needs not appeal to in addition to evidence from

[61] Too much signposting might be confusing for the reader, but just to say that chapters 6 and 7 will take up again *II)* and *III)* and will show their relevance for *confirmation* purposes, in the framework of a collaboration between IBE and Bayesianism. More precisely, chapter 6, which takes up and continues the rationale of *III),* will argue that - *VI)* mechanistic evidence can increase the resilience of probability functions for the hypotheses backed up by such evidence. Finally, chapter 7 will take up and continue the rationale of *II),* arguing that –*VII)* mechanistic evidence could constrain the prior and likelihood probabilities, for hypotheses backed up by such evidence.

population studies). (Howick 2011, p. 938)

e) mechanisms cannot solve on their own the problem of extrapolation (contrary to the claims of RWT proponents, e.g. Russo and Williamson, 2007) and, as a general conclusion, too much epistemic burden is put on mechanistic evidence by RWT (Howick, 2011, p. 934).

I will argue that some of these claims are based on a misunderstanding of RWT and the arguments motivating it. Some others, however, criticize successfully the initial form of RWT, even when taking into account Illari's (2011) disambiguation. Nonetheless, they are unsuccessful with the revised form of RWT I propose.

Claim a) is correct, but it just does not concern RWT. RWT is a thesis about *establishing causal claims,* whereas a) discusses cases of prevention or cases of medical treatment where causal claims have not been entirely established. In cases of prevention, naturally, if there is enough evidence of potential harm, medical action can be taken, *before* the causal action of the potentially harmful agent was established. Accordingly, preventative actions do not concern RWT. Similarly, in serious diseases, we might use treatments whose mechanism of action is not known, but which seem to have positive effects upon evaluation by population studies. But this does not mean that we should not *seek* to know more about the respective mechanisms, and Howick draws his examples from areas of medicine (psychiatry and neurology) in which we could really use more mechanistic knowledge. But in these cases in which we lack completely mechanistic knowledge are cases in which the causal claims have not been established, and we thus arrive at Howick's b) claim.

Claim b) is partly based on a misunderstanding, but partly it also hits the target. It is based on a misunderstanding because mechanistic knowledge, contributes to individualizing the causal factors assessed by population, and also because population studies can only be sufficient to establish causal claims in ideal situations in which we would have at disposal for RWT an infinite randomization. Outside of such ideal situation, in real life that is to say, one also needs evidence of production from mechanisms in order to rule out spurious causation (Russo and Williamson, 2007).

But partly, claim b) hits the target, because the old form of RWT cannot explain why is it that, in such ideal situations, the evidence of population studies would be sufficient. It cannot explain it because, as I have argued in the section 1 of this chapter, by sticking to the definition of mechanisms as productive only, the consequence of ontic causal pluralism is unavoidable. Hence, from evidential pluralism (the advisable position to take concerning evidence) one cannot fall back on ontic causal monism - the advisable ontic position to take, in which evidence of both production and difference making concerns the *same* causal relation – but only on ontic causal pluralism. On the production only definition of mechanisms (either of mechanistic causation or of mechanistic evidence) evidence of production concerns production causation only, whereas evidence of

difference making concerns difference making causation, which is bound to be a different causal relation from the productive type. In other words, evidence from population studies and evidence of mechanisms turn out to concern different causal relations. And it follows finally, that in the ideal case of infinite randomization for population level RCTs, evidence of difference making would be sufficient to establish causal claims.

This is directly relevant for the claim c) above, that population studies cannot establish the existence of a mechanism. Given a framework of ontic causal pluralism and a definition of mechanisms as productive only (either concerning mechanistic evidence or concerning mechanistic causation – the consequences are the same), population studies cannot establish the existence of a mechanism, and Howick's criticism hits the target.

Importantly, it hits the target even when taking into consideration Illiari's (2011) disambiguation of RWT. As stated in the *Introduction* to this thesis, Illari distinguishes two senses of evidential pluralism, with respect to the *type* of evidence (either production or difference-making, and with respect to the *source* of evidence (either population studies or laboratories studies). This disambiguation of the meaning of evidential pluralism also disambiguates RWT. In Illari's view, RWT requires no longer evidence of difference making from population studies and evidence of mechanisms (i.e. of production) from laboratory studies (as in the initial version set out in Russo and Williamson, 2007). It requires instead evidence of difference making that could come either from population studies or laboratory studies, and evidence of mechanisms (i.e. production) that could come either from population studies or laboratory studies.

Yet, this disambiguation is not sufficient to respond to the criticism in c), because mechanisms are still considered by Illari as productive only (see also Illari and Williamson, 2011), and we have all the consequences of ontic causal pluralism outlined above. In other words, we are told that evidence of mechanisms could be searched for at the level of population studies but the conceptual terms of the discussion do not allow that this evidence could ever be found there. If mechanisms are concerned with production only, and accordingly with the production type of causal relation, how could evidence of them be found on the level of populations, which give us evidence of difference-making (probabilistic dependencies established on population studies can only mean difference-making), and hence are concerned with the a different type of causal relation, the difference-making one?

The revised version of RWT I have proposed, which requires for establishing causal claims evidence of both production and difference-making, while evidence of production comes from laboratory studies, whereas evidence of difference-making comes from populations studies *and* laboratory studies (on mechanisms), since mechanisms are concerned with *both* production *and*

difference-making, can withstand the criticism of claim c), because it is unmistakably grounded in ontic causal monism. Hence, in the ideal case in which we would have infinite randomization for population level RCTs, the establishing of the causal relation allows to infer the existence a mechanism, producing this same causal relation.

Interestingly, by insisting that evidence of population studies could never give us insight into the existence of a mechanism, Howick turns out to adhere implicitly to the framework of ontic causal pluralism - because only in this framework such an inference from population studies to mechanisms is impossible. This echoes the point I made earlier in §1 of the present chapter, that the general position of EBM theorists on the role and value of mechanistic evidence indicates embracing ontic causal pluralism (although an overly conscientious pursuit of the empiricist ideal of 'value-free', 'objective' science could also be at play).

Claim d) is also correct and it hits the target. In certain (rare) cases, evidence of mechanisms is sufficient for establishing causal claims. This contradicts evidently the initial form of RWT. It is however consistent with my revised RWT. Evidence of a mechanism provides evidence of both production and difference-making, which is just what is required by the revised RWT. These cases are rare because, most often, evidence of difference making coming from mechanisms should be backed up by evidence of difference-making coming from population studies, for the reasons laid out at the end of the previous section.

And cases of difficult extrapolation are certainly cases in which we need evidence of difference making from both laboratory studies looking at mechanisms, and population studies. We arrive thus at Howick's claim d), which is partly based on a misunderstanding, and partly hits the target. It is based on a misunderstanding because the very logic of RWT demands not that evidence of mechanisms be used alone (even if the rare cases in which is used alone can be accommodated) but that this mechanistic evidence be attended by population studies whenever possible. In my scheme, it would also be necessary that one had as great and detailed knowledge of a mechanism as possible (*pace* Darby and Williamson, 2011). However, claim e) also partly hits the target, because RWT proponents have claimed that mere knowledge of the existence of a mechanism is sufficient for extrapolation purposes (Darby and Williamson, 2011). The latter position is based on a view of mechanisms as productive only. But as I will argue in chapter 5 - looking at another of Howick's contributions (Howick *et al.* 2013) and taking also in account the discussion of extrapolation in Clarke *et al* (2014) – we need the difference-making of mechanisms, in combination with the evidence of difference making from population studies, in order to be able to respond to the difficulties raised by Howick's critique.

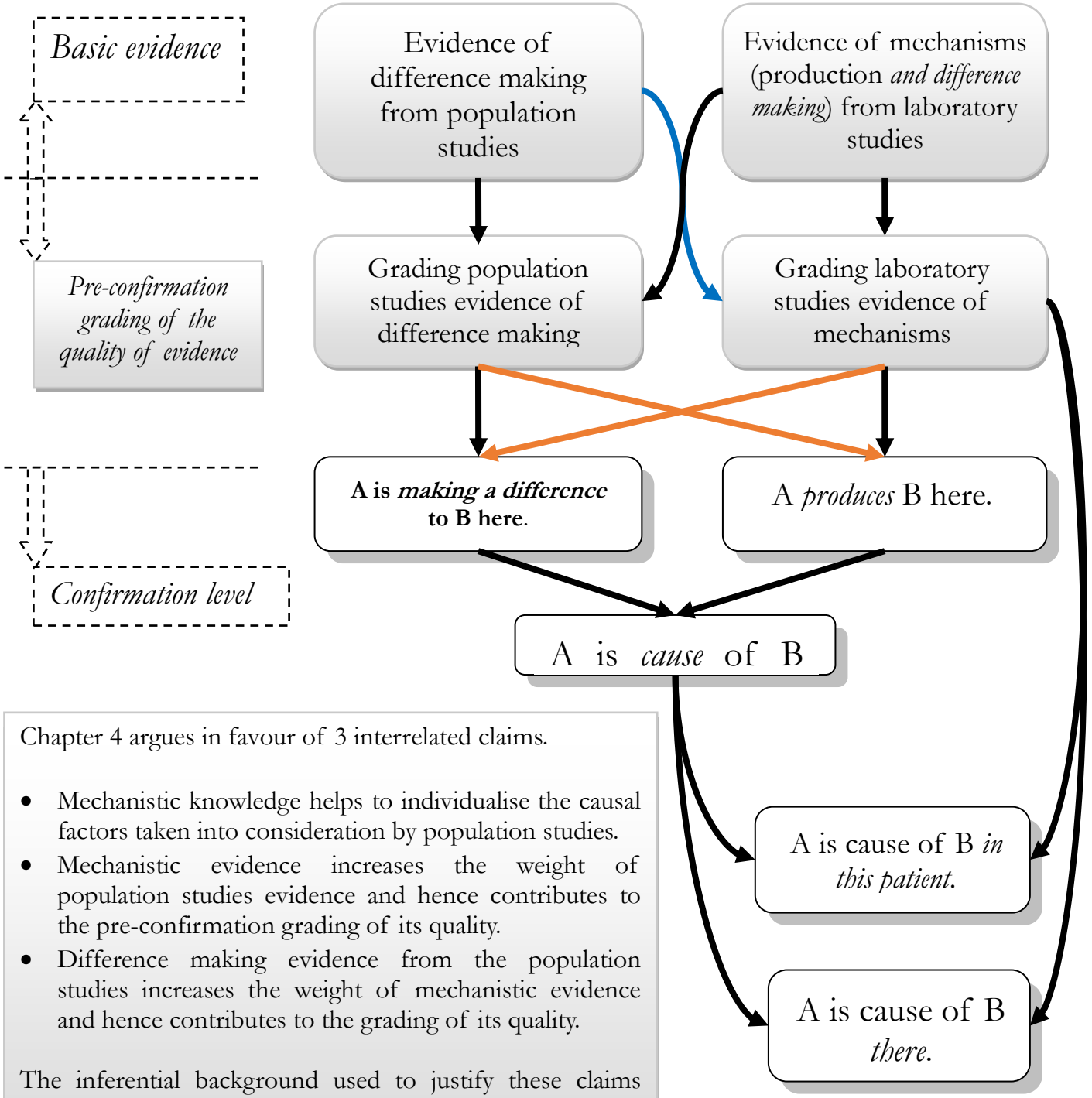
However, before embarking on the problem of extrapolation, the next chapter will allow us

to take a closer look at the evidential interplay between laboratory and population studies, in the RWT terms I have proposed.

Conclusion chapter 3

I have argued that difference making should be viewed as part of mechanistic causation. I have proposed accordingly a difference-making definition of what a mechanism is (modifying the definition of Illari and Williamson, 2011) and a revised form of RWT. I have discussed the problem of pre-emption and I have argued that it does not pose a threat to my view of mechanisms. Finally, I have defended the revised form of RWT against criticism raised in Howick (2011).

Chapter 4



Chapter 4 argues in favour of 3 interrelated claims.

- Mechanistic knowledge helps to individualise the causal factors taken into consideration by population studies.
- Mechanistic evidence increases the weight of population studies evidence and hence contributes to the pre-confirmation grading of its quality.
- Difference making evidence from the population studies increases the weight of mechanistic evidence and hence contributes to the grading of its quality.

The inferential background used to justify these claims concerning the pre-confirmation grading of evidence is Lipton's interpretation of IBE as a guide to confirmation.

Chapter 4. The weight of evidence and the evidential interplay between populations and mechanisms

Introduction

The previous chapter has listed a number of five epistemic advantages of RWT in the ontic framework of mechanisms as difference-making. These are that *I)* evidence of difference making from the level of population studies can usefully complement the evidence of difference making obtained from laboratory experimentation, in order to ensure that pre-emption and analogous situations do not distort our assessments of the causal actions of mechanisms, *III)* mechanistic evidence could increase the weight of population studies evidence and hence could contribute to the pre-confirmation grading of its quality, *IV)* difference-making evidence from the population studies could increase the weight of mechanistic evidence and hence could contribute to the grading of its quality, and *V)* evidence of difference-making from population studies could fortify the evidence of difference-making from laboratory in order to identify robust mechanisms, which should be better prepared to face the problem of extrapolation.

I) was discussed above in the context of the pre-emption discussion, and we have seen *II)-V)* preliminary at work in the above section on Howick's criticism of RWT. However, *II)-V)* are worth further discussion, both in order to clarify their content, and in order to underlie the contrast between the revised form of RWT and the previous use of RWT, in particular as deployed in Clarke *et al.* 2014. The present chapter will look in more detail at *II)*, *III)* and *IV),* while the next chapter will take up *V)* (and also, as we shall see, *I),* from a different perspective).

Now, the basic intuition behind *II)* is that *which* causal factors are to be ascertained in relation to *which* effects in population studies is substantially influenced by our mechanistic knowledge. In turn, the basic intuition behind *III)* and *IV)* is that mechanistic evidence from laboratory studies and population studies evidence reinforce reciprocally their quality and weight when they cohere. In other words - *III)* evidence of population studies backed up by detailed mechanistic evidence should be qualitatively superior to the evidence of population studies backed up by poor mechanistic evidence, and conversely – *IV)* the mechanistic evidence whose findings are reflected in population studies assessments should be graded higher than the mechanistic evidence whose findings are not reflected in the assessments from the level of populations.

Among others, these intuitions should lead us to extend the criteria of evaluating the quality and weight of mechanistic evidence we have discussed in chapter 2. The Clarke *et al.* criteria - which were analysed and justified in chapter 2 using *T*IBE (the *testimonial* usage of IBE) - concerned

strictly the laboratory studies in which mechanistic evidence is obtained. However, given the basic thrust of RWT, we should expect that evidence of population studies have some role to play as well in grading mechanistic evidence (and vice-versa, contrary to the ethos of EBM).

One justification for why this mutual reinforcement of quality and weight of evidence should be in place (or rather, the conceptual possibility of this mutual reinforcement) was laid out in the previous chapter, in relation to the construal of mechanisms as both productive and difference making - in which in which the productive *and* difference making causal relations of mechanisms are *the same* ontically as the causal relations reflected epistemically in the dependencies of population studies. This construal of mechanisms also makes possible the individualization by mechanistic knowledge of the causal factors taken into consideration by population studies.

In the present (short) chapter, two additional justifications are put forward (maintaining and continuing to take advantage of the construal of mechanisms as difference-making). One justification is straightforward – the fact that in the practice of medicine we can actually find this mutual reinforcement of quality, as well as this way of individualizing the causal factors. The other justification uses the framework of IBE, as in chapter 2 - with the difference that instead of the testimonial side of explanatory inferences, it will appeal to what Lipton has dubbed 'the guiding role' of IBE, as well as to IBE's capability of taking into account both the balance and the weight of evidence.

As it might be recalled from chapter 1, Lipton acknowledged that in most cases of theory confirmation, given its preponderantly *qualitative* use and conclusions, IBE will not be sufficiently fine-grained to allow us to confirm or disconfirm the hypotheses at stake.[62]Accordingly, Lipton afforded to IBE a more modest role, namely that of guiding, of pointing out towards the right theories. In plain terms, this is not a *confirmation* role but amounts to a supporting contribution *before* the confirmation stage. However, as it happens, even if this guiding role cannot be sufficient in itself for confirmation purposes, it suits perfectly the pre-confirmation stage of grading the quality of evidence with which we have been concerned in chapter 2, i.e. it is perfectly suited as a justificatory framework of the criteria or 'rules of thumb' that should point towards quality evidence.

In the particular context of this chapter, the criterion to be justified is that mechanistic evidence backed up by population studies evidence is of a better quality than mechanistic evidence that is not backed up by evidence from population studies. It is an 'external' criterion for mechanistic evidence, as it were, in contrast to the 'internal' criteria of mechanistic evidence of

[62] An exception to this are the Inferences to the Only Explanation, which are basically IBEs that apply to particularly rich evidence, being able to eliminate the alternative candidate hypotheses and pick out the right one. Chapter 1 has provided some exemplifications as applied to mechanistic evidence.

Clarke *et al.* (i.e. criteria that concern solely the laboratory studies). The 'guiding use' of IBE will also justify a corresponding 'external' criterion for evidence of population studies, saying that population studies evidence backed up by mechanistic evidence is of better quality than population studies not backed up by mechanistic evidence (or less detailed mechanistic evidence).[63]

The rationale of justification for both of the 'external' criteria above will be straightforward. In Lipton's *causal* interpretation of IBE, we are of course seeking to infer the right explanation, but given the role played by Mill's criteria, we could directly state that we are looking for the causes (or causal theories) to explain effects. And it will turn out that the two 'external' criteria are justified because, at bottom, what they state is that having multiple epistemic ways of access into the cause or into the cause effect-relations, is better than just having one single epistemic way of access.

With respect to the form of argumentation, given the *guiding* role of IBE in Lipton's same interpretation, IBE will provide the respective justifications by drawing its explanatory conclusions with an 'as it were' clause. The argumentative scenario will be to suppose that IBE was allowed to pick out the right hypothesis, in cases in which the rival hypotheses seem to have the same balance, but a different weight. Again, given the guiding role of IBE, the fact that hypothesis H_1 would be picked out by IBE as the right theory (rather than its rivals), will just mean that 'as it were', theory H_1 is the confirmed theory. At the *pre*-confirmation level of grading evidence of mechanisms, this will *actually* mean that the evidence in favour of H_1 has more *weight* than the evidence in favour of the rival hypotheses. And this will mean in turn that it is of a better *quality* than the evidence in favour of the rival hypotheses - thereby justifying the two 'external' criteria of grading evidence mentioned above, which, as I will show, are already (more or less tacitly) employed in medical practice.

In the following chapters, we will move beyond this pre-confirmation level. In chapter 5, the 'external' criterion of grading mechanistic evidence will be put to use for *extrapolation* purposes. In chapters 6 and 7, the 'external' criterion of grading population studies evidence will be put to use for *confirmation* purposes, in the context of the collaboration between IBE and the Bayesian theory.

It should be underlined that in the context of the present chapter, we just cannot move beyond the pre-confirmation level and the 'as it were' clause of IBE conclusions cannot be eliminated. That is to say, this guiding use of IBE cannot be a confirmation use, given the predominantly qualitative aspect of IBE-type inferences. To draw confirmation conclusions, one needs also the quantitative aspect, in order to distinguish the cases in which the balance is indeed equal between rival hypotheses, to adjudicate in cases in which the balance is not equal, and, when the weight of

[63] Or, mechanistic evidence established using less different methodologies. We could employ at this point all the 'internal' criteria of mechanistic evidence of Clarke *et al,* in order to better define the role that mechanistic evidence plays in the 'external' criterion for the evidence of population studies.

evidence comes into play, to reflect this qualitative side of evidence in quantitative terms. Chapters 6 and 7 will defend in this respect a form of collaboration between IBE and Bayesianism, in which the explanatory part of confirmation inferences does justice to the weight of evidence by constraining priors and/or likelihoods, and by increasing the resilience of hypotheses backed up by mechanistic evidence.

The present (short) chapter then marks our final discussion in the present thesis of the pre-confirmation stage, and will be structured as follows. §1 will lay down a necessary, preliminary discussion of the relation between IBE and the weight/balance of evidence pair of concepts, using as an example an episode from the early history of atherosclerosis; in addition, it will prepare the way to for looking into tacit criteria of grading evidence that are used in medical practice. §2 will look at the evidential interplay between mechanistic studies and population studies, in the background of the revised RWT, showing how the IBE-based assessment of the weight of evidence points towards the reciprocal reinforcement of quality for mechanistic studies and population studies (thus treating points *III)* and *IV*) from the above classification of the epistemic advantages of the revised RWT). Examples will also be drawn from the history of atherosclerosis research. §3 will draw contrast with the previous treatment of RWT in Clarke *et al.* (2014) and will discuss the way mechanistic knowledge can individualise the causal factors taken into consideration by population studies (thus treating point *II)* from the above classification of the epistemic advantages of the revised RWT).

§1 IBE and the balance/weight distinction

As mentioned in the Introduction to this thesis, one of the crucial conceptual distinctions in the epistemology of evidence is that between the balance and the weight of the evidence, the latter depending crucially on the reliability of evidential sources (Joyce 2005, Kelly 2008 and 2014). A strong balance of evidence for a certain outcome in a chance set up, for instance, might have little weight if the operation or experiment is performed just a few times. Repeating the experiment would have the consequence of increasing the weight of the evidence, even if the same outcome was obtained (the balance of evidence remaining the same). To come back to the causation cases, a population level correlation might have a strong balance, and yet, for various reasons, its weight might turn out to be quite weak—the size of the population might be too small, or the other potential causal factors might not have been sufficiently screened off (by not choosing the right subjects in a case control or observational study, or by not randomizing and double blinding accurately in an RCT).

It is not very often mentioned that the differentiation between the balance and the weight of

evidence was put forward by John Maynard Keynes, who was, among many others, an early proponent of (logical) probabilities. In 1921, Keynes wrote: 'As the relevant evidence [for a hypothesis] at our disposal increases, the magnitude of [its] probability may either decrease or increase, according as the new knowledge strengthens the unfavorable or favorable evidence; but something seems to have increased in either case—we have a more substantial basis on which to rest our conclusion.... New evidence will sometimes decrease the probability of [the hypothesis] but will always increase its "weight"' (Keynes, 1921, p. 77, *apud* Joyce, 2005, p. 158).

The worry was that in computing probabilities, one might lose sight of the qualitative side of evidence, of how reliable the sources of evidence are. Indeed, initially, Keynes just did not know how to incorporate the weight into his formulas (Cohen, 1986, pp. 264-267), with Karl Popper alleging subsequently that the notion of weight gives rise to insoluble paradoxes in the theory of probability (O'Donnell, 1992, pp. 44-47). Parenthetically, the balance/weight distinction remains problematic for probabilistic theories of evidence assessment. Bayesian theorists still have problems with how to calculate the balance of evidence (with different measures of it being proposed),[64] and continue to be accused often of not taking (sufficiently) into account the weight of evidence.[65]

Is it the same for IBE? As we have seen in the chapter 1 - when exemplifying the use of IBE in mechanistic laboratory research— the weight or quality of evidence is always taken into account when inferring the best explanation. Suppose we have two theories H_1 and H_2 that seem to have the same balance. If H_1 is preferred to H_2 on explanatory grounds, this means that in the actual process of inferring that H_2 explains better the overall evidence than H_1, it was already taken into account that the sources of evidence in favour of H_1 were superior in terms of reliability to the sources of evidence in favour of H_2. As we have also seen in chapter 1, in which Harman and Psillos' comparison between enumerative induction and IBE-based ampliative induction was presented, one does not first draw the explanatory inference that 'all As are Bs' or that 'As cause Bs' and then enquire whether the samples in favour of As being Bs, or of As producing Bs, are biased or not. Whether the samples are biased or not is really part of the explanatory inference that adjudicates between different hypotheses at stake. This is so both for the rare cases in which IBEs allow us to actually confirm a hypothesis, by being able to rule out all alternative hypotheses (what I have called

[64] For an overview, see for instance, Glass (2012). Since, as mentioned in chapter 1 and chapter 2, I do not take IBE to be a rival of Bayesianism but rather its companion in those situations where the numerical data is not sufficient or the evidence is rather qualitative than quantitative, I will not go into the proposals that IBE itself can offer numerical expressions of the measure of confirmation, advanced by Glass himself in his (2007) and (2012), and also more recently in by Douven and Wenmackers(2015) and Douven and Schupbach (2015), Douven and Schupbach (2015b) interesting as these proposals may certainly be. My purpose in this section is to show that IBE does not have a *conceptual* problem with the balance/weight distinction—in other words, that it can coherently deal with and incorporate these aspects of the evidence.

[65] See for instance Joyce 2005 (and also the discussion in chapter 6). One possible solution for the Bayesian is to derive the weight of evidence from the balance, but it is problematic, since the weight of evidence entails a qualitative aspect that the balance *per se* does not contain.

in chapter 1, following Alexander Bird's terminology, Inference to the Only Explanation), and for the 'as it were' cases in which IBEs have just a guiding role.

Take an example from the early stages of the atherosclerosis research - the strong correlations noticed in rabbits in early XXth century by Anitschkow and other researchers (like Wesselkin, Wacker and Hueck) between the level of cholesterol and the atherosclerotic lesions in arteries (Kritchevsky, 1995). The general medical community did not consider these rabbit findings as reliable, good quality source of evidence. Why? Because, among other things, the atherosclerosis results obtained from animal experimentation with various other species were quite ambiguous, i.e., the other results either did not show a correlation or showed only a very weak one (Furie and Mitchell, 2012, p. 2185). The argument for not accepting Anitschkow's findings could well be interpreted as IBE with a guiding role (where the best available explanation in that period was the senescence theory), an inference that would count rabbit experiments in vivo as unreliable source of evidence of mechanism(s).[66]

Now, what is rather striking is that how the grading the rabbit evidence as lacking quality (or as having a low weight) contributed to the conclusion (drawn by the medical community at the time) that the senescence theory is to be preferred was kept more or less tacit. It is a fact that medical researchers rarely make entirely plain their rules of reasoning (or they used to until very recently). The situation is not ideal, and we could usefully recall here Psillos' (ironical) comments on the search for the Holy Grail of scientific confirmation and reasoning that seems to be present in recent philosophy of science: one would like to have rules that are transparent and algorithmic, and whose following would just be a matter of grasping their logical form (and the Bayesian theory, adds Psillos, seems to respond, at least in part, to this ideal). However, in messy real-life situations, in which most often, scientist employ explanatory reasoning, what we have are the (to a certain extent) imprecise and tacit models of the IBE. However, scientific explanations, claims Psillos, are imprecise and tacit because they are hardly amenable to an abstract, context-free algorithmic form; yet, they should be sufficiently precise insofar as one looks at real-life cases, uses the background knowledge of each science, and pays enough attention the particular whole context experimental situations (Psillos 2007, pp. 441-447).

The case of medicine certainly seems to suit Psillos' description. There are tacit ways of using explanatory inferences in medicine, both with respect to the quality evaluation of the evidence and with respect to the confirmation of hypotheses. At the same time, this does not mean that one should not try, as much as it is possible, to make explicit the explanatory inferential rules at stake, even if one will not be able to reach the ideal of abstract, context-free rules. In our case, what we

[66] Indeed, pace authors like Steinberg, the decision not to accept the lipid theory at that stage of research and evidence was justified.

should do is to look in particular at the tacit rules of evaluating the quality of evidence, in order to make explicit how the evidential interplay between mechanistic studies and population studies is explanatorily understood in inferences that take into account the weight of evidence.

To illustrate what I mean - when Bradford-Hill articulated his nine criteria of medical causality in 1956, there was, of course, a clear display of his gifts as an exceptional statistician. However, at the same time, his work spelled out a series of criteria that were already present in the epidemiological medical research—tacitly or less clearly and systematically articulated. As argued in chapter 1, IBE in general seems to correspond de facto to scientific research and evaluation. If then there are tacit criteria and rules for taking into account the weight of evidence with respect to the interplay between laboratory studies and population studies, then looking at how IBE reflects previous research into mechanisms and population causal claims might well help us to bring to the surface and articulate these more or less tacit criteria and rules. It is what we will do in the next section.

§2 Weight or quality of evidence, and the interplay between populations and mechanisms

What I am seeking to make explicit from the history of atherosclerosis should be no secret to the reader by now. I am looking for mutual reinforcement of quality between laboratory evidence and the population studies evidence, as well as illustrations of the individualisation of causal factors in population studies, driven by mechanistic knowledge (underlined by explanatory reasoning). We do not have to look very far. The key moments in this history offer readily relevant material.

Take first the way in which population studies influenced the mechanistic research. What happened was that after the sceptical period of Anitschkow in the 1920's and 1930's, and a lack of research activity until after the end of the WWII, a series of population studies in the 50's maintained interest in the cholesterol hypothesis, even if there were no major mechanistic discoveries in the respective period. The population studies were not conclusive, but they showed interesting correlations between the level of cholesterol and lipids in general, and cardiovascular events, sometimes in quite distinct environments and contexts.

To take the clearest example, a famous (in medical circles) study called 'The Seven Countries', which started in 1958, took into account epidemiological data from countries as diverse as Finland and Japan, looking at the correlation between the cholesterol diet intake and the frequency of cardiovascular events (Steinberg, 2007, pp. 34-35). Such population studies addressed obviously the numerical, quantitative basis of the correlation between cholesterol (and lipids in general) and cardiovascular diseases (i.e. addressed the 'balance' of the evidence for the cholesterol hypothesis, so

to speak. But given the diversity of patients taken into consideration, they addressed also the qualitative side of the evidence (i.e. its 'weight') in favour of the respective hypothesis. And the increase in weight brought about by studies such as the Seven Countries study and others (the Framingham Heart study, began in 1948, the Minnesota trial, began in 1947) influenced also the mechanistic research and evidence.

The main problem of mechanistic evidence, as mentioned in the previous section, was its heterogeneity, its decreased qualitative aspect –although strong correlations were observed in one type of animal experimentation, much weaker or no correlation were observed in the other types of animal experimentation, involving other species. In spite of these problems, research into the mechanisms associated to cholesterol continued in the 50's and the 60's, due to the interesting results of population studies. In other words, the increase in weight of population studies in that period, due to the diversity of evidential sources taken into consideration, transferred onto the mechanistic evidence and research, whose problem had been precisely a low weight due to lack of diversity in evidential sources.

Take second the way in which laboratory research advancement influenced the population studies. One evident way in which, in general, laboratory research advances (and in which the corresponding quality of evidence increases) is expressed in the Clarke *et al* criterion of the number of features known of a mechanism. Discovering entities or intermediate chains in the mechanism itself - along the micro-structural research on mechanisms whose reliance on Mill's methods I discussed in the chapter 1, and in perfect agreement with the understanding of mechanisms as difference-making I have outlined in chapter 3 – increases the weight of mechanistic evidence. But does it not increase also the weight of population studies evidence?

It should do so, given that although population studies are a different *source* of evidence, it is evidence *of* the same causal relations that laboratory studies are concerned with, in our framework of causal monism. And what the history of atherosclerosis shows is that, when, starting from the mid-60's, more and more chains in the mechanism of atherosclerosis were discovered (e.g., various types of lipoproteins, the LDL receptor, the scavenger receptor), the hypothesis of the pathogenity of cholesterol gained more acceptance*, even if* the general results from the population level remained the same (in other words, even if the *balance* of evidence afforded by the population studies remained the same; Stehbens, 2001).[67]As David Steinberg, one of the scientists involved actively in research at that time, observes, one just needed a decently detailed mechanism, in order for the

[67] What Stebhens shows is that, due to the success in mechanistic research, the medical community tended to favour the trials showing cholesterol significance and to disconsider the trials with different results; see also Ravnskov (1992). Unfortunately, the history of atheroscelosis has had such episodes, and the continued animosity between proponents of the general view and contesters did not help much to to clarify the issue of which studies were (or should have been) taken into consideration.

causal hypothesis formulated initially by observing correlations on the level of entire organisms, to be accepted (Steinberg, 2007, p. 89).

Why is it that discovering more intermediate chains within mechanisms contributed to the general acceptance of the cholesterol hypothesis (which had become already in the 80's a hardly disputed dogma) and increased the weight of the evidence in favour of the cholesterol causal claims? On an IBE approach, the answer is straightforward—because mechanisms explain, and the more fine-grained a mechanism supporting a certain hypothesis is, the more explanatory the hypothesis becomes in comparison with other hypotheses.[68] This is guiding IBE inference, that does not take the responsibility of confirming the more explanatory hypothesis, but justifies the 'external' criterion of grading higher the quality of evidence from population studies when the evidence from laboratory research converges in similar results.

Why is it that the weight of evidence of mechanistic evidence was increased by corroboration from population level correlations? Because, an IBE-based answer would go, what we look for in evaluating hypotheses is ultimately to find the right causes, or the right causal law in order to explain effects, and having two epistemic ways of access into the causes at hand (i.e. laboratory research into mechanisms, and population studies) makes a hypothesis more explanatory when these two epistemic sources converge in their results. And this justifies the 'external' criterion of grading higher the quality of evidence of mechanisms from laboratory studies when the evidence from population studies converges in similar results.

Naturally, this reasoning only works in the framework of the revised RWT given ontic causal monism, and when viewing mechanisms as both productive and difference making. In the next section, I will compare my scheme of justification for the two 'external' criteria of grading evidence with the treatment of the interplay mechanisms-population studies provided in Clarke *et al* (2014), in the framework of the initial RWT, in which mechanisms are viewed as productive only. This will allow us to bring to the fore the last advantage of the revised RWT to be discussed in the present chapter, namely the individualisation by mechanistic knowledge of causal factors dealt with in population studies - what was classified in the beginning of the chapter as the advantage *IV)* of the revised RWT - and will help us make the transition to the next chapter discussing extrapolation.

§3 Individualisation of causal factors

In order to see this, consider how this interplay is portrayed in Clarke *et al.* The authors argue that both mechanisms and population level assessments should be taken into account when

[68]The reader might recall that in the first and second chapters of this thesis, I have underlined that 'individuation' or 'precision', as Psillos calls it, is one of the main explanatory virtues employed in IBE.

evaluating evidence in medicine for two main reasons a) evidence of population studies can clarify complicated mechanistic evidence and b) mechanistic evidence can rule out spurious correlations established in population studies.[69] In a bit more detail:

a) when mechanistic evidence is unclear due to the problems of complexity and masking, population correlations could help. The problems of complexity and masking say that for mechanisms with too many internal paths and for mechanisms cancelling each other's effects, it might be hard to detect the causal relations as such and also the direction of causal relations (whether A causes or prevents B, or has no influence at all on B, or finally, whether B does not in fact cause A in turn). In these situations, population studies could provide the net contribution of a putative cause to a putative effect and the direction of causation.

'Even where a mechanism linking A to B is well established and known in some detail, it can be hard to infer whether A has a positive effect on B, or A prevents B, or indeed whether A has any net effect on B at all. This is particularly true in cases where the mechanism is complicated: where there are several links on a pathway from A to B or where there are several pathways from A to B. It is also a problem where a mechanism is known to be non-robust over time or over other changes in situation. It is typically evidence of correlation that is crucial for determining whether any causation is positive or negative and what the net effect is. Thus evidence of mechanisms should be used in conjunction with evidence of correlation, not on its own, to infer causal claims.… The human body is a complex system, and the more we discover about it the more it seems that it is very common to have multiple mechanisms operating. If there are multiple mechanisms operating, they may impact on each other, and one or more may mask the effects of the mechanism you have discovered' (Clarke *et al.* 2014, pp. 349-351).

b) in turn, mechanistic evidence can rule out the danger of 'false positive' results due to spurious correlations recorded on a population level. This danger arises because *positive* correlations might be due to confounding or accidental co-variation, due to variables that are semantically connected, to logical, physical and mathematical connections. In these cases, proof of a mechanism or of the non-existence of a mechanism could rule out the 'false positive' correlations—the true causal claims are 'clinched' by evidence of a mechanism (Clarke *et al.* 2014, pp 340-351).

These two points are important and they suit the parts of the history of atherosclerosis we looked at in the previous. a) After the ambiguous laboratory results from Anitschkow's period due to the complexity and masking of the mechanisms at work in different animal models,[70] the population studies initiated after WWII showed a correlation sufficiently robust between various lipidic factors and cardiovascular events to warrant further mechanistic research. b) The plethora of mechanistic discoveries of intermediate chains backed up the above correlations and contributed to the increasing acceptance of the hypothesis that high cholesterol is pathogenic, because the

[69]There are some other reasons, like the fact that mechanistic knowledge allows the application of general schemes of treatment to particular patients, or that mechanisms help with the problem of extrapolation. As mentioned, I will look at extrapolation in the next chapter, in which I will also consider the views of Clarke et al on this matter.

[70] Indeed, in those early animal experiments, the ambiguous results were due to the fact that some species possess a very effective mechanism of metabolising cholesterol, such that their cholesterol level could not possibly have corellated to atherosclerotic formations.

presumption of spurious correlations on the population level was, for the most part, withdrawn.

Clarke *et al* have drawn points a) and b) in the framework of the initial RWT and by viewing mechanisms as productive only. There would be several (interconnected) contrasts to underlie (some of which will be developed in the following chapters) when comparing their scheme of the evidential interplay between population studies and mechanistic research with the scheme of the revised RWT.

The first contrast to be underlined is that, while a) and b) are valid points, which recognised in the evolution of medical knowledge, they could not justified in the scheme of Clarke *et al* and in the framework of the initial RWT. As argued in chapter 3, the initial RWT, insofar as it is associated with a definition of mechanisms as productive only is ontically a causal pluralistic account. This means that the difference making and the production causal relations are *different* causal relations. Hence, mechanistic evidence could not justifiably bear upon the evidence from population studies, and neither the latter could bear upon the former. They could bear upon each other in an ontic monistic framework, which entails that mechanisms should also be difference-making.

The second contrast is that Clarke et al. do not discuss this evidential interplay in terms of the balance/weight distinction, which is present in the analysis provided above as a consequence of the revised RWT. Apparently, a) corresponds in content to what I have termed above the 'external' criterion of grading mechanistic evidence, whereas b) seems to correspond in content to what I have termed above as the 'external' criterion of grading population studies evidence. Undoubtedly, there is an area of overlap (and, as throughout the present thesis, the work of Clarke *et al* has been a source of inspiration, even if, given the necessity to distinguish my claim, it has been the critical part that has mostly emerged). But I would maintain that the understanding of the two 'external' criteria in terms of reciprocal increase of *quality* or *weight* is more fruitful than putting forward the establishment of the nett effect or direction of causation as a) does, or the ruling out of spurious causation, as b) does.

As far as a) is concerned, if adopt the revised RWT understand it as referring to the nett *difference making* (what else could measure as nett the dependencies of population studies?) then the idea of increase in quality or weight of the mechanistic evidence makes sense because population studies, in conjunction with laboratory research, could be used to differentiate *fragile* from *robust* mechanisms. A robust mechanism would be a mechanism which manifests its difference making even when the causal context changes, as opposed to a fragile mechanism. This might matter substantially for the issue of extrapolation, and I will look at this aspect in more detail in the next chapter, chapter 5.

As far as b) is concerned, there is more to the contribution of mechanistic evidence to

population studies evidence than ruling out spurious causation. Suppose spurious causation is ruled out. But is not is more that mechanistic evidence contributes? If there is an increase in weight, in different degree, beyond the stage at which at which spurious causation has been ruled out, then this increase of weight might well matter for confirmation purposes, and I will look at this aspect in chapter 6.

Finally, as far as both a) and b) are concerned, sticking to a definition of mechanisms as productive only makes us loose an important phase in the interplay between population studies and mechanisms. It is the phase in which the discoveries from laboratory studies are used to refine and focus the enquiries at the population level.

Recall that according to Clarke *et al.*, the problems of masking and complexity are problems of *mechanisms*, which can be solved by appeal to population correlations. This is entirely true, but the problems of complexity and masking concern not just mechanisms - these problems could show up *also* at the level of population assessments. And it might happen, on the contrary, that when the issues of complexity and masking are present on the level of populations, knowledge of mechanisms in turn could help solve these issues from the population level.

Indeed, what is interesting about the early history of atherosclerosis is that not *all* population assessments showed proof of correlation between the causal factors at stake. Some did not show any correlation at all. Whereas, in some cases, this failure to obtain conclusive results can be attributed to various set-up biases,[71] in other cases, no significant methodological flaws seem to have been in place. For instance, the 1968 British Medical Research Council (MRC) Study showed no significant benefit at all of lowering cholesterol for cardiovascular events; in the two groups, the difference of events was 74 versus 62 (Steinberg, 2007, pp. 54-55). Steinberg points out that, due to the small number of patients taken into consideration (400), what the study showed was that the effect was less than a 50 percent reduction in cardiovascular events. However, mixed results have always existed, and apparently, they tended to be overlooked, especially after the 1970s.[72] Moreover, real-life examples aside, as a matter of principle, population level studies, even the most sophisticated double blinded randomised ones, cannot always discover correlations.[73] This could well have been the case in the history of atherosclerosis research because, in such a complex, multifactorial disease—involving several systems, including the metabolic and the cardiovascular

[71] For instance, consider the lack of patient compliance in the Finnish Mental Hospitals Study, 1968, which reached levels of significance for women in just one of the hospitals. Similar lack of compliance was reported for the Minnesota Coronary Survey, 1968 (Steinberg 2007, pp. 36-37)

[72] Even in 1992, Ravnskov was arguing in the British Medical Journal that 'apart from trials discontinued because of alleged side effects of treatment, unsupportive trials were not cited after 1970, although their number almost equaled the number considered supportive…. [A]uthors of papers on preventing coronary heart disease by lowering blood cholesterol values tend to cite only trials with positive results' (Ravnskov, 1992, pp. 15-19).

[73] See Worall (2007), Worall (2010) for discussion.

system, and lifestyle and diet elements—taking into account all the causal factors involved was (and perhaps still is) very difficult.

What this suggests, again, is that the problems of complexity and masking can manifest themselves not only at the level of mechanisms but also on the level of population correlations, i.e., that these problems can 'propagate' from the microstructural level to the macrostructural one. Two mechanisms can cancel or mask each other, and the pathways of a mechanism might be so complicated that the nett result and difference making is not visible on a population level. It is for this reason that, in the overall interplay, mechanistic discoveries can help population assessments not just by singling out, among the positive results of populations studies, the ones that are truly causal (and hence ruling out the 'false positive' results). Mechanistic discoveries can also help population assessments by reconfiguring the space of causal hypotheses when the results from the population level are inconclusive, or even show no correlation at all.

A good example is the case of the 1969 Wadsworth Veterans Administration Hospital Study, in which due to the fact that strokes and advanced peripheral arterial disease were not included among the putative effects of atherosclerosis, the results did not reach statistical significance. Once evidence of mechanisms was gained showing that cerebral strokes are also consequences of atherosclerosis, the results of the study were retrospectively evaluated as statistically relevant (Steinberg 2007, pp. 184-186). Yet another case in point is the whole array of population studies that followed the discovery of lipoproteins, the LDL receptor, the scavenger receptor, and the other crucial mechanistic chain in the physiopathology of atherosclerosis (Steinberg, 2007, pp. 102-104).

A final observation that is in order here is that the case of the 'propagation' of the problems of masking and complexity to the level of population studies could be easily overlooked if one adopts a definition of mechanisms as productive only, given ontic causal pluralism that follows from the latter position. Why there could be a 'propagation' of the problems of masking and complexity to population studies is clear, if one adopts the definition of mechanisms as both productive and difference-making, and one insists accordingly that the mechanistic causal relations are *the same* relations that are reflected epistemically in population level studies. If these causal relations are *the same*, then whatever epistemic problem might show up at one level of epistemic access into these relations, i.e. the laboratory studies into mechanisms, could also show up at the other level of epistemic access into these relations, i.e. the population studies.

And, with respect to the issue of mechanisms and pre-emption, which we discussed in the previous chapter, it is somehow ironical that mechanistic knowledge could help solving the problems of complexity and masking from the level of populations. The existence of possible pre-emption, as well as of analogous problems like complexity and masking, was brought forward by

proponents of the initial RWT as part of the argument that mechanisms are productive only and that accordingly we need the difference-making of population studies in order to solve or compensate for this problems (Russo and Williamson, 2007, and also Clarke *et al.* 2014). The argument is ingenious, and I have also appealed to a form of it in chapter 3, in order articulate the epistemic advantage *I)* of the revised RWT. To briefly recall, *I)* states, on the hand, that when experimental interventions are not possible in order to make manifest the difference-making of particular mechanisms, one can appeal epistemically to the overall difference-making of entire organisms, assessed by population studies, in order to establish overall causal claim, and on the other hand, that, even when experimental interventions are possible, one can use population studies to eliminate confounding and make manifest the difference-making of mechanisms.

However, to come back to our present discussion - given the issue of the 'propagation' of the problems of complexity and masking at the level of population, we are facing a complementary scenario in which, on the contrary, it is the causal knowledge from the level of mechanisms that could help disentangle complexity and masking from the population level. In such a complementary scenario, if there are experimental interventions that can solve the problems of masking and complexity, and/or that discover additional, intermediate chains in mechanistic interactions, the mechanistic knowledge thereby acquired can pull population studies out of the limbo of ambiguous, inconclusive results by reconfiguring the space of causal hypotheses, individualising the causal factors taken into account and/or controlled for in population studies, and helping to divide the tested population into the right groups.

This is also a sort of increase in the weight of evidence brought about by mechanistic studies. One consequence of it, which I will look at in chapter 7, is that explanatory considerations drawn from such mechanistic studies could be used to constrain the priors and the likelihood of hypotheses adjudicated by the Bayesian theory of confirmation.

Conclusion chapter 4

I have shown that IBE can deal with the balance/weight distinction, and that the revised form of RWT can reveal to us three interesting features of the interplay between mechanistic studies and population studies - which is what was designated in the previous chapter as epistemic advantages *II), III)* and *IV)* of the revised RWT. I have argued that - *II)* mechanistic knowledge could help to individualise the causal factors taken into consideration by population studies, that - *III)* mechanistic evidence could increase the weight of population studies evidence and hence could contribute to the pre-confirmation grading of its quality, and that - *IV)* difference-making evidence from the population studies could increase the weight of mechanistic evidence and hence could contribute to the grading of its quality.

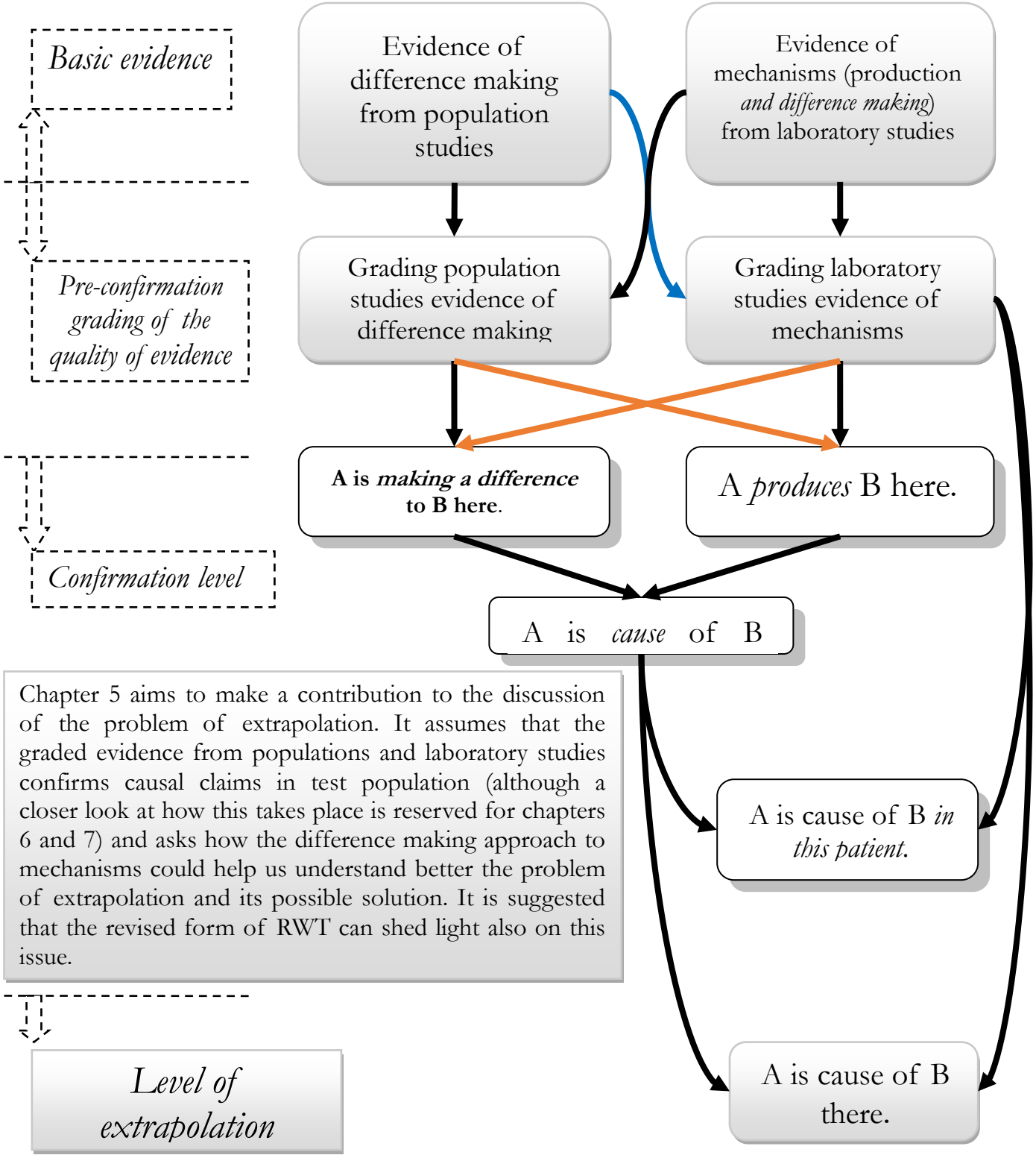
From the next chapter on, our discussion moves from the pre-confirmation level of grading evidence to the level of establishing and extrapolating causal claims. But the results and lines of reasoning of the present chapter will be taken up, heuristically used and continued.

The next chapter will take up the epistemic advantage *IV)* stating that population studies could increase the weight of mechanistic evidence, and will show how the evidence of difference-making from population studies could fortify the evidence of difference-making of mechanisms, such that the latter be better prepared to face the problem of extrapolation – the latter being, in our classification of the advantages of the revised RWT, the epistemic advantage *V)*.

Chapter 6 will take up the epistemic advantage *III)* stating that mechanistic evidence could increase the weight of population studies evidence, and will argue, in the framework of a collaborative use of IBE and Bayesianism for confirmation purposes, that *VI)* - the increase of weight brought about by mechanistic evidence would influence the resilience of probabilities functions of hypotheses established by the Bayesian theory taking into account population studies evidence.

Finally, chapter 7 will take up the epistemic advantage *II)* stating that mechanistic knowledge could help to individualise the causal factors taken into consideration by population studies, and will argue, in the framework of the same collaborative use of IBE and Bayesianism for confirmation purposes, that *VII)* - mechanistic evidence could be used employed to constrain the prior and/or likelihood probabilities established by the Bayesian theory taking into account population studies evidence.

Chapter 5



Chapter 5 aims to make a contribution to the discussion of the problem of extrapolation. It assumes that the graded evidence from populations and laboratory studies confirms causal claims in test population (although a closer look at how this takes place is reserved for chapters 6 and 7) and asks how the difference making approach to mechanisms could help us understand better the problem of extrapolation and its possible solution. It is suggested that the revised form of RWT can shed light also on this issue.

Chapter 5. What difference could mechanisms make for the problem of extrapolation?

Introduction

The present chapter aims to have a closer look at the problem of extrapolation, which, as we have seen in chapter 3, is the subject of one of the main charges advanced by Howick in his 2011 against RWT. The subject deserves a closer look because RWT strictly speaking, has been concerned with *establishing* causal claims. On the other hand, RWT proponents have spoken to a less extent about the claims of *extrapolation*. I aim to show that my construal of mechanisms as difference-makers can throw new light on the problem of extrapolation as well, and can show us in a clear way how RWT extends in such cases (which is what was designated in the previous chapter as epistemic advantage *IV)* of the revised RWT). Finally, this chapter aims to put more flesh onto the bones of one crucial criterion of grading mechanistic evidence advanced by Clarke *et al.*, namely the criterion of 'robustness/fragility' and shows how this criterion can be usefully understood in relation to the problem of extrapolation.

Generally speaking, in the philosophy of science, the problem of extrapolation is the problem of using results obtained in specific testing contexts, in order to justify inferences about the same results holding in untested, target contexts. This problem has been discussed from various and detailed perspectives(Guala 2010, Jimenez-Buedo 2011, Cartwright 2007) and the discussion has received a vital impulse through Daniel Steel's work on process-tracing (Steel, 2008), which claims that a suitable use of mechanism can solve this problem, although serious criticism has been mounted against his approach (Reiss 2009 and 2010).

In the philosophy of medicine, the basic insight behind the whole mechanisms has followed Steel's approach and could be resumed as follows - given certain causal results, obtained either in laboratories (animal experimentation or other physiopathological assessments), and/or in controlled studies, these causal results could be justifiably extrapolated for other target populations, if we knew that the testing and the target contexts share similar mechanisms (Thagard 1999, Steel 2010). Proponents of RWT have follows the same line (Darby and Williamson 2011, Clarke *et al* 2014). In addition to his (2011) criticism of RWT - which touches upon the problem of extrapolation, as we have seen in the previous chapter - Howick also has authored, in collaboration, a large-scale critique addressed exclusively to the use of mechanisms in extrapolation (Howick, Glasziou, Aronson 2013). This critique synthesizes both the general difficulties of Steel's approach, and also takes into

account the particularities of medical mechanisms.

Howick *et al.* argue in their (2013) that our understanding of mechanisms is often incomplete, that medical mechanisms can behave erratically, and that Steel's well-known methodology of process tracing fails. The criticism in Howick *et al.* (2013), as I will show, is successful when applied to the very brief treatment of extrapolation provided by RWT proponents in Clarke et al (2014). My main proposal in this chapter is that, when viewing mechanisms as difference-making, the *revised* RWT has a natural extension for the realm of extrapolation, and can withstand this criticism. This extension of the revised RWT says that, when dealing with such untested, target contexts, one needs evidence from both mechanistic studies and population-studies, in order to circumscribe *robust* mechanisms - where robust mechanism are mechanisms whose *difference-making* is manifested, or is likely to be manifested, even when the causal context changes.

The reason why the extension of the revised RWT can withstand the Howick *et al* (2013) criticism is that both Howick *et al* and the proponents of Steel's approach assume the construal of mechanistic causation as being just production. Accordingly they miss an aspect of mechanistic causation that should be vital for extrapolation claims (an indication of this being that the very form of extrapolating claims is counterfactual – were an organism with such and such mechanism be present in such and such different circumstances, it will produce such and such effects).

The main argument I will defend in this chapter - that the evidence of difference-making from population studies could fortify the evidence of difference-making of mechanisms and accordingly circumscribe robust mechanisms which are better prepared to face the problem of extrapolation – takes up and continues the argument from the previous chapter on the epistemic advantage *II)* of the revised RWT – namely that population studies could increase the weight of mechanistic evidence.

The plan of the present chapter is as follows. In §1, I present the problem of extrapolation in more detail, appealing most importantly to the pivotal contribution of Steel (2008) as well as to the criticisms raised in Howick *et al.* (2013). In §2, I look at a case study proposed in Clarke *et al.* (2014),and show that the criticism raised in Howick *et al.* (2013) applies to their case study and discussion as well. In §3, I discuss why viewing mechanisms as difference makingshould allow us to better conceptualise the problem of extrapolation. In §4, I come back to the case study provided by Clarke *et al.*(2014), and show that it can be more charitably understood as adequately describing the role of mechanisms in extrapolation, once the construal of mechanisms as difference making is taken in.

§1 Extrapolation and mechanisms in medicine

When are we warranted to extrapolate our causal claims, i.e. to use causal results obtained in specific *testing* contexts in order to in order to infer that the same results hold in *untested*, *target* contexts?[74] The paradigmatic cases of concern for extrapolation are those of experiments made in the highly protective media of laboratories, whose results might or might not hold *outside* laboratories, in untested contexts. The behaviour of electrons, for instance is of course well established for any potential interaction, in the isolated media of experimentation. But outside such media, the behaviour might be different, and Nancy Cartwright has illustrated how this change of behaviour outside the testing area might take place (Cartwright's 1999, pp. 58-61). In fact, Cartwright has made the point from very early on in her research that such erratic behaviour might show up for the entities and activities we are accustomed to think as governed by the immutable laws of physics (Cartwright, 1983). Since her work initially dealt with physics, her worries on extrapolation problems might have seem to some commentators as exaggerated or rhetorical (e.g. Kline and Matheson, 1986). However, in the special sciences, the issues of extrapolation are clearly present, and they need to be dealt with. It is still Cartwright who was among the first to draw attention to the pitfalls of engineering social and economic policies and applying them to contexts in which local factors might render them either inefficient or harmful (Cartwright 1999, 2010).

In medicine, the problem of extrapolation shows up both for population studies (RCTs included) and for the results of laboratory experiments. All results of animal experimentation, as well as of biochemical, biophysical, genetic, physio-pathological experiments (on the side of laboratory research research) and all results of controlled studies made in order to test drugs and prophylactic measures, etc., (on the side of population level research) - all can be subject to the worry of extrapolation. How can we make sure that when we move from the tested to the untested contexts the causal relations we have established remain the same? The thought is that, in principle, the test and the target contexts should share approximately similar *mechanisms*. But then the difficulty just seems to resurface on a different level. How can we make sure that these two types of context share similar mechanisms?

One option would be to study in detail the mechanisms in the target context. However, in this case, the whole point of trying to extrapolate causal claims seems to vanish, and the research done initially in the testing context appears to be redundant. It would be redundant because we

[74] The same problem is sometimes formulated in terms of "validity" as the problem of using the "internal validity" of test results to justify inferences regarding their "external validity". In light of Jimenez-Buedo's (2011) challenge to the biased utilisation of the concepts of "internal validity" and "external validity", I will speak from now on in terms of extrapolating causal claims from *tested* contexts (i.e. the *results inferred* from certain experiments in well-circumscribed circumstances) to *untested,* or *target* contexts (see Jimenez-Buedo 2011, pp. 274–275). I will make an exception from this terminological usage when directly quoting or paraphrasing the authors I discuss, who trust for the most part the concept of validity.

would not anymore project the causal relations from the testing context into the target one, but we would directly establish the existence and nature of these causal relations in the target. This is what Daniel Steel has dubbed as the problem of "extrapolator's circle" (Steel, 2008, p. 4). A related challenge is that the mechanisms in the study and target context might be just partly similar (as often is the case in animal experimentation for human medical purpose) and yet this partial similarity might just be sufficient in some cases to back up extrapolation claims, provided that the differences between the mechanisms in the target and in the study context are not significant. But then, when are such differences significant and when they are not? This is, again in Steel's terms, the "problem of differences" (Steel 2008, 78–79).

According to Steel's process tracing approach, what one needs to do is first to learn the mechanism in the model organism by reconstructing step-by-step the production paths between end points (either initial cause or final effect) and identifying intermediary nodes. Then, instead of an overall comparison with the mechanism in the target population, one selects for comparison specific stages and nodes, which background knowledge indicates as most likely to differ between the model and the target. The nodes selected for comparison could often be situated close to the end-point, or to final output of the mechanism in question. The rationale is that differences upstream the nodes in question matter only if they generate differences further downstream. If the mechanistic production issued at these critical points or nodes is similar, this offers a warrant for the extrapolation claim. If the production is dissimilar, the extrapolating claim should be viewed with greater suspicion. Admittedly, by looking just at the critical nodes, we leave aside what is going on upstream in the mechanism. But that is the catch. One needs not know everything about the mechanism in the target in advance, and the "extrapolator's circle" appears to be avoided. Similarly, since the relevance of any differences upstream the critical nodes only matters if the output at the critical node is modified, the "problem of differences", also seems to be avoided (Steel 2008, pp. 89-90).

Steel's proposal has faced various objections. One set of objections have been put forward by Howick, Glasziou and Aronson, who argue that his process tracing does not avoid the "extrapolator's circle." Against Steel, Howick *et al.* argue that the method of process-tracing depends on knowledge of whether the *significant* nodes have been chosen for comparison between model and target (as for instance, in order to rule out that intervening on the mechanism under study does not trigger an alternative mechanism whose production goes downstream by bypassing the node in question and changing the final output). However, such knowledge can only be obtained by studying in detail the target and evaluating the consequences of intervening upon the mechanism under study. In other words, either the process-tracing does not work as a method for extrapolating

causal claim (in case the desideratum not to know in detail the mechanism in the target is maintained) or one has to drop the above desideratum, and one falls into the extrapolator's circle (Howick *et al.* 2013, pp. 283, 286).

More generally, Howick *et al.* argue that in medicine mechanisms cannot solve the problem of extrapolation, because it is rarely possible to identify *all relevant mechanisms*, because studies of mechanisms themselves (whether in animals or humans) suffer from their own problems of external validity, and finally because mechanisms can behave erratically. First of all, ignorance of all relevant mechanisms is proven by recent medical history. For instance, based on mechanistic considerations, patients were treated anti-arrhythmic drugs after myocardial infarction, which a clinical trial showed to increase significantly the mortality from arrhythmias or cardiac arrest, as well as the all-cause mortality (Howick *et al.* 2013, pp. 282-283). Then, studies of mechanisms themselves suffer from their own problems of external validity because they are not governed by the type of universal laws we find in the exact sciences. Whereas most definitions of mechanisms in the philosophical literature emphasize regularity and stability of the final output, medical research shows that results obtained in laboratory research do not always show up in evaluations on a population level, as in the case of the putative capacity of *Hypericum perforatum* (St. John's wort) to reduce the concentration of androgenic steroids, by inducing the activity of cytochrome P450 isoenzymes. Finally, the lack of regularity and stability is epitomised by the paradoxical behaviour of certain drugs, which can worsen the condition they were supposed to alleviate, as for instance the antidepressants or antiepileptic drugs (Howick *et al.* 2013, p. 284).

Now, Howick *et al*'s criticisms are powerful, but not decisive. They also seem to apply, as I will show in the next section, to the case study provided in one of the latest contributions of RWT proponents, namely Clarke *et al.* 2014, to which I turn next, before moving on to show, in §3, §4 and §5 why, after all, Howick *et. al.*'s arguments are not decisive, in the framework of RWT.

§2 Clarke *et al*'s case-study

In its primary scope, RWT is concerned with *establishing* causal claims. In Clarke *et al.* 2014, RWT proponents also discuss, albeit briefly, the problem of extrapolation. Since my approach to this problem supposes the framework of the (revised) RWT, I need to touch here upon their own treatment as well.

Citing Steel, the authors maintain that mechanistic evidence has a role in ascertaining the external validity, as it can indicate how medical drugs and measures work or should work in the test population, and whether, and to what extent, mechanisms from the test population are also present

in the target population. As an illustration, they take the case of the hypertension treatment with calcium-channel blockers (CCBs). The example is important, and I provide below the full quote.

"Mechanistic *evidence helps to ascertain the external validity of treatments*. Mechanistic evidence can indicate how the intervention works (or is supposed to work) in the test population, and whether, and to what extent, such mechanisms are also present in the target population (i.e. outside the trial) – see Steel (2008). A good example of this can be seen in clinical guidelines governing prescribing practices for anti-hypertensive drugs in the UK. Recent research has suggested that different drugs should be used for patients from different ethnic groups. NICE guidelines therefore state that treatment should differ depending on ethnicity: Offer step 1 anti-hypertensive treatment with calcium-channel blocker (CBB) to people aged over 55 and to black people of African or Caribbean family origin of any age... (NICE 2011c, p.5) This recommendation was based on RCTs that had been designed to test the efficacy of different treatments in these ethnic groups. *In turn, these trials were based upon the plentiful evidence suggesting the operation of different pro-hypertensive mechanisms operating in different ethnic groups* (see NICE 2011b, pp. 248-250 and citations)" Clarke *et. al.* 2014, p.347, italics added

According to Clarke *et al.,* this anti-hypertension guideline was based on RCTs testing the efficacy of drugs in different ethnic groups and ages (in this regard being cited the study of Kshirsagar 2006, which showed a higher risk of hypertension for older people and people of African or Caribbean family origin). In turn, Clarke *et al.,* claim that these RCTs were based *on evidence of different mechanisms* for hypertension in different ethnic groups. That is to say, according to Clarke *et al.*, this evidence of such different mechanisms for hypertension should have been pivotal in extrapolating claims about the efficacy of calcium-blockers (CBBs) from the model of their test population, to the groups of elderly, Caribbean and Afro-American patients, with RCTs playing the role of control for this basically mechanistic extrapolation claim.

Clarke *et al.* are entirely right that in the last decade at least, medical research has furnished us plentiful mechanistic evidence not just suggesting, but proving the operation of different pro-hypertensive mechanisms in different age and ethnic groups. But if we look further in the history of cardiological treatments, what we discover is that it was *not* the mechanistic evidence which triggered a specific treatment for these special groups. It was, on the contrary, the *failure* of an extrapolation claim, based on apparently well-established mechanistic evidence for the use of beta-blockers (BBs), i.e. the drugs used before the CBBs.

BBs have an interesting history of their own which I should briefly recall, since their other uses are yet another illuminating illustration of how apparently safe mechanistic knowledge can hide epistemic surprises, and are also suggestive of a certain aura of "wonder drug" that BBs used to possess. BBs were synthesized in 1962, following the discovery of the beta receptor in 1948, and were used initially to reduce stress in patients with angina pectoris, and also for arrhythmias. In the early 1970's, they started to be used as anti-hypertensive treatment with evident success. The mechanism was fairly well understood. BBs diminished the hearts' capacity to pump by decreasing the simulating action of the sympathetic system over the cardiac muscles.

It is worth recalling briefly from chapter 2 that beta-blockers even came to be used in heart

failure – that is to say, in a clinical context in which the heart pumps much less blood than normal. Previously, they were formally contraindicated in heart failure (how could one use a drug that decreases the pump function when there is already not enough pumping?). In other words, the mechanistic knowledge available at the time was entirely against their use in conditions with diminished ejection rate of the heart. The beneficial effect in heart failure was discovered accidentally, in 1973, when it was administered to a 59 year-old woman patient with tachycardia caused by acute pulmonary oedema owing to dilated cardiomyopathy, and this dramatically improved her condition (Waagstein, 1975). Increasingly complex clinical trials confirmed the beneficial effect in heart failure, and this entailed substituting or complementing the paradigm of the mechanism of heart failure in itself and its treatment – from a hemodynamic model to a neuro-hormonal model (Davies & Bashir, 1999).

However, to return to the case of hypertension – there were *clinical* reports in the late 70's that BBs do not work well for people of African or Caribbean origin (Saunders, 1988). This was confirmed by trials, the first being the so-called Veterans trial in 1982. There was no inkling of a mechanism, except for the very general fact that drugs can act differently on different ethnic groups. Hence, the fourth Joint National Committee in 1988 issued the recommendation to substitute diuretics for BBs (with CCBs being a supplement), as first line of treatment, in African-American and older patients. This recommendation was maintained even when later angiotensin-converting enzyme (ACE) inhibitors came in vogue in the mid 1980's, albeit for different reasons – one needed between 2 and 4 times the same dose of an ACE inhibitor to produce in an African-American patient the same effect as in Caucasian patients (Weir *et al.* 1998). The same change from beta-blockers to CBBs happened for elderly patients, sometime later, at the fifth Joint National Committee in 1996, with some studies suggesting that beta-blockers might even be detrimental in the long run (Messerly *et al.*, 1998, Khan & McAlister, 2006, Bangalore *et al.* 2007).

Of course, mechanistic explanations were subsequently discovered. On the one hand, the profile of elderly hypertension is that of a low cardiac output and high peripheral resistance. On the other hand, African-American patients tend to have low renin levels, higher sensitivity to sodium, reduced Na+/K+ ATPase activity, and expansion of plasma volume. More recently, the responsible genes were also identified, amongst which a different gene for the beta receptor (Johnson 2006).

However, my point here is that the history of this use of BBs (and of their substitution with CBBs) suggests that Clarke *et al.*'s hypertension example is *not* a case in which mechanisms contributed decisively to the problem of extrapolation by triggering differentiated treatments for specific groups. It suggests, on the contrary, that it was a case in which BBs, whose mechanism was fairly well known, encountered the problem of extrapolation, and it was in response to this problem

that the CCBs were substituted for them in the above-mentioned guidelines.

Indeed, it might perhaps be also worth mentioning that CCBs have had in their history problems of extrapolation as well, since it was discovered that short-acting CCBs increase the chances of cardiac infarction, especially for young patients with non-atherosclerotic lesions, due to paradoxical stimulation of arterial contraction. Yet another finding was that they increase the risk of death for patients, when administered post myocardial infarction or for unstable angina (Psaty *et al.* 1995, Furberg *et al.* 1995).

In the end, the sort of example used by Clarke *et al. s*eems to play directly into the hands of the criticism provided by Howick *et al.* It is an example in which BBs (and short-acting CBBs) acted erratically, or did not act at all, or acted insufficiently in certain circumstances, even if the mechanistic evidence at that time did not have anything specific to say about these circumstances. Of course, it has been known that specific age and ethnic groups might have clinically-significant particularities in their response to medical treatments, but this sort of general knowledge is tantamount to the knowledge (and worry) that problems of extrapolation might exist. And it is one thing to have the worry of extrapolation, and another to trigger measures such that it could be prevented or mitigated.

In the line adopted in Howick *et al.* 2013, one can always formulate the following meta-induction: if the previous set of drugs, based on mechanistic evidence, failed, then we have no assurance that the next set of anti-hypertensive drugs will always work, or do no harm. And, further on, if this is the case for anti-hypertensive drugs, then it could be the case for any other drugs or other medical treatments. I now turn to my positive suggestion, which is advanced in the broad framework of the initial RWT, but views mechanisms as difference-making and hence follows the revised RWT proposed in chapter 3.

§3 Mechanisms as difference making and a new look at extrapolation

Howick *et al.* seem to leave us in a dire situation as to the extrapolation problem, at least as far as mechanisms are concerned. As they warn at the end of their (2013) paper "A possibility that has been implied throughout this paper is that we have to learn to live with a much higher degree of uncertainty and scepticism about the effects of many medical interventions, even those whose effects have been established in well-controlled population studies" (p.288). But perhaps there is a better way of thinking about this problem. The philosopher of science should, of course, be extremely careful when advancing proposals to tackle the issue of extrapolation, and tragic examples such as that of Thalidomide (Howick *et al* 2013, p. 283) should always be borne in mind as a

reminder that serious dangers might be associated to medical treatments. But at the same time, there is something that does not feel exactly right about the arguments typified by Howick *et al* in their (2013).

For instance, in the example discussed above from Clarke *et al.* (2014) – which I tried to do justice as much as I could to Howick *et al.*'s line of argumentation - one cannot help but think that after all, the hypertensive treatment is one of the most detailed and efficient treatments in modern medicine. Similarly, the associated (anti-) hypertensive mechanisms are nowadays part of the most well known physiopathological schemes. Again, clinical medicine is no place for philosophers to brazenly throw in radical proposals and re-interpretations (and PhD students should be even more careful), but I think the view proposed in this thesis of mechanistic causation as both productive and difference-making could throw some light on the problem of extrapolation. The purpose here is not to 'solve' the problem of extrapolation (which is yet another subject that would require an entire thesis to develop) but to suggest that by paying attention to what happens in medical practice, and by adopting the correct view of mechanistic causation, we might be able to conceptualise more properly what the problem of extrapolation is and what its possible solutions might be.

As mentioned in chapter 3, most philosophers of science differentiate mechanistic causation from difference-making causation, maintaining that mechanisms act by production only (Williamson 2011; Illari and Russo, 2014; Craver and Tabery, 2015). Howick *et al.* are no exception, providing in the beginning of their paper a list of the *production* accounts of mechanisms current in the literature (Howick *et al.* 2013, p. 279). They later define 'mechanistic reasoning' as an inference about an *intervention* producing an effect – which inference, since mechanisms as understood in terms of production, just articulates the attempt to use production in order to ground, among others, extrapolation claims. "Regardless of how they are characterized, mechanisms must have some action if they are to be used to support claims that an intervention produces some effect. Following previous work, we define 'mechanistic reasoning' as an inference about an intervention's clinical effect from alleged knowledge of relevant mechanisms and how they relate to one another" (Howick *et al.* 2013, p. 279). Admittedly, there is a grain of difference-making in the way Howick *et al.* define "mechanistic reasoning", since they make use the language of an "intervention". This grain of difference-making is not developed at all, however. Nor are Howick *et al* discussing at all the very distinction between production and difference making, setting aside any foray into the specific functioning of mechanisms with the clause "regardless of how they are characterised".

On his part, Steel develops one of the classical approaches to mechanisms in terms of production (namely the so-called "MDC" approach), while aiming of course to look in much more detail at how production is enabled and preserved across contexts (Steel 2008, p.42).Just like

Howick, Steel uses, in addition to the MDC approach, the language of intervention, in order to explain how the process-tracing method can be applied to particular nodes of mechanistic production. However, the whole weight of his argumentation rests on the *process* side of his process-tracing method, and this grain of difference making is not developed. As for the RWT proponents, Clarke *et al.* adopt in their paper the non-committal definition of mechanisms provided in by Illari and Williamson (2012, p. 120) and discussed in chapter 3 "A mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon". It is a neutral and thin definition, that could be developed to make room explicitly for difference making, but as things stand, is rather geared towards covering only the variety of production approaches.

Could one use this construal of mechanisms as difference-making that I have put forward in chapter 3 in order to think of the problem of extrapolation from a slightly different perspective - while adopting the framework of RWT, and also the criteria of mechanistic evidence provided in Clarke *et al* (2014)? One of the crucial criteria advanced by the latter in order to organise evidence of mechanisms is the differentiation between *fragile* and *robust* mechanisms (p. 357). If we take this criterion, coupled with the insight into the dimension of difference-making of mechanisms, we arrive at the conception that a *robust* mechanism, as opposed to a *fragile* mechanism, is the one that can preserve its difference-making through various tests of stress and across a certain variation of contexts.

The natural semantics of the notion of 'mechanism' can only reinforce this conception of robustness. A part of the very meaning of the notion of mechanism is linked to the idea of a set of factors that are 'shielded' from variations in external factors (Cartwright, 1999, pp. 29, 50, 87-90). It is an idea we owe in large part to the seventeenth century 'mechanicist' philosophy, and it is certainly reflected by mechanisms in physics and chemistry. Obviously, the degree of 'shielding' of mechanisms in biology is lower than the one in physics and chemistry. However, my point is simply that, just by unilaterally insisting on our incomplete knowledge of medical mechanisms, on their potentially paradoxical effects and on their potential change of action following change in external conditions (as Howick *et al.* do), one should not lose sight of the fact that *we still have* the distinction between fragile and robust as concerns the mechanisms of biology and medicine, and an adequate way of conceptualising this distinction by way of maintaining or not the difference making.

Recall the discussion of pre-emption from chapter 3. This discussion is not only important in order to show that pre-emption is not a serious problem for ontic construal of mechanisms as both productive and difference-making. It is also important from the point of view of extrapolation, because taking into account phenomena such as pre-emption might be one of the keys to

understand why the problem of extrapolation shows up and what we need in order to mitigate it. One of the causes of the problem of extrapolation are precisely phenomena such as pre-emption or analogous events, in which other causes neutralize or modify the effect of the cause we are interested in.

In such cases, because the difference-making of mechanisms is no longer *actually* present, what we are actually left with, if the mechanism is still functioning at all, is production. The sort of infinitesimal differences that Howick *et al.* describe as being sufficient to make mechanisms behave erratically, presumably leave in place spatio-temporal contiguity, the transmission of signs and energy, or whatever other means one might employ in order to mark out the continuity of processes. It is conceivable that in the case of BBs discussed in §2, the marks of production were in place for people of African-American origin, for instance. But the difference-making specified by any of the counterfactuals from above was missing.

What we need in order to tackle the problem of extrapolation is a bit of modal force, as it were. We need a mechanism that produces indeed an effect, but following a true counterfactual that is actualized and that warrants that the effect *would* happen, were such and such conditions in place. As mentioned in the Introduction to this chapter, any extrapolating claim expresses itself in a counterfactual way, since it states that a certain effect (observed in the test context) would be produced in the target context. This is the sense in which, arguably, we need a *robust* mechanism – a mechanism whose difference-making is not influenced by pre-emption, neutralization, antidotes, etc.

How to check whether we are dealing with a fragile or a *robust* mechanism, i.e. with a mechanism whose difference making can reasonably be expected to manifest itself across varying causal contexts? Here, we can usefully rehearse the reasoning behind the revised form of RWT. One access to difference making is provided by the laboratory studies, including animal experimentation, and it is on the laboratory studies that theorists of extrapolation have placed the burden of solving the problem (leading ultimately to skepticism about a possible solution). But the revised form of RWT tells us that population studies are a different epistemic way of access into the difference-making of the same causal relations. If *both* mechanistic evidence and evidence of population studies concern difference-making and the same type of causal relations (although mechanistic evidence is assessed experimentally, whereas evidence of population studies is assessed statistically), then evidence coming from population studies could be used to tell us something about mechanisms themselves. One is thereby in position to extend RWT from the area of *establishing* causal claims, to the area of extrapolation. The extension of RWT would say that, when dealing with the untested, target contexts of interest for extrapolation, one *needs* evidence of difference making from both mechanistic studies and population-studies (as well as evidence of production

from mechanistic studies), which are thus both necessary, although not necessarily sufficient, to solve the problem of extrapolation.

In order to see how precisely the interplay between population studies and microstructural research offers us a way of access into the robustness of mechanisms, we should go back, with a more charitable reading, to the Clarke *et al.*(2014) case study.

§4 RWT, extrapolation and the Clarke *et al.* example

Recall now the case of the hypertension treatment with calcium-channel blockers (CCBs), put forward in Clarke *et al.* (2014) as an example of how, in their view, mechanisms help solve the problem of extrapolation. This example seems to offer us an important dimension of the interplay between mechanisms and population studies because, as Clarke *et al.* argue, it is primarily mechanisms that seem to contribute to the extrapolation claims formulated in the context of population studies. Clearly, we have evidence of CBBs coming from physiopathological laboratories, animal experimentation, etc. about (what appear to be) *robust* mechanisms, and this mechanistic evidence that is used in the context of population studies. In my own terms, we have evidence of difference-making obtained from a microstructural research that is used to ground and reinforce claims about difference making made at the level of population studies in order to extrapolate them.

My point in §2 was that this is not the whole story, and that the history of the hypertension treatment does not suggest that it was a case in which mechanisms solved the problem of extrapolation. It suggests, on the contrary, that it was a case in which the BBs, the previously employed treatment for hypertension, whose mechanism was fairly well known, encountered the problem of extrapolation, more precisely failed the test of extrapolation, and it was in response to this problem that the CCBs were substituted for them in the guidelines cited by Clarke *et al*; the subsequent population trials strengthened the claim as to the robustness of CBBs (eliminating, of course, the short-acting CBBs).

But this critical point of view, which is typified in Howick *et al* 2013, ignores the very basics of RWT and the *epistemic* interplay between mechanisms and population studies – an epistemic *interplay* that is made possible, *once* mechanistic causation is construed in terms of both production and difference making. The issue becomes clearer if we recall that in his (2011) critique of RWT, Howick repeats the same charge of mechanisms not being able to solve *alone* the problem of extrapolation (Howick 2011, p. 930 *et passim*), as I showed in chapter 3. Again, this means to ignore the gist of RWT, which presupposes the joint-use of both controlled population studies and mechanisms, all the more intelligible in a monistic (and not pluralistic) framework of causation.

Let me give you a final illustration of this critical approach, which tends to see in isolation

the possible means of tackling the problems of extrapolation. In the programmatic part of their (2013), Howick *et al.* list five possible solutions to the problem of extrapolation that are to be examined: simple induction, n-of-1 trials (in which a single patient randomly receives the experimental treatment or the control), pragmatic randomised trials, clinical expertise, and mechanistic knowledge. Each of these possible solutions are then taken in turn and shown to be problematic, and I have laid down in section 1 Howick *et al.*'s argumentation about mechanisms. But what is extremely interesting for my purposes here is that at no point do Howick *et al* suggest that this set of possible solutions (or a subset of them) all with their proper weaknesses, admittedly, could be used *together*, reinforcing each other, complementing each other's weaknesses, etc. (Howick *et al.* 2013, p. 278). One seems to be in the search for the 'magic wand'. A panacea does not exist though. What we can try to do is to find the adequate way of conceptualizing where the problem lies, and to use all the means at our disposal in trying to alleviate the problem (crucially, including difference-making evidence coming from both mechanisms and population studies), always keeping an eye to what happens in clinical *practice*.

In the end, to come back to the example of Clarke *et al.,* when looking at the guidelines which recommends long acting CCBs as a first line treatment for elderly patients and patients of African-American origin, what we see is not mechanisms alone solving the problem of extrapolation. We see rather the *collaboration* between the population studies and the mechanistic research - a collaboration which results in the tacit definition of CCBs as entailing a *robust* mechanism in the first line treatment of hypertension, in contrast to BBs, which are tacitly defined as entailing a *fragile* mechanism in the first line treatment of hypertension.

And we also see that happened *in practice* was that the difference-making of BBs, although apparently pretty constant and well-defined on the level of microstructural research in laboratories, manifested erratically in real-life situations involving differing biological contexts. One learnt about it following clinical reports, and this erratic manifestation (of the laboratory established difference making) was confirmed by RCTs. On the other hand, the difference making of CBBs, also well-defined on the level of microstructural research, was confirmed by RCTs as manifesting itself in the same way outside of laboratories (i.e. on the population level), and hence the mechanism of lowering hypertension associated with them was deemed *robust*.

This approach to robustness and extrapolation is consistent with the criteria for the quality of evidence laid out by Clarke *et al.* and also coheres with their global picture of the evidential interaction between mechanisms and populations (Clarke *et al.* 2014, p. 355). If we added to the *to-and-fro* between mechanisms and populations of RWT the insight that the difference-making should be seen as issuing also from mechanisms, we would be better equipped conceptually to approach the

problem of extrapolation.

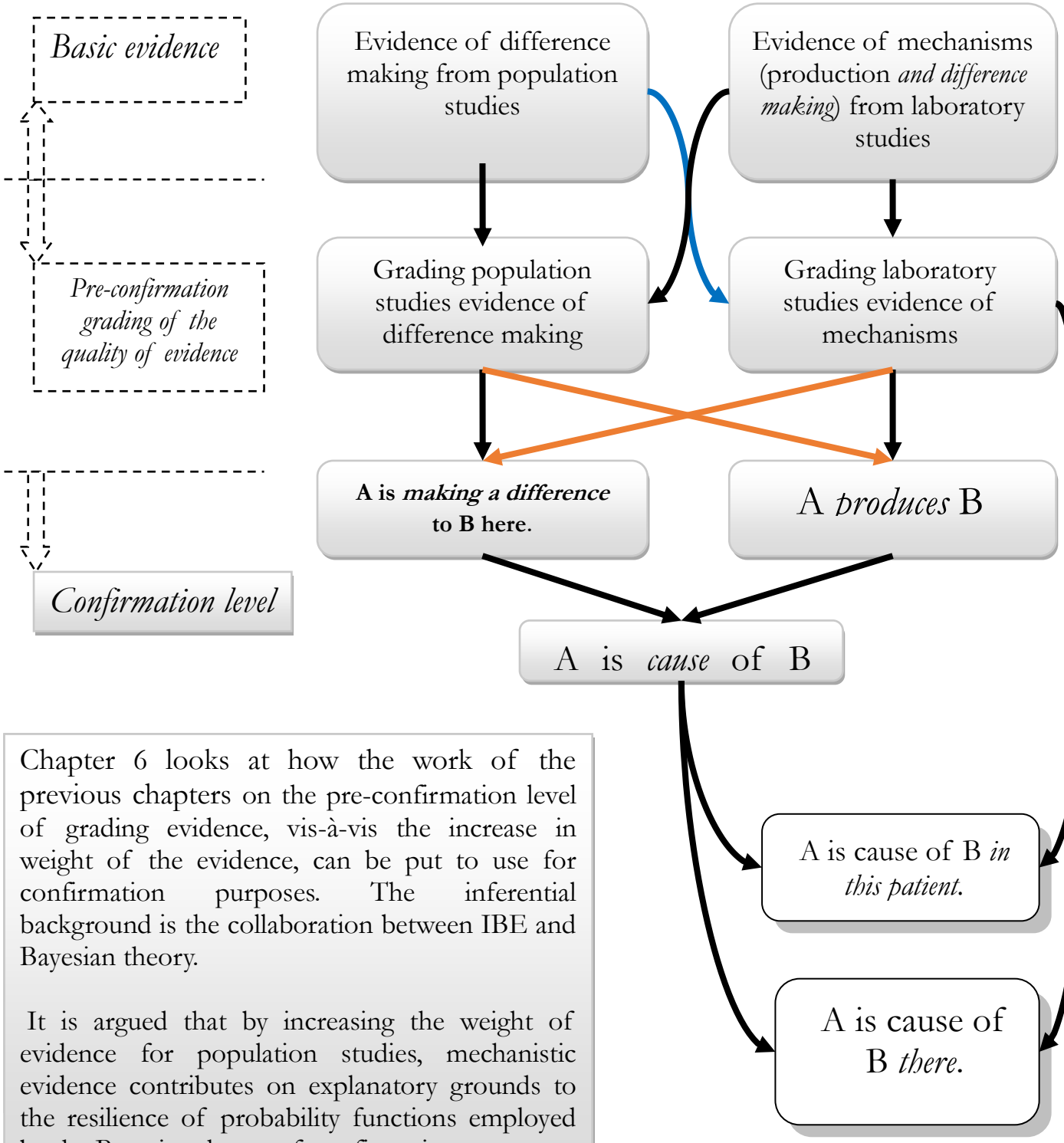
Furthermore, using such a conceptual approach, I suggest, one could take full advantage both of the emerging and very promising literature on applying Bayes nets to microstructural mechanisms (Clarke *et al.* 2014b, Casini *et al.* 2011) – Bayesian nets which are thereby measuring the difference-making on the level of mechanisms – and the global approach of systems medicine (Williamson, *forthcoming*). The perspective here is that the Bayes nets applied to mechanisms and/or the global results of systems medicine could provide a numerical expression of the *robustness* of mechanisms and accordingly of their potential to overcome extrapolation worries.

Conclusion chapter 5

In this chapter, I have argued in favour of the epistemic advantage *V)* of the revised RWT, namely that joint evidence of population studies and of mechanisms as difference-making can be helpful to face the challenges of the problem of extrapolation. This joint evidence can help us circumscribe *robust* mechanisms – where a robust mechanism is a mechanism which is likely to preserve its difference-making across a certain range of contexts outside the testing area.

The emerging research on systems medicine and on the application of Bayes nets to mechanisms – which could be easily integrated into the framework of the revised RWT – could offer a fruitful development of this approach to the issue of extrapolation.

Chapter 6



Chapter 6 looks at how the work of the previous chapters on the pre-confirmation level of grading evidence, vis-à-vis the increase in weight of the evidence, can be put to use for confirmation purposes. The inferential background is the collaboration between IBE and Bayesian theory.

It is argued that by increasing the weight of evidence for population studies, mechanistic evidence contributes on explanatory grounds to the resilience of probability functions employed by the Bayesian theory of confirmation.

Chapter 6 Medical Mechanisms and the Resilience of Probabilities

Introduction

The present chapter aims to move the discussion of the evidential relevance of mechanistic evidence and of the revised RWT from the pre-confirmation stage (with which it has been concerned in the previous chapters) to the confirmation stage. As noted in chapters 1 and 4, Lipton construed IBE as mainly a *guide* to scientific inference, given that, by being concerned mostly with the qualitative aspects of evidence and lacking a quantitative expression, it could not be sufficiently fine grained to let us adjudicate in complicated cases of confirmation (as opposed to the simpler cases in which, due to the explanatorily rich nature of evidence, IBE could eliminate all alternative hypotheses and pick out the right one, leading as it were to the 'only explanation', as was shown in chapter 1).

One possible extension to confirmation for IBE would be a form of association with the Bayesian theory, which in principle, would usefully compensate for its lack of quantitative expression, and Lipton himself hinted at this possibility, maintaining that IBE and Bayesianism are compatible, and rather than rivals, they should be 'friendly companions' (Lipton 2004, pp. 107-117). But questions still remain to be answered. As shown in chapter 1, in principle, IBE should confirm hypotheses that best explain the evidence (on the basis of explanatory virtues like simplicity, scope, theoretical unity, and individualization, plus Mill's principles, if one adopts Lipton's particular interpretation). In return, the Bayesian theory takes hypotheses as being more confirmed to the degree to which their probability is raised by conditionalising on evidence using Bayes' theorem (or by employing various measures of confirmation that revolve around this increase in probability). How precisely could they be used together?

This question has sparked heated debates in the literature (e.g. Okasha 2000, Iranzo 2008, Romeijn 2013, McCain and Poston, 2014), and the question's relevance extends beyond the confines of a strictly theoretical discussion, well into our particular area of concern, i.e. medicine. This is evident if we take into account that on the one hand, Bayesianism is consistent with EBM (Ashby and Smith, 2000) and, given its tremendous popularity in the general philosophy of science, it has been increasingly used to interpret the results of controlled population studies. One the other hand, IBE is (or should be) a natural ally for RWT proponents who seek to emphasize the importance of mechanisms. The reason is that a traditional role that mechanisms play is precisely that of explaining (Williamson, 2011). Moreover, given the work done in chapter 2 on how the Clarke *et al* criteria of

mechanistic evidence can be shown to function and be justified on explanatory grounds, a natural question is whether these criteria (which originally have a *pre*-confirmation role, as 'rules of thumb' for assessing the *quality* of mechanistic evidence) could not play some role in the confirmation as such of medical causal relations.

This is then the context due to which the question of the collaboration between Bayesianism and IBE is worth investigating in the present thesis. The Bayesian theory of confirmation is and should still be the first choice for controlled population studies in the special sciences. But if mechanisms are taken to explain (as they should), and if they are taken as both productive and difference-making (as they should be) - being involved in the *same* causal relations as those reflected epistemically by the dependencies of population studies - then the revised RWT proposed in this thesis requires one to look at the explanatory features that could be integrated into the Bayesian theory of confirmation. More specifically targeted, the question is how to integrate the explanatory features, which are expressed by the Clarke *et al.* 2014 criteria of mechanistic evidence, in the Bayesian framework of confirmation.

I will not insist here on the point that I have already discussed in the previous chapters – the evidential interplay between mechanistic evidence and evidence from laboratory studies is only possible in the framework of ontic causal monism, and accordingly, only in the framework of mechanisms as difference making. Otherwise, one is bound to conclude that laboratory, mechanistic studies, and population studies, draw evidence of *different* causal relations, which not only make impossible evidential aggregation from different sources, but also, as far as the present discussion is concerned, would make impossible the collaboration between IBE and Bayesian theory (in a way useful for RWT, at least).

Now, it should be admitted that the issue of the collaboration between IBE and Bayesianism is controversial in the philosophical literature. And the present thesis has gone already far into controversial issues. Therefore, my discussion of theory confirmation in these two last chapters will assume a prudent development. In effect, the present chapter try to be as non-committal as possible when it comes to possible interpretations of the compatibility and 'friendly companionship' between IBE and Bayesianism, leaving the discussion of the more contentious aspects for the next, final chapter.

The description in §1 of the state of the art of this debate in general philosophy of science will show that Lipton's initial proposals as to how the 'friendly companionship' should look like (proposals also developed by various other authors), namely that explanatory elements should play a role in the assignment of priors and likelihoods,[75] bear with them a bone of contention. Contention

[75]See Lipton 2004, pp. 107-117, but also the exchange of articles and replies in Lipton (2001) and Salmon (2001), and

and disagreement do not mean of course that such issues should not be further pursued and I will do so in the next chapter. In the present chapter though, while noting the initial plausibility of using IBE to reinforce Bayesian confirmation in the way suggested by Lipton and like-minded theorists - in the sense in which the Clarke *et al.* criteria of mechanistic evidence could be employed in order to gear the abovementioned assignment of priors and/or likelihoods – it will be advanced a less contentious proposal on how to construe the 'friendly companionship'.

More precisely, the proposal put forward in §2 is based on Brian Skyrms' initial discussion of the 'resilience' of probabilities (Skyrms, 1977), and on an analogical discussion of the 'stabilising' role that mechanisms play over causal relations, on an ontic level, inspired by Russo and Williamson's account of epistemic causation. In brief, the proposal says that mechanistic evidence can increase the 'resilience' of Bayesian probabilities, where 'resilience' means the stability of one's credences in the face of *new* evidence.

This proposal coheres with and strengthens recent arguments advanced in McCain and Poston (2014, and *forthcoming*) as to how explanatory information is evidentially relevant, and the first part of McCain and Poston's discussion will be presented in §3. In the thread of the present thesis, this proposal takes up and continues the epistemic advantage *III)* of the revised RWT, as outlined in the previous chapters - stating that mechanistic evidence could increase the weight of population studies evidence.

Going back in §4 to the medical aspect of the discussion, the developed suggestion will be that, when dealing with the results of controlled population studies in medicine, the stability of one's credences in the face of new evidence should be greater for the population studies results, where we also have attending evidence of a mechanism for the causal relation(s) in question - an increasingly greater stability to the degree to which the criteria of mechanistic evidence of Clarke *et al.* are satisfied. This, in effect, is precisely what was designated in the previous chapters as the epistemic advantage *VI)* of the revised RWT.

The next chapter, as mentioned, will discuss the more contentious aspects of the 'friendly companionship' between IBE and Bayesianism. Taking its cue from the second part of McCain and Poston's discussion of evidential relevance, it will suggest that, after all, mechanistic evidence could play a role in the assignment of priors and likelihoods, if we adopt a strong interpretation of the objective status of explanatory values.

§1 State of the art - the relationship and compatibility between IBE and Bayesianism

Any presentation of the possible compatibility between IBE and Bayesianism and of the ways

---

further discussion in Iranzo (2008) and Weisberg (2009).

of articulating it, should start, surprisingly, with van Fraassen's initial diagnosis that the two are actually *incompatible* (van Fraassen, 1989, p. 169). The reason for starting in this way is that all the subsequent discussions of compatibility - including McCain and Poston's discussion on evidential relevance, the first part of which will be outlined in more detail in §3 - attempt, in one way or another, to respond to van Fraassen's charges. One such charge, for instance, has been that of the 'bad lot' (van Fraassen, 1989, p. 143), holding that, all our attempts to find the best explanation notwithstanding, the right explanation might be outside of the pool of our available hypotheses. It was in response to this charge that Lipton rolled back to interpreting IBE as inferring the *probable* truth of the best explanation (Lipton, 2004, p. 58), and this cautious move has also been adopted by other IBE theorists (e.g. Niiniluoto, 1999).

As far as the theme of the present chapter is concerned, it is another of van Fraassen's lines of attack against IBE which stands out: explanatory considerations can play no role in the context of Bayesian conditionalisation, on pain of drawing unwarranted inferences. If the posterior probabilities of more explanatory hypotheses were given 'bonus points' post-conditionalisation, due to their superior explanatory content, this would violate the demands of Bayesian rationality and lead to Dutch-Book results - a set of bets guaranteed to make one lose money in all scenarios (Van Fraassen, 1989, p. 169).

An interesting answer to this irrationality charge came from Okasha (2000), followed by Lipton (2001) and Lipton (2004).[76] Lipton and Okasha argued that associating explanatory elements to Bayesian reasoning, in the sense of taking these explanatory features into account *after* the posterior probabilities have been calculated, is actually putting the cart before the horse (Okasha, 2000, pp. 702-704, Lipton, 2004, pp.106-117). The right way to construe the 'friendly companionship' need to be tighter than merely taking into account the explanatory features of hypotheses *after* conditionalisation. These explanatory features should play a part in the conditionalisation itself, in the sense that explanatory considerations should influence the assignment of priors and/or likelihoods.

We have seen in the previous chapters that, in the framework of IBE, the explanatory value of hypotheses is standardly cashed out using several virtues as simplicity, scope, theoretical unity and individualization, entailing, respectively, that a more explanatory hypothesis is simpler, covers a greater part of the evidence, converges with previous knowledge, and provides a mechanism or a fine-grained description of the causal enchainment(s) in question, respectively. Now, according to

[76]Proponents of the view that Bayesian conditionalisation should include explanatory elements also include Iranzo (2008), Weisberg (2009) and Romeijn (2013). In order to keep this brief this introductory section on the state of the art, I will refer in the main text mainly to Lipton and Okasha, and I will note in footnotes the relevant aspects adduced to the main discussion by the just-mentioned authors.

Okasha's proposal, if one or more of these features or virtues sets a hypothesis higher than its competitors, then, *before* applying Bayes theorem to obtain the posterior probability (p(h/e)=p(h)p(e/h)/p(e)), one could assign to the more explanatory hypothesis either a higher prior (p(h)) (say, if it converges more to previous knowledge than its rivals) and/or likelihood (say, if it has a greater scope).

Here is one of Okasha's examples (an example with a medical context, as it happens). Suppose a mother brings her child to a doctor with leg pain complaints. The doctor has at his disposal two hypotheses: the first is that a muscle was strained, the second is that ligaments are torn. After examining the child, he opts for the second hypothesis, because for children of that age straining of muscles is very rare (background knowledge), and because the hypothesis that ligaments are torn best explains the symptoms he notices after examining the child (greater scope and individualisation). In probabilistic terms, this translates to the second hypothesis having a greater prior probability than the first, and to the probability of the evidence conditional on the second hypothesis (the likelihood) being higher than the probability of the evidence conditional on the first hypothesis. After conditionalisation, the posterior probability of the second hypothesis will be rendered higher, with no apparent violation of the Bayesian rationality.

Admittedly, not *all* hypotheses with a high prior and likelihood are also highly explanatory. But, in turn, all highly explanatory theories should have either a higher prior or likelihood, or both (Okasha, 2000, p. 705). This strategy applies with a vengeance to cases in which scientists come up with radical new hypotheses, because the Bayesian's usual strategy of retorting that 'today's priors are yesterday's posteriors' does not apply. In such cases, the 'Bayesian model is silent', on Okasha's stern diagnosis (Okasha, 2000, p. 709 *et passim*).[77] A less drastic observation would be to say that, in the context of new hypotheses being devised, the Bayesian has problems in assigning prior probabilities and likelihoods, and the 'friendly companionship' should hold in these cases also, with explanationist features helping out to circumscribe the probabilities in question (Lipton, 2004, pp. 115-119).

This applies for various other situations in which, although the hypotheses in question are not new, we still have incomplete data or evidence; and various theorists have followed up this path, suggesting that explanatory constraints should be employed to make the transition from subjective to (some sort of) objective Bayesianism and Bayesian interpretation of probabilities. Weisberg (2009, p. 141) has argued that explanatory considerations should be used to constrain *a priori* probabilities, along with the *Principle of Indifference* and the *Principal Principle* (in cases in which the background knowledge, and the available evidence, are such that, even if the *Principle of Indifference*

[77] Okasha views such cases of paradigm shift or inventing of new hypotheses as cases in which IBE can apply alone, leaving aside the Bayesianism.

applies, we do not get a unique value of the posterior probabilities of hypotheses).[78]

Another brand of objective Bayesianism which incorporates explanatory elements is the one proposed by Jon Williamson. Of the three main requirements of his objective Bayesianism (probability, calibration and equivocation), upholding the norm of equivocation (which asks that one's degrees of belief should equivocate as far as possible between the elementary outcomes; Russo and Williamson, 2007, p. 168), amounts to introducing a classical explanatory constraint, namely simplicity and parsimony. This can be seen by looking at the corresponding account of epistemic causation developed by Williamson. On this account, the norm of equivocation asks that a causal graph be as non-committal as possible about what causes what 'A causal graph C is maximally non-committal, from all those in E, if there is no other causal graph D in E which makes fewer causal claims (including both arrows and gaps) than C' (Wilde & Williamson, 2016). Williamson's objective Bayesianism leaves enough room to introduce other explanatory elements, and we will shortly come back to his insightful relationship between causation and Bayesianism.

Now, problems for this approach show up because, in certain cases, it might be wrong to use the explanatory virtues to constrain the priors and/or the likelihoods. Here is one example put forward in Weisberg (2009). Suppose you find one day your house in disarray, with the lock broken and valuable belongings missing. The first hypothesis that comes to mind is that a burglar broke in, stole the valuables and while searching for them, caused all the mess in the house. But there is another, improbable hypothesis, and it happens that it is the true one. One burglar did force the lock and entered the house, but happened to encounter in that very moment another burglar who had penetrated through the window. They started to fight, turning everything in the house upside down, until a police officer came in. They both took off, and it is the policeman who decided to take advantage of the situation and take the valuables, planning to cast the blame on the two burglars (Weisberg, 2009, pp. 129-130). Clearly, the first hypothesis is best explanatory, in terms of simplicity, theoretical unity, etc.; and yet it is the wrong one. Why pay any attention after all to explanatory values at the level of confirmation?

Examples such as Weisberg's seem to be devastating. On the other hand, it should be noted that the mere fact that the hypothesis that scores higher in terms of the explanatory virtues might in certain cases be false, does not directly show that those virtues should not be used to constrain the priors and/or likelihoods. It simply shows that the explanatory virtues can be misleading. Compare the analogous reasoning - there can be cases where your total evidence supports the hypothesis that is false as a matter of facts, but it does not follow in such cases that you should not use your total evidence. Thus, based on particular cases, one should not deny the use that explanatory values

[78] An analogous move of shifting away from subjective Bayesianism, in order to integrate IBE, is made in Iranzo (2008)

might have in the assignment of priors and/or likelihoods.

Moreover, Weisberg's example described above exploits one of the most vulnerable of the explanatory virtues, namely simplicity, and it is a vulnerability recognized by IBE theorists. Lipton has dubbed the corresponding objection 'Voltaire's objection', in connection to Voltaire's diatribe against the best of the possible worlds theory of Leibniz (Lipton, 2004, pp. 144-147). Why should the world (and our true hypotheses about its causal structure) be simple? But simplicity is just one of the explanatory virtues. When our explanations are based on mechanisms (which embody simultaneously various explanatory virtues, since mechanisms-based explanations are arguably more theoretically unified, have greater scope and greater individualisation), then it is much harder to reject as *ad-hoc* (or as stemming from some sort of idealist perspective) the inferences to the best explanation that favour the hypotheses backed up by mechanisms over the hypotheses are not thus backed up. Hence, moving on to the friendly companionship between IBE and Bayesianism, to say that, in the realm of medicine, one should increase the likelihood or priors for these hypotheses which also provide a mechanism, is not susceptible *prima facie* to Voltaire's objection and is not betraying some sort of straightforward, simpleminded use of the explanatory virtues of IBE.

However, one has to admit that there is still a bone of contention, and I leave the discussion of this use of explanatory values in confirmation for the next chapter. In the present chapter I will focus on what appears to be a less controversial aspect of the Bayesianism/IBE possible collaboration, which I will introduce in the next section.

§2 Resilience - of mechanisms and probabilities

Using mechanisms in order to control the priors or likelihoods is not an uncontroversial move to make for the theorist of medicine. Still, mechanisms (and the criteria for mechanistic evidence) should have some role in the confirmation of causal claims in medicine. But what would this role be? In §1, I mentioned Williamson's objective Bayesianism, as an example of how one could attempt to strengthen the subjective basis of the Bayesian theory with norms that could be derived from (or resemble) explanatory considerations. But Williamson's account is a source of inspiration at this point from yet another point of view, precisely because, as we have discussed from a different perspective in chapter 3, his account of objective Bayesianism is coupled with an account of *causation* (in which the requirements of objective Bayesianism have corresponding requirements for circumscribing causal relations).[79] The lesson to be drawn immediately is that the to-and-fro

[79] I am using here Williamson's account (of objective Bayesianism and of epistemic causation) due to its heuristic richness. This is not to say that the respective account is not without its problems; see for instance [*reference suppressed for blind review*]

between a theory of subjective credences/objective probabilities and a theory of causation can be illuminating by allowing us, via various analogies, to circumscribe, or rather suggest, characteristics of credences/probabilities that parallel ontic characteristics of causal relations. In other words, such a to-and-fro between the epistemic and ontic levels could be useful heuristically in order to suggest features that our theories of confirmation should possess, in relation to the ontic structure of reality that science uncovers (or, in our particular discussion, in relation to the causal structure that we know to be in place ontically in biology and medicine).[80]

Now, what is the main characteristic of mechanisms in medicine and biology as far as the causal relations are concerned? Perhaps a univocal answer to his question cannot be provided, but at least it would be safe to say that one of the main characteristics of mechanisms is that of 'stabilising' causal relationships. The idea is present in the founding (2007) paper by Williamson and Russo (in which the basis for the equal treatment of mechanisms and population studies was put forward). Moreover, the literature on biological mechanisms abounds in references to the 'resilience' of mechanisms (Bechtel and Abrahamsen 2005; 2011, Folke 2006)[81] – where their 'resilience' means that they can resist and adapt to variations in inputs and parameters, both internal to the organisms, and external, through different paths of feedback and self-maintenance, such that the causal relations that mechanisms produce are conserved.

Moving on from this ontic level, it just so happens that we have in the literature a related notion on the epistemic side, namely the notion of the 'resilience' of probability, developed by Brian Skyrms (1977; 1983). Very briefly put, Skyrms arrived at this notion in an attempt to characterize the probabilities that figure in statistical laws. More precisely, the attempt was to provide a notion of law-like, objective chance (or 'propensity' in his own terms) to be distinguished from the mere subjective credence, by pinning down the limited variation of such objective chances in law-like statements or propositions, relative to any other statements, i.e. by pinning down the resilience of the respective probabilities when conditionalised on other statements. The resilience of a proposition q having a probability a, was defined as 1 minus the greatest difference between a and the probability of q, conditional on any truth function of other propositions which is consistent both with q and its negation (Skyrms, 1977, p. 405, Mellor, 1983, p. 101).

For our purposes, it matters less how exactly Skyrms defined numerically the resilience of

[80]The epistemic/ontic distinction, as heuristically drawn above, in order to map the probabilities/causation differentiation (and simplify the presentation), is in order with the classical Bayesianism approach to *subjective* probabilities. Of course, on certain other interpretations of probabilities (including Williamson's brand of objective Bayesianism) one would also want to assign to probabilities a higher rank than merely the 'epistemic' category. But again, the epistemic/ontic distinction is drawn above heuristically, and the reader should be able to follow the role that this distinction plays in presenting the above proposal.

[81] The natural semantics of the notion of 'mechanism' can only reinforce this conception. A part of the very meaning of the notion of mechanism is linked to the idea of a set of factors that are 'shielded' from variations in external factors; *cf.* for instance Cartwright, 1999, pp. 29, 50, 87-90.

probabilities. What matters, however, is to follow the close parallel that links analogically the epistemic characteristic of the resilience of probability functions conditional on other evidential statements (which, in the meantime, has become a rather uncontroversial characteristic of the dynamics of credences for Bayesian theorists) with the ontic characteristic of mechanisms, capable of conserving the causal relations they produce, in spite of variations in the internal and external environment.

Given this parallel, the following proposal can be suggested, bearing also in mind that the explanatory function of mechanisms in the framework of IBE is based on the causal relations that mechanisms produce. The suggestion is that the explanatory considerations that could be integrated by a Bayesian theory of confirmation, refer to the evidential contribution of mechanisms to the resilience of Bayesian credences. More precisely, on this proposal, mechanistic evidence can increase the resilience of Bayesian probabilities, where 'resilience' means the stability of one's credences in the face of new evidence.

It is worth noting that, beyond the immediate purposes for which Skyrms used the notion of resilience initially, this notion was subsequently applied by Joyce in order to distinguish between the weight and the balance of evidence in a Bayesian framework (Joyce, 2005), by Leitgeb in order to differentiate between the quantitative and the *qualitative* aspect of evidence and belief (Leitgeb 2014; 2017), and by McCain and Poston, in order to defend the evidential relevance of explanatory considerations (McCain and Poston, 2014).

Interestingly, Joyce, Leitgeb, and McCain and Poston never linked the resilience of probabilities or credences to the evidence coming from mechanisms, in spite of the evident epistemic/ontic analogy mentioned above. However, McCain and Poston's discussion of evidential relevance (2014 and *forthcoming*) - which draws, among others, on Skyrms (1983), and Joyce (2005), and whose first part I will briefly present and discuss in the next section - will be very useful to situate better within the literature (and the current debates) the proposal as to the resilience brought about by evidence of mechanisms. In turn, §4 will go back to the medical context and will flesh out the consequences of the proposal for the medical interplay between mechanisms and medical controlled population studies.

§3 McCain and Poston's discussion of evidential relevance

The state of the art discussion in §1 began with van Fraassen's critique of IBE, as providing the reference point for all subsequent discussion on the collaboration between IBE and Bayesianism. This applies also to McCain and Poston's discussion, because they, too, start from a critique advanced in the vein of Van Fraassen, namely Roche and Sober's charge that explanatory

considerations are evidentially irrelevant (Roche and Sober, 2013, p. 659). As a matter of fact, Roche and Sober's critique of IBE is just van Fraassen's argument on the irrationality of adding 'bonus points' in disguise.

Roche and Sober's aim is to show that explanatory properties do not influence probabilities at all. In order to show that they cannot and should not influence probabilities, they formulate the issue in Bayesian terms. More precisely, they take a hypothesis H, the observations relevant to it O, and formulate the proposition E: if H and O were true, H would explain O. In their terms therefore, E encompasses the explanatory relation between the hypothesis and the observed data. They ask subsequently, as the Bayesian confirmation theory asks, whether Pr(H/E&O) > Pr(H/O) - in other words, whether the explanatory features add anything to the confirmation of H. Their answer is negative - nothing is added, because E is screened-off by O, or in formal terms Pr(H/O&E) = Pr(H/O); 'the explanatoriness of H is evidentially idle, once the truth of O is taken into account' (Roche and Sober, 2013, p. 660). Notably, the target of Roche and Sober's screen-off thesis does *not* concern *prior* probabilities, but posterior probabilities. The point is that learning E (i.e., that H would explain O if H and O were true) *after* having already learned O should not lead one to change your credence in H, which was precisely the crux of van Fraassen's irrationality argument.

Roche and Sober's example is, again, a medical one, concerning smoking and lung cancer. They take two scenarios, one in which smoking causes cancer, and another in which, according to Fisher's ancient supposition, smoking and cancer are both due to a common cause. In the first scenario, E is in place, because H explains O. In the second scenario E does not hold, as H does not explain O since both H and O (or the fact corresponding to the respective propositions, as one might want to read them) are the result of the common cause.

Now, on the level of strictly observed correlations, we will have, say: Pr(S smoked at least 10,000 cigarettes before age 50 | S got lung cancer after age 50) = c. And Roche and Sober's point is that c is not modified in any way if E is added: Pr(S smoked at least 10,000 cigarettes before age 50 | S got lung cancer after age 50 &*if S smoked at least 10,000 cigarettes before age 50 and S got lung cancer subsequently, then the smoking would explain the lung cancer*) = Pr(S smoked at least 10,000 cigarettes before age 50 | S got lung cancer after age 50). The explanationist story does not change the probabilities, so it is evidentially irrelevant.

Going back to McCain and Poston's reply, their main point is that, even if the screening off holds, the explanation features are still evidentially relevant. In other words, they concede as much as they can to Roche and Sober as to their screening off claim (although, as we have seen in the previous section, there is space of discussion as to whether explanatory features can influence the

priors of the hypotheses in question). However, McCain and Poston contend, there is a sense in which the explanatory story does have an influence on confirmation or prediction, without modifying as such the probabilities. What they influence is the *resilience* of the Pr function, by making certain of the probabilities more stable, or less volatile, given new evidence, or rather new information (and we shall see shortly why the distinction between evidence and information might be useful).

Again, the rationality of the Bayesian conditionalisation is not violated. Consider once more the probabilities of Roche and Sober, in a simplified form. We first have: Pr (S will get lung cancer/S has smoked i cigarettes to date) > Pr (S* will get lung cancer/S* has smoked j cigarettes to date), for all i > j. On a Bayesian analysis of confirmation this inequality implies that S being a heavier smoker than S* is evidence that S has a greater chance of getting lung cancer than S* (or, strictly speaking, that the hypothesis of S getting cancer is more probable than the hypothesis of S* getting cancer). At this point, one can assign different measures of confirmation, and it is commonplace that there is disagreement among Bayesians as to how exactly to frame this measure of confirmation (see for instance Glass 2012). McCain and Poston leave this point aside also, and postulate this measure of confirmation be some *d*.

Now consider the probabilities to which the explanatory element has been addedPr (S will get lung cancer/S has smoked i cigarettes to date *& smoking causes cancer*) > Pr (S will get lung cancer/S has smoked j cigarettes to date *& smoking causes cancer*), for all i > j.McCain and Poston's claim is that, even if the degree of confirmation *d* has not changed by integrating the explanatory story, the frequency data correlating smoking and lung cancer is 'reinforced', without being changed. It is not some paradox or a mere play of words, and in order to clarify what this means, McCain and Poston appeal to the useful distinction between balance and weight of evidence, where 'balance' is related to the strictly quantitative, numerical aspect of evidence, whereas 'weight' is related to the *qualitative* aspects of evidence.[82]

Simply put, in McCain and Poston's rationale, the explanatory side, while leaving the balance untouched (i.e. the degree of confirmation), adds to the *weight* of evidence – a notion which was at the center of our discussion of the evidential interplay between mechanisms and population studies in chapter 4. McCain and Post introduce the notion of weight as follows. Suppose one has a two-sided coin, which looks fairly typical, and one assigns a value of 0.5 to the next flip resulting in heads. Suppose next that one flips the coin a million times and it lands heads approximately one-half of the time. What is now the chance, ask McCain and Poston, that the next flip is heads? The

[82]McCain and Poston draw on Joyce (2005) in order to outline the distinction between balance and weight. As mentioned above, Joyce was one of the authors to put to use Skyrms' notion of resilience; useful discussions of this distinction can also be found in Kelly (2005) and Kelly (2008).

answer is the same, 0.5. Pr_t(heads on next flip/k) = 0.5 and Pr_{t+e}(heads on next flip/k& e) = 0.5, where k is one's background evidence, e is that there have been roughly 1/2 heads among a million flips, Pr_t is one's initial probability assignment and Pr_{t+e} is the probability assignment after learning only e.

The *balance* of the probability has remained unchanged; strictly speaking, the probabilities are the same, and the million flips are 'screened off'. However, e has added to the *weight* of evidence, in the sense that the one has learned that the two-sided coin is unbiased.[83] In other words, the strictly numerical, quantitative aspect reflected by the balance has remained constant; but the *qualitative* aspect reflected in the weight of evidence has changed. As a consequence $Pr_{t+e}$is more resilient in the way it changes in response to new *information*. Suppose that after the one million flips, the subject gets next an improbable sequence of five heads in a row (I am changing McCain and Poston's example slightly, to make it more similar to the next example they use, which I will present shortly). Due to the resilience of the probability $Pr_{t+e}$the credence of the subject that the next is heads should remain 0.5. Citing Joyce's claim that '[t]he weight of this evidence is reflected in the tendency for credences *to stably concentrate on a small set of hypotheses about the proposition's objective chance*" (Joyce 2005, p. 176, italics added), the two authorscontend that that the resilience brought in by explanatory stories works in the same way, by *fixing* the data about the relevant objective chances.

> "Explanatory information fixes the data about the relevant objective chances. It does not change what the data says the relevant objective chances are. Thus, if one has frequency data about the relevant objective chances, that is a feature of one's evidence that indicates what the objective chance is (obviously). But, once one acquires an explanatory-cum-casual story *that indicates that the objective chances are the same as the frequency data shows, the weight of the evidence is significantly increased even though in both cases the direction of the evidence is the same.* Adding the explanatory story significantly changes the probabilistic role of the Pr-function when updating on future information." (McCain and Poston, 2014, p. 6, italics added)

A two-sided coin being flipped one million times does not really amount to an explanatory story (although in an enlarged sense of explanation, it could be reckoned with this way, if it is taken to explain that the coin is not biased and that is why the chance of heads remains 0.5). Hence, McCain and Poston provide another example, in which the explanatory story is neatly delineated, and which is all the more useful because it will help us see the transition to mechanisms and the quality of evidence.

---

[83]The example of coins with the balance remaining unchanged after innumerable flips, in a different formulation (a normal coin vs. a magical coin) is also used by Romeijn (2013), in order to make the point that bare probabilities need supplementation in order to account for the (bare) observations and evidence. Romeijn does not go however on the path of the weight vs. balance distinction, but understands the need for *qualitative* aspects of confirmation as the need of 'theoretical notions', which are then connected to the explanatory features of the hypotheses at hand, ending up with the solution of modifying the priors, which I have discussed above. Interestingly, Romeijn also brings in (a certain construal of) mechanisms, in the sense of the underlying features of different chance-set ups, determining different outcomes, as indeed it happens in the simplest case with a magic coin and a normal one (Romeijn, 2013, pp. 435-437). But mechanisms as such are not explicitly invoked by Romeijn.

In the new example, we have Sally and Tom being informed that there are 1,000 x-spheres in an opaque urn. Sally, but not Tom, knows that blue and red x-spheres must be stored in exactly equal numbers because the atomic structure of the x-spheres is such that if there are more or less blue x-spheres than red, the atoms of all of the x-spheres will spontaneously decay resulting in an enormous explosion. Now, a random sampling of ten x-spheres without replacement provides five blue and five red spheres, which are replaced in the urn. Given the data, both Sally and Tom should assign Pr (blue|random draw)= 0.5. However, whereas Tom only has at disposal the *frequency* data, Sally also has an explanation for why the probability of drawing a blue x-sphere at random is .5. Yet, this evidential difference does not show up, given the initial data, which prompts both to assign Pr (blue|random draw)= 0.5. Suppose, however, that 10more x-spheres are drawn at random and they are all blue. Given her knowledge about the atomic structure of the x-spheres, Sally's rational credence in blue given a random draw remains the same - $Pr_{Sally+E}$ (blue|random draw)=.5, where this is her revised probability, upon learning the result of the new draw, E. Yet Tom's rational credence changes significantly. Whereas Pr_{Tom} (blue|random draw)=.5 for the initial round of testing, Pr_{Tom+E} (blue|random draw)=.75 when he learns about E. So, Tom's probability function is more *volatile* in the face of the new *information*. Sally's probability function is more *resilient*.

In the terms of Joyce (2005), adopted by McCain and Poston, Sally'credence stably *concentrates* on the hypothesis that the objective chance of blue x-spheres is 0.5. Moreover, at this point, one could usefully recall Lipton's insightful observation that a serious problem for Bayesians is that they cannot distinguish which is the *relevant* evidence (Lipton, 2004, p. 116).[84] One would thus be tempted to add that the 10more x-spheres which are drawn at random and are all blue, while constituting *information*, do not constitute relevant *evidence* (where relevant evidence is that which can affect one's credences). It is a line of thought that McCain and Poston do not adopt explicitly, although it might well lie in the back of their specific use of the term 'information', and of their repeated differentiation between frequencies, epistemic and non-epistemic, objective chances and epistemic probabilities. At any rate, McCain and Poston conclude that explanatoriness is evidentially relevant *even* when the screening off condition holds, by increasing the total weight of evidence.

We can see that our proposal with respect to mechanisms and resilience, is consistent with, and naturally strengthens, McCain and Poston's rationale as to how explanatory aspects are evidentially relevant. In a sense, our proposal is the natural continuation of the series of suggestions made by Skyrms, Joyce, and McCain and Poston, as well as of the discussion of evidential weight provided in chapter 4. If (*a*) one needs to take into account the phenomenon of resilience of probabilities (as Skyrms urges); (*b*) resilience is an important symptom of the weight and quality of evidence, to be

[84] 'Let us begin with evidence. Bayes' theorem describes the transition from prior to posterior, in the face of specified evidence. It does not, however, say *which* evidence one ought to conditionalise on' (Lipton, 2004, p. 116).

distinguished from its balance (as Joyce has maintained); and (*c*) it is the explanatory features that increase the weight of evidence and the resilience of probabilities (as we have just see McCain and Poston arguing), then it would only be natural to take the evidence of mechanisms as the main type of evidence that can increase the resilience of probabilities – and again, the discussion in chapter 4 on how mechanistic evidence increases the weight of the total evidence we have, can only support the above rationale. It is mainly mechanisms that explain, and they do so in virtue of stabilizing causal relations – which is an ontic phenomenon that parallels the epistemic phenomenon of 'resilience' for Bayesian probabilities.

One could even claim, with respect to McCain and Poston's discussion that, in their examples, the evidence brought about by mechanisms is almost tacitly present. In the example with the x-spheres for instance, it is mechanistic evidence which is at stake in the difference between the resilience of the probability functions of the two participants at the experiment. When reading such examples, the notion of mechanistic evidence is, as it were, on the tip on the tongue; yet, the notion just remains unarticulated by McCain and Poston. We can now close the circle and return to the mechanistic evidence and the criteria for its quality provided in Clarke *et al.* 2014.

§4 Evidence of mechanisms and the resilience of credences

We have seen in the previous section that McCain and Poston ground their argument in favor of the resilience of probability functions on the distinction between weight and balance, where weight is related to the *qualitative* aspect of evidence, which arguably cannot be captured in pure probabilistic terms. Now, a crucial piece in the puzzle of how to apply in a medical context the rationale of the resilience of probability functions, is that the criteria of mechanistic evidence advanced in Clarke *et al.* are related precisely to this qualitative aspect. As mentioned at various points in this thesis, these criteria are means to assess and provide a hierarchy for the *quality* of medical mechanistic evidence. They take into account traits such as independent methods, different research groups, proportion of features found, knowledge of analogous mechanisms, and robustness, defined in terms of being reproducible across a wide range of conditions.

We have seen in chapter 2 that each of these traits plays a role in terms of pluses and minuses, and in the pre-confirmation stage they are used for grading the 'quality' of evidence. For instance if we take a case in which the mechanistic evidence comes from two different teams reporting two mechanistic features, using the same research methodology - this evidence will be graded higher than the evidence of two mechanistic features provided by a *single* team, but will be graded lower than evidence coming from two different teams reporting two mechanistic features using *different*

methodologies; in turn, the latter evidence will be graded lower than evidence coming from two different teams reporting *three* mechanistic features using different methodologies, and so on.

Now, if the quality of mechanistic evidence is traced back to the balance/weight distinction, then we can see immediately the way to apply to our medical context the proposal concerning the resilience of probabilities brought about by mechanistic evidence. The way to apply it would be to say that, for the medical correlations obtained from the controlled population studies, the corresponding Pr function should be more resilient, to the degree to which the criteria of Clarke *et al.* for *mechanistic* medical evidence are satisfied. In other words, on this rationale, the quality of medical mechanistic evidence will be reflected in the resilience of the Pr function from the controlled population studies, which will be higher for those hypotheses for which we also have mechanistic evidence obtained using different methodologies, different research teams, with more mechanistic features found, etc.

Take again the example of smoking and lung cancer, as discussed by Roche and Sober. In plain terms, according to them, the mechanism linking smoking and lung cancer does not have any relevance for the correlation between smoking and cancer, no matter whether Fisher's supposition was adequate or not. This might be right, but, using again the plain language of causation, it is evident that the mechanism in question has relevance for whether smoking does cause lung cancer or not (and whether any intervention or manipulation can intervene beneficially). It sounds almost like a truism, and denying this truism is a direct consequence of what Roche and Sober say.

It is this causation worry which underlies McCain and Poston's distinction between frequency and non-epistemic chance, and finally their reasoning as to the resilience of probability functions when backed up by explanatory considerations. It was also this causation worry which triggered the first formulation of the thesis that evidence of mechanisms should be placed alongside (and on the same footing as) evidence from controlled studies, in the original (2007) article of Russo and Williamson, as we have discussed in chapter 4. More precisely, in that article it was claimed that, in the absence of evidence *of the existence* of a medical mechanism, one cannot rule out that a correlation, however rigorously established on a population level, could be spurious (Russo and Williamson, 2007, p. 159) and the claim of ruling out spuriousness is repeated in the more recent Clarke *et al.* 2014 (p. 343).

Recall now my suggestion made in chapter 4, in the context of the weight of evidence discussion - as far as the mechanistic evidence is concerned, there should be more work they could do for *confirmation* purposes than just ruling out cases of spurious causation. In chapter 4, I have postponed developing my suggestion, since the context there was the pre-confirmation grading of evidence. The proposal in the present chapter is precisely a development advanced in chapter 4.

Indeed, referring more precisely to the Clarke *et al.* criteria of mechanistic evidence advanced in the same Clarke *et al.* - these criteria can do more than that, in combination with the evidence of correlation from population studies. It is perfectly acceptable to say that the existence of a mechanism helps establish (or even establishes) that a medical correlation is non-spurious. But mechanisms also explain causal relations, and ontically, they stabilize causal relationships (by offering a degree of protection against interfering factors, using various feed-back paths, etc.). Moreover, beyond the simple *existence* of a mechanism, this explanatory relation is all the more direct, the more the criteria of mechanistic evidence of Clarke *et al.* are fulfilled. How could this explanatory relation of mechanisms be put to work at the level of confirmation? At the very least, I propose, it is in making more resilient the probability functions resulting from a Bayesian interpretation of the acclaimed RCTs of EBM, in direct correspondence or dependence with the Clarke *et al.* criteria.

Suppose that a medical study S provides evidence that for the hypothesis H_1 that C_1 causes E, with a strong resulting balance – e.g., that $P(E|C_1) = 0.7 > P(E) = 0.1$. Hence we have that $P(E|C_1\ S) = 0.7$. Now, some future evidence (study T) might cast doubt on this causal relation, and thus we would have, say, $P(E|C_1\ S\ T) = 0.3$, where 0.3 would be a compromise between 0.7 and 0.1 = P(E). The evidence of the study T would favour another hypothesis H_2, according to which it is C_2, and not C_1, which causes E.

However, if at the time of doing the study S, one had good, quality evidence M_1 that there is a mechanism of action by which C_1 causes E, then although this mechanistic evidence would make no difference to the original probability obtained after conditionalising on the evidence of S, namely $P(E \mid C_1\ S\ M1) = 0.7$, it would make that probability more resilient under the putative future evidence T, such that, say, $P(E \mid C_1\ S\ M_1\ T) = 0.65$. And having good, quality evidence M_1 in favour of hypothesis H_1 would mean that the Clarke *et al.* criteria of mechanistic evidence are satisfied, and that the proportion in which they are satisfied makes it far superior to any alternative mechanistic evidence M_2 that may be put forward in favour of rival hypotheses such as H_2.

We can now return to another point we left in suspension in chapter 4 – the prospects of theory confirmation in the current controversy as to the role of cholesterol in atherosclerosis. Recall that, against the widely accepted view that cholesterol has a pathogenic role in atherosclerosis, a number of (rather isolated) scientists and public figures have advanced the hypothesis that cholesterol is actually an innocuous factor. On this innocuousness hypothesis, the observed correlation between cholesterol and cardiovascular diseases is due to some common cause, like lack of physical activity, mental stress, smoking and obesity (Ravnskov, 2002). Now, in adopting the pathogenetic hypothesis on cholesterol and rejecting the hypothesis on innocuousness, we do not

seem to be dealing (only) with the issue of ruling out spurious cases of causation. Moreover, proponents of the innocuous hypothesis seem to be on the track of a mechanism; it is not entirely implausible that mental stress, smoking, obesity and lack of physical activity could induce directly atherosclerotic lesions.

However, when it comes to the application (and fulfillment) of the Clarke *et al.* criteria, the pathogenetic hypothesis fares far better than the innocuousness hypothesis. The number of research groups and different methodologies used to pin down the pathogenic mechanism of cholesterol far outreaches that used for the opposite hypothesis. What is even more important, the number of discovered features of the pathogenetic mechanism (including thirteen Nobel prize discoveries; see Endo, 2010) is far superior to the number of mechanistic features backing up the innocuousness hypothesis. Population studies grouping participants with very different characteristics and background (age, race, dietary habitudes, etc.) have shown a high degree of robustness and stability for the pathogenetic hypothesis (Steinberg 2005; 2005b) that contrasts with the fluctuation of results of studies (mostly systematic reviews) in favour of the innocuousness mechanism.[85]

All in all, the quality of mechanistic evidence for the pathogenic hypothesis is superior to the quality of mechanistic evidence for the innocuousness hypothesis. Since the explanatory role of mechanisms is all the more direct, the more criteria of Clarke *et al.* are satisfied, we should have at the level of confirmation that the Pr function resulting from a Bayesian interpretation of population studies in favour of the cholesterol hypothesis is more resilient than the Pr function of the innocuousness hypothesis, thus reflecting the far superior quality of mechanistic evidence (no matter what the balance of population studies shows).

Parenthetically, one of the advantages of this resilience approach is that it leaves room for revolutionary theories (and this is one of the reasons I have chosen to return to the example of the cholesterol controversy). Proponents of the innocuousness hypothesis have often complained of being censured and of having limited access to the academic publishing sphere (and this is a complaint which, of course, given the tremendous influence of pharmaceutical companies, might have some grain of truth in it). However, on the resilience approach, theories such as the innocuousness hypothesis have still the door open.

---

[85] For the latest contribution in favour of the innocuousness hypothesis, see Ravnskov et al. 2016. A response from mainstream proponents of the pathogenic hypothesis, can be found at http://www.sciencemediacentre.org/expert-reaction-to-systematic-review-reporting-lack-of-an-association-between-ldl-cholesterol-and-mortality-in-the-elderly/
The response of the Oxford center for evidence based medicine can be found at http://www.cebm.net/cebm-response-lack-association-inverse-association-low-density-lipoprotein-cholesterol-mortality-elderly-systematic-review-post-publication-pee/

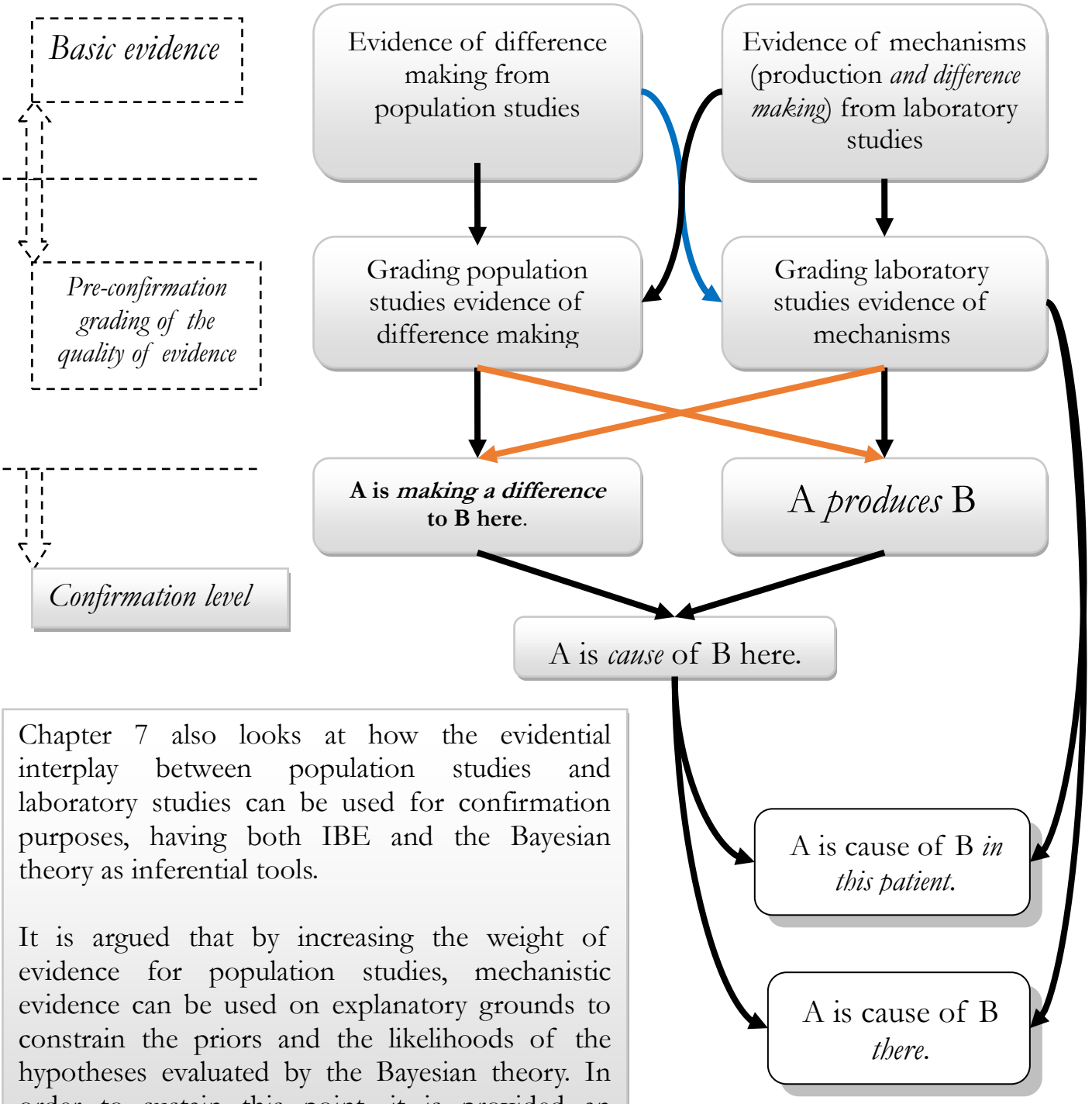
What I mean is that, it is the case that quality mechanistic evidence should increase the resilience of corresponding hypotheses. It is not the case that, once good mechanistic evidence is brought in, the rival hypotheses and contrary evidence will simply be ignored. But what is asked from such rival hypotheses (beyond the results of population studies) is that they provide quality *mechanistic* evidence. Given the subtle and penetrating nature of the Clarke *et al.* criteria, such mechanistic evidence could be judiciously assessed, and promising evidence would (and should) be given appropriate heed. From this perspective, the message to be sent to the proponents of the innocuousness hypothesis would just be – tell us a bit more about your mechanisms, and we will see what we can do about the resilience!

Subsequent work in the application of Bayes nets to mechanisms themselves (of which Clarke et al. 2014b is a very encouraging starting point) as well as, as one should hope, developments of McCain and Poston's own proposal, are likely to make this resilience role of mechanisms more explicit. But it is, I suggest, a very fruitful path of research.

Conclusion chapter 6

There are various ways in which one can conceptualise the 'friendly companionship' between IBE and Bayesianism in the realm of confirmation. I have presented here several options and pursued one proposal which seems the least contentious of all, namely that explanatory features increase the resilience of probability functions. I have shown its relevance for the medical issue of the interplay between mechanisms and population studies and I have proposed that, for the correlations obtained from the population studies, the corresponding Pr function should be more resilient, in proportion to the degree to which the criteria of Clarke *et al.* for medical mechanistic evidence are satisfied, thereby developing the epistemic advantage *VI)* of the revised RWT. The next chapter will look at the more contentious aspects of the collaboration between IBE and Bayesianism.

Chapter 7



Chapter 7 also looks at how the evidential interplay between population studies and laboratory studies can be used for confirmation purposes, having both IBE and the Bayesian theory as inferential tools.

It is argued that by increasing the weight of evidence for population studies, mechanistic evidence can be used on explanatory grounds to constrain the priors and the likelihoods of the hypotheses evaluated by the Bayesian theory. In order to sustain this point, it is provided an alternative justification for the use of explanatory values in inference

Chapter 7. On constraining priors and likelihoods

Introduction

The present chapter looks at the more contentious aspect of the collaboration between IBE and Bayesianism. As noted in the previous chapter, the use of explanatory values in order to constrain the priors and likelihoods of hypotheses, advocated among others by Lipton (2004) and Okasha (2001), faces the great obstacle that intuitively, with respect to theory confirmation, it is hard even to imagine how one could provide an *objective* assessment of *values* and use them at all for confirmation purposes - irrespective of whether those values are explanatory values, ethical or pragmatic.

This intuitive difficulty is present not just in particular debates such as the possible use of explanatory values to constrain priors and/or likelihoods – it also comes to the surface in general discussions about the principled compatibility between IBE and Bayesianism, beyond any particular proposal as to the details of their collaborations. Indeed, Lipton (2004) and Okasha (2001) might have provided a valid response to van Fraassen's charge of incompatibility and irrationality, as discussed in the previous chapter. But no matter how ingeniously and insightfully IBE theorists could argue in order to respond to charges such as van Fraassen's, the intuitive difficulty mentioned above concerning the very use of values in theory confirmation is hard to set aside, especially for Bayesian authors. The critical discussion in Roche and Sober (2013) - the first half of which we have surveyed in the previous chapter - stems from the same conviction that, at bottom, IBE could have nothing to add to Bayesian inference, except for making the latter invalid.

Interestingly, some IBE theorists have argued, picking the other end of the stick, that IBE would be dissolved as an autonomous referential method, if it was used complementary with Bayesianism, as a means to plug in explanatory considerations into the Bayesian machinery (Psillos, 2004). This latter possibility has been a constant source of worry for proponents of the compatibility between IBE and Bayesianism (e.g. Iranzo, 2008), and naturally, has been exploited by adversaries of this compatibility thesis.

The key of my argument in the present chapter is that, by providing a sufficiently strong defense of the *objectivity* of explanatory values, one can bring to relief the both general worry of the compatibility between IBE and Bayesianism and the particular worry that IBE might lose its substance by being mingled with Bayesian conditionalisation, and one can accordingly justify the constraining of priors and likelihoods by explanatory considerations.

The argumentation I will employ in favour of the objectivity of explanatory values uses, just like in the previous chapter, but from a different perspective, the heuristically rich account of epistemic causation developed by Russo and Williamson, and appeals also to David Lewis's construal of scientific laws. It offers a way to conceptualise the companionship between IBE and Bayesianism, explaining *how* IBE and Bayesian inference are compatible, while remaining *different* types of inference, and also supports the more contentious proposal of Lipton and Okasha that these explanatory values could constrain priors and likelihoods.

The plan of the present chapter is as follows. §1 offers the first line of argumentation in favour of the objectivity of the explanatory values employed in IBE. It appeals to an ideal scenario revolving around David Lewis's construal of scientific laws (or, in the terms I will rather employ, of the nomological structure of the world). §2, compares my Lewis-derived strategy of defence of the objectivity of explanatory values (and accordingly of the IBE based inferences) with Russo and Williamson's strategy of defending the epistemic causality associated to *objective* Bayesianism, which is based also on an ideal scenario argumentation. I show that the two strategies are, in an important respect, *analogous*, while being from another important perspective, *different*. My rationale is that, if both these strategies (and accordingly, the inferences they are supposed to justify) are analogous, defendable and yet different, they show the way for conceptualizing IBE and Bayesianism as *distinct* and yet *complementary* ways of selecting the right causal hypotheses, in real-life situations in which the ideal scenarios *do not hold*.

In very brief, I argue that these inferences are analogous since they are justified by the appeal to ideal scenarios and can be construed as hunting down the ultimate nomological structure. They are defendable insofar as, in real-life situations in which the ideal scenarios *do not hold,* their results can be seen as progressively approximating this nomological structure – all the more, as science progresses. They are different insofar as one of them, the Bayesian inference, hunts down for the quantitative aspect of this nomological structure, whereas the other, i.e. IBE, hunts down the qualitative aspect of this nomological structure. The analogy and the difference result in complementarity due to the fact that it is the *same* nomological structure that is hunted down (or towards which both methods of inference are guiding).

§3 looks at the second part of the dispute (and exchange of articles between) McCain and Poston, on the one hand, and Roche and Sober, on the other hand; I show that Roche and Sober's claim that explanatory features are evidentially irrelevant hinges on a neglect of the abovementioned contrast between ideal and real-life scenarios.

In §4 I return to the medical side of the discussion and and lay down the proposal which corresponds to the point *VII)* of the list of epistemic advantages of RWT, namely that the Clarke *et*

al. criteria of mechanistic evidence could be employed to constrain the prior and/or likelihood probabilities established by the Bayesian theory taking into account the population studies evidence.

Now, since the pair of concepts of ideal and real-life scenarios will play an important role in this chapter, a word of preliminary clarification is due in this introduction. *Ideal scenarios* are epistemically transparent scenarios in which one has access either to the *total evidence* tout court, or alternatively, to all the particular evidence relevant for particular inferences, such that the reliability of these inferences is guaranteed *a priori* by the respective epistemic *transparency*. Such ideal scenarios are useful as thought experiments in order to settle various disputes with a *metaphysical* background - we can think here, for example, of the way the possible worlds apparatus is used in the metaphysics of causality, in relation to the counterfactual approach to causation *à la* Lewis, or in the metaphysics of natural kinds, in relation to the Kripke/Putnam account of reference. More precisely, these ideal scenarios allow the theorist to speak about (metaphysical) facts, without the need to count in the possible ignorance or epistemic opacity on the part of the theorist (although of course epistemic opacity can be also brought in, as in the two-dimensional semantics). It is the famous 'God's eye view'.

So for instance, Russo and Williamson use a type of ideal scenario in order to show the coherence of their own account of causality, based on *objective* Bayesianism. On my part, I will be using an analogous scenario in §1, in order to lay down, in Lewis' footsteps, a justification of the use of values in IBE, on the ground of these being *objective* values. I will also appeal to the discussion of such ideal scenarios in §3, in order to disentangle the metaphysical and epistemic aspects of Roche and Sober's thesis on the evidential irrelevance of explanations, using the Kripke/Putnam account of reference as a clarifying illustration for the way in which the notion of epistemic transparency has a bearing on Roche and Sober's thesis.

On the other hand, real-life scenarios assume explicitly the epistemic opacity of evidence. The real-life scenarios are obviously useful to consider because in science we are actually dealing only with such scenarios, in which we do not have access to the total evidence (or to all the evidence relevant for particular theories and hypotheses), and therefore the reliability of inferences is not guaranteed by the respective epistemic opacity.

To underlie, the distinction between ideal (or epistemically transparent) and non-ideal (or epistemically obscure) scenarios is *not* be reduced to the distinction between availability or non-availability of the *total* evidence (although the latter distinction is a paradigmatic illustration of the former), but applies also to the distinction between availability of all *relevant* evidence for a particular theory or hypothesis, and its non-availability. This is why the use of ideal and non-ideal scenarios is common in a multitude of philosophical discussions in which there is no mention of *total* evidence,

as for instance in the philosophy of language. It is for this reason that the Kripke/Putnam account of reference will find its place in my argumentation in §3. Actually, most discussions as to the correct way our words refer, employ examples in which a certain term or expression is analysed with respect to its capacity to refer to certain state of affairs, where our access to the respective state of affairs is framed in terms of the ideal scenario (on the part of the theorist), and our use of the respective term or expression entails the possibility that the non-ideal scenario holds, in the sense at least that one might not be aware of how and to what one's terms refer (does 'water' refer to H_2O?; *cf.* for instance Quine's distinction between referentially opaque and transparent contexts, or Tarski's truth definition). On the other hand, theories of explanation in the philosophy of science are also congruent with the distinction between ideal and non-ideal scenarios in that the ontic theories of explanation suit the ideal scenarios whereas the epistemic conception of explanation suits the non-ideal scenario.

Finally (and hoping the reader will not have lost her patience with this long, but necessary introduction to this chapter) to take the distinction from a different angle, in ideal scenarios one can *be safe* to speak about objective and logical probabilities, or objective chances; in real-life scenarios one can only *be safe* to speak about subjective probabilities and frequencies. Ideal scenarios are limiting cases in which the use of explanatory values in IBE justified. On the other hand, in real life scenarios, IBE can only be a *guide* to inference, and not the ultimate inference itself, just like, in such real-life scenarios, the Bayesian inference needs to assume important caveats in its practical application. It is this difference between ideal and real-life scenarios, as I will round off the general discussion in §3, that shows the useful complementarity of Bayesianism and IBE as *different* ways of guiding our search, in the real-life situations, to inferences that approximate as much as possible those in the ideal-scenarios. And, as I will suggest in §4, drawing Bayesian inferences in which the priors and likelihoods are constrained by the available mechanistic evidence, via the Clarke *et al.* criteria, might just be such a way in which one uses the advantages of both ways of guiding our search for medical causes, in order to come as close as possible to the nomological structure that characterizes medicine.

§1 Justifying the explanatory values

How are explanatory values *justified*? One frequently cited argument in the literature, which I have also laid down in chapter 1, starts from the common observation that much scientific discovery and acceptance takes places on the lines of IBE, and proceeds by meta-induction to the actual use of these values in science is justified (Niiniluoto, 1999). Alternatively, value epistemology (including its neo-Aristotelian tenet) could offer plausible ways of integrating these values in a

normative framework (Wilkenfeld, 2014). However, one should admit that very use of terms such as 'values' in a sensitive and heated discussion as that concerning confirmation in the philosophy of science is likely to carry along an aura of arbitrariness. To use Jonathan Vogel's words 'under these circumstances, believing a hypothesis because of its explanatory value would not be much better than believing it because someone thought it up on your birthday' (Vogel, 1998). Or, more closely to our topic, Henderson argues that 'one might still stipulate […]that we should constrain priors in such a way as to agree with IBE […] this means giving 'normative primacy' to IBE and effectively recommending a new form of objective Bayesianism. This is a possible response to van Fraassen's challenge, but without independent motivation, it appears more like a *stipulation* designed to ensure that Bayesianism is compatible with IBE, rather than an explanation of why it is. It has the consequence that IBE is merely accommodated in, rather than explicated by, the Bayesian framework. And it also raises the question of why this new form of objective Bayesianism should be preferred over previous attempts to provide objective Bayesian norms' (Henderson, 2014, pp. 10-11, italics added).

So again, how could one bring in values into such a sensitive discussion as the confirmation of scientific theories? I think the most fruitful way of justifying explanatory goodness is not by tying them *directly* to the epistemic subject and the (reliable) inferences s/he is supposed to draw, but seeing primarily these values as part of the ontological description of the nomological structure of the world.[86]

Let me begin with what might appear as a startling observation (for the reader with a positivistic slant). Take the famous account of scientific laws provided by David Lewis - the laws of nature are the axioms or theorems of a true deductive system that achieves the a best combination of simplicity, strength and fit. Simplicity means here conciseness and lack of additional assumptions, strength means covering as large as possible an area of phenomena in the world, and fit means suiting with the actual outcomes of world history (Lewis 1973, p. 73, Lewis, 1994). Now, on a closer look, what we see in this account of laws, put forward by one of the most austere philosophers of science, are values, or qualitative characterizations of the nomological structure that scientific laws should represent. I mean to say that, arguably, Lewis' 'simplicity' parallels the explanatory value with the same name, his 'strength' corresponds precisely to the explanatory value

[86] I would like to stress that I do not wish to diminish the importance of meta-induction arguments from the actual use of these values in current scientific practice, or disconsider the arguments coming from value epistemology. But I am bound to admit that in the literature on theory confirmation, IBE and its explanatory values are still regarded with a certain degree of skepticism (especially by Bayesian authors). Thus, my attempt in the present paper is to offer an alternative justification of explanatory values, using a strategy that is analogous to the strategy used by some established Bayesian authors in justifying their brand of objective Bayesianism (e.g. Russo and Williamson, 2007, Wilde and Williamson, 2016). It goes the same with my attempt in the following to link the objective status of explanatory values to Lewis' account of scientific laws – an account which surely has more proponents among Bayesian theorists than IBE itself has.

of 'scope', his 'fit' is very close to the explanatory value of 'individualisation', and the explanatory value of 'theoretic unity' is built into the very idea of a deductive system with theorems and axioms.

Importantly, Lewis conceives of his Best System analysis as an outcome of knowing virtually everything, the entire world history. It is, that is to say, an ideal scenario in which an omniscient being hierarchizes the universal generalisations that can be extracted from the entire world history. And plausibly, the way these generalisations are obtained and hierarchized by the omniscient being follows a pattern that corresponds to the pattern of explanatory values at work in IBE.

To put it differently, the explanatory values could arguably be taken to correspond to the qualitative aspect of the mapping of the world realized by the Best System analysis. The Best System is not reducible to these values, since, for instance, its quantitative or numerical aspect is not fully taken into account by the pattern of explanatory values.87 But again, it is still the case that this pattern plausibly constitutes a qualitative way to describe the nomological mapping of the world in Lewis's ideal analysis.

The idea of linking values to this nomological framework should not seem very surprising. The role of laws is to systematize, bring phenomena under their cover, describe them accurately and at the same time be simpler than a purely factual description of the world. If we think about causal laws, the rationale is even more intuitive. For a multiplicity of given effects, an etiological analysis reduces this multiplicity to a simpler set of causal factors which has the respective effects under its scope, is descriptively adequate insofar as the causal effect-relations are brought to the fore and hence are individualized, and the more the etiological analysis is pursued, the more likely the set of causal factors is to manifest theoretical unity (by being linked up with another set of causal factors, by turning up to be themselves effects of a higher set of causes, etc.).

To sum up, when looked at through the lenses of Lewis' account of scientific laws, the explanatory values have an objective status, that parallels (and derives from) the objective status of scientific laws. Of course, we can disagree about what precisely the ontological status of these scientific laws is. That is to say, one can view them as relations between universals (*à la* Armstrong, 1978) as characterisations of essential properties of natural kinds (*à la* Lowe, 1989 and Ellis, 2001) as systematisations of causal relations founded in anti-Humean dispositional properties (*à la* Bird 2007), or indeed, as Humean generalisations carrying with them a sort of necessity or non-contingency, to be described (but arguably not explained!) by the possible worlds semantics (as

[87]To clarify, this emphasis on the *qualitative* aspect is proper to the pattern of explanatory values, which provide non-numerical, non-quantitative description of scientific laws. The quantitative aspect can also be found in Lewis' account of laws, via the notion of fit (which, besides this numerical aspect, is also linked, on the qualitative side, to the explanatory virtue of individualisation, as I have stated above).

Lewis himself maintained). However, one needs not settle here the issue of the ultimate construal of laws. I have chosen Lewis's account above for the ease of exposition, since it offers us a clear view of the justificatory framework provided by ideal scenarios, and it is an account that is more neutral than other, more metaphysically loaded, accounts of laws. In spite of the sophisticated disagreement of philosophers of science over how the fine-grained rendition of laws should look like, what is generally accepted is that scientific laws have their own strain of objectivity, no matter whether one takes such laws as ontologically primary or derived from more fundamental entities. And the point of the present section is that, *a fortiori*, the explanatory values have themselves the same strain of objectivity, as providing a qualitative description of these laws.[88]

Such a defence of explanatory values, and accordingly of the reliability of IBE as a method of inference, is stronger than the one provided by arguments driven by the actual success of scientific practice, or the ones advanced within value epistemology. The reason is that (with the help of ideal scenarios such as Lewis') one can bring to the fore a metaphysical side of discussion that should tip the balance within an epistemic debate (not to mention that in most strictly epistemic discussions, this metaphysical side is also present in a tacit way).

To take a commonplace example, Hume's epistemic arguments against the reliability of induction are also based, among others, on a metaphysical stance that denies the existence of causal powers. Were one to provide an argument in favour of the *objective* existence of causal powers, one would thereby *also* argue in favour of the reliability of induction, even though more steps would have to be undertaken, as for instance the argumentative step in which one shows that we can be acquainted with these causal powers (see Ellis, 2001).

This is not, however, to enter into the discussion of the existence or non-existence of causal powers. It is just to use a commonplace example in order to illustrate how an apparently pure epistemic discussion has one of its roots into (and can be settled by) a metaphysical argument,[89] or more precisely, how the *objectivity* justified by the metaphysical side of the discussion can be used to shed light on the corresponding *reliability* of a certain type of inference. One thing that this example shows, however, is that the move from objectivity to reliability discussed above is not immediate or direct, but, on the contrary, entails intermediate steps. The anti-Humean theorist will have to go through the intermediate step of providing an account of acquaintance with the causal powers. On its side, the IBE theorist has to face the gap between the ideal scenario and the real-life situations in which our access to evidence is incomplete. By and large, the optimal combination between simplicity, theoretic unity, scope and individualization (the latter, *inter alia*, fulfilled by way of

[88] I will come back to Lewis' account in the concluding remarks of this paper, when rounding off the main thread of my argumentation.

[89] One could of course invoke here Devitt's famous imperative: put metaphysics first! (Devitt, 2010).

providing mechanisms) should track down the truth of the inferred hypotheses (Lipton, 2001, Glass, 2012, Douven and Wenmackers 2017). But limit cases, part of real-life scenarios, could also be envisaged, and one just needs to recall here the burglars case envisaged in in Weisberg (2009) and discussed in the previous chapter.[90]

Cases like this could not contribute to knock-out arguments against the principles use in inference of explanatory values, but they seem to pose difficulties to deriving the complete reliability of IBE in real-life situations from the objectivity of explanatory values circumscribed in the ideal scenario. However, in fact, such cases and examples underlying the difference between the ideal and the real-life scenario provide actually the key to understand the complementarity between IBE and Bayesianism. In order to see why, I will look in the next section at an ideal scenario employed in a Bayesian framework.

§2 Analogies and differences with another ideal scenario

It is time to use again the heuristical resources of epistemic causation. Compare the above justification of the values involved in IBE with the strategy of defending this account associated to the brand of *objective* Bayesianism proposed by Russo and Williamson (Williamson 2006, Russo and Williamson, 2007). Recall that the epistemic account of causality takes causal relations to be the causal beliefs that an agent with access to *total* evidence should adopt (Russo and Williamson, 2007, p.167). And the connection with the objective Bayesianism is straightforward: the causal beliefs in question should be represented by a directed acyclic graph whose nodes are the variables of interest and whose arrows correspond to direct causal connections (Williamson 2006, pp. 75-82), where this graph is constrained by evidence and should otherwise be as non-committal as possible as to what causes what.

The three requirements embedded in such an account, namely *acyclicity* (one's causal claims should be representable by an acyclic graph C) *calibration* (one's causal claims should fit evidence: C $\in$ E, the subset of acyclic graphs that fit evidence) and *equivocation* (C should otherwise be as non-committal as possible about what causes what) correspond to the three main requirements of objective Bayesianism, namely *probability* (one's degrees of belief should be representable by a probability function P_E), *calibration* (one's degrees of belief should fit evidence: $P_E \in$ E, the subset of probability functions that fit evidence) and *equivocation* (one's degrees of belief should otherwise equivocate as far as possible between the elementary outcomes) (Wilde and Williamson, 2016, p. 6,

[90] Examples such as Weisberg's are used to discuss how one or another explanatory value might be the wrong guide to inference are frequent in the literature. Parenthetically, I think it is misleading to use just examples in which *one* of the explanatory values fails inferentially. On the contrary, one should discuss the inferential use of the optimal combination of these explanatory values, in the same way in which, in Lewis' Best System, there is an optimal balance of certain qualitative characteristics. But due to limited space, I will not go further in this direction.

Russo and Williamson, 2007, p. 168).

Crucially, both the objectivity of this strand of Bayesianism, and the objectivity of the corresponding epistemic account of causation, are derived from the limiting case in which an omniscient rational agent has access to the total evidence. In Russo and Williamson's words: 'Causal relationships are to be identified with the causal beliefs of an omniscient rational agent. This gives *a view of causality that is analogous to the objective Bayesian view of probability, according to which probabilistic beliefs are determined by an agent's evidence, and probabilities themselves are just the beliefs that an omniscient agent should adopt*.' (Russo and Williamson, 2007, p. 168, italics added).

We can see that the pattern of explanatory values drawn out of Lewis's Best System, as discussed in §1, and the objective Bayesian account of epistemic causality, as described above, are ways of (nomologically) mapping the world, in ideal scenarios in which one supposes the availability of the total evidence for an omniscient being. From this point view, we have a strong *analogy*.

On the other hand, arguably, there are *differences* between these ways of mapping. The pattern of explanatory values provides a *static* map, concerned, as I said in the previous section, with the *qualitative* aspect of the ideal nomological structure. The Bayes nets representing the beliefs of the omniscient being, even though capturing important qualitative features, provide a *dynamic* map that primarily takes into account the quantitative or *numerical* aspect of the ideal nomological structure represented by the causal laws.[91]

To summarize, the two strategies discussed above are analogous from an important perspective, and yet differ. They are analogous insofar as they are framed using ideal scenarios with an epistemic subject having access to the *total evidence* and mapping the facts of the world history in a (causal) nomological structure. They differ, *while being complementary*, insofar as the Lewis-derived strategy for justifying the objectivity of explanatory values captures the qualitative, static dimension of the nomological structure, whereas the epistemic causality strategy captures the dynamic, quantitative (or numerical) aspect which is characteristic of Bayesian networks.

Now, crucially, Russo and Williamson argue that the conclusions drawn in the epistemically ideal scenario are defendable in a non-ideal scenario, insofar as the causal relations - corresponding to the actual causal beliefs that we form - can be seen as approximating (and progressing towards) the ideal Bayesian inferential structure revolving around the omniscient being. Again, in Russo and Williamson's words: 'It might be thought that such a view renders causal relationships unknowable, for none of us can be omniscient, *but it is quite plausible that, roughly, the more we know, the closer our rational causal beliefs will correspond to the causal facts*, i.e., correspond to the causal beliefs of an omniscient rational agent. If so, then causal knowledge is possible.' (Russo and Williamson, 2007, p.

[91]Russo and Williamsondo not discuss about causal *laws* but such laws have an easily ascertainable place in their framework of epistemic causation.

168, italics added). In an analogous way, in the case of explanatory values, one can make the move from the ideal scenario to the real-life scenarios in which they should guide IBE inferences. What IBE and these explanatory values guide us towards, one would say, the more we know about the facts of the world, is getting closer and closer to the qualitative description of the nomological structure discussed above.

But getting closer and approximating the nomological structure implies that a perfect matching, in various stages, will not be in place, as far as the inferences of objective Bayesianism are concerned, on the one hand, and as far as the IBE inferences guided by explanatory values are concerned, on the other hand.

I have mentioned in the previous section cases of mismatch of failure for IBE and the associated explanatory values, and there are similar cases for the Bayesian inference, in which, in particular contexts, evidence in favour of the correct hypothesis brings about no change in the conditional probability, or even a decrease of it (Cartwright, 2007, Achinstein, 2001, McCain and Poston, 2014, Strevens, 2014).[92] When seeking to approximate and get increasingly closer to laws and the nomological structure they describe, for this perfect match, it seems more than reasonable that one should appeal to all forms of support available. This means that both the qualitative (i.e. explanatory) and the quantitative (i.e. Bayesian) guides could and should be used together in the search for the perfect mapping between our hypotheses and theories and the nomological structure.

A medieval saying has it that 'All roads lead to Rome'. The saying refers to the *Milliarium Aureum*, the golden milestone built up by Augustus in the central forum of ancient Rome, out of which all roads of the Empire were said to originate. If a group of medieval pilgrims wanted to reach Rome to deliver a message or simply visit the 'Eternal City', they would have better picked out at least two different roads (and kept in touch through whatever means of communication were then available). They would have thus minimized the risk of sidetracking, tiredness, famine, and other such opposing factors, and at least a few of could have finally reached the city. In our case, the saying translates into the useful convergence (and complementary use) of different methods of inference, as IBE and the Bayesian inference, which should increase the chances that the way we treat our real-life evidence and hypotheses resembles and increasingly approximates the way an omniscient being draws out of the total evidence of world history its nomological structure.

[92]Cartwright (2007) and Achinstein (2001) discuss cases in which we have a decrease of probability for the right hypothesis, brought about by relevant evidence. We have seen in the previous chapter that McCain and Poston (2014, and forthcoming) focus on the inability of probabilities to reflect the distinction between weight and balance of evidence. Strevens (2014) has an interesting discussion of the contrast between the logical omniscience assumption of Bayesianism and the 'humanized' Bayesianism of real-life situations. Interestingly, Russo and Williamson's argument that population studies are unable to rule out *spurious* correlations in the absence of evidence of mechanisms (Russo and Williamson, 2007) could easily be interpreted as pointing to the need to integrate the *qualitative* aspects of evidence in the 'inferential machine' or Bayesianism. I will come back to this in the final section of this chapter.

Again, this is not to ignore that failings and vulnerabilities of both IBE and Bayesianism. But the point is that their joint use, under the form of probabilistic conditionalization constrained in one way or another by explanatory values, should compensate for at least a part of these vulnerabilities.[93] For instance, Daniel Kahneman has famously documented how many intuitive uses of explanatory values can be vitiated by ignorance of the basic axioms and rules of probabilistic calculus (e.g. Kahneman, 2011); in the joint use of IBE and Bayesian inference, the quantitative strengths of Bayesianism should compensate for and remedy this vulnerability of explanatory reasoning. To take an example from the other side, Bayesian inference has notorious problems in dealing with the weight of evidence (Joyce, 2005), which is not reducible to (or reflected in) the simply numerical aspect of probabilities; in the joint use of IBE and Bayesian inference, the qualitative strength of explanatory values should compensate for and remedy this vulnerability of probabilistic reasoning.

I have nothing particular or technical to add here to the proposals that explanatory values should constrain the priors/likelihoods (Lipton 2004, Okasha, 2000), that they should increase the resilience of posterior probabilities (McCain and Poston, 2014), and/or that they could contribute to a brand of objective Bayesianism (Iranzo, 2000, Romeijn, 2013).

However, the above argumentation should provide a justificatory background for pursuing such proposals and seeking a way to applying them in particular contexts. It should also provide an incentive to always bear in mind in argumentation the difference between the epistemically ideal scenarios and the real-life, opaque ones. I will argue in the next section - by looking at the second part of the dispute between McCain and Poston vs. Roche and Sober - that it is precisely due to ignoring this difference that Roche and Sober are able to claim that the explanatory features are evidentially irrelevant, and that one should just stick with the Bayesian framework of confirmation.

§3 The dispute Roche and Sober *vs.* McCain and Poston

. We have looked in the previous chapter at the first part of the dispute Roche and Sober *vs.* McCain and Poston. This section will look at the second part of their dispute and show that the disagreement between these authors can be clarified when viewed through the lenses of the previous discussion from §1 and §2 on the difference between ideal and real-life scenarios.

Recall that Roche and Sober's main aim is to show that the explanatory goodness of a

[93] In cases in which i) the hypothesis favoured by IBE would differ from ii) the hypothesis favoured by the Bayesian approach, which in turn would differ from iii) the hypothesis favoured by combining the Bayesian approach with the use of explanatory values, I would be tempted to say that iii) provides us with more chances of hitting the target than i) and ii). I am aware however that this would require supplementary argumentation, and I am happy to stick with the more prudent position claiming that we should go with iii) in those cases in which ii) simply cannot give us a precise answer and ambiguity is to be dispelled; see Douven (2014) for discussion.

hypothesis cannot influence its *posterior* probability.[94] They propose that we take a hypothesis H, the observations relevant to it O, and formulate the proposition E: *if H and O were true, H would explain O*. In their terms therefore, the proposition E should encompass the explanatory relation between the hypothesis H and the observed data O. They ask subsequently, as the Bayesian confirmation theory demands, whether Pr(H|E) > Pr(H), or rather, whether Pr(H/E&O) > Pr(H/O). They ask, in other words, whether the explanatory features add anything to the confirmation of H. The answer is negative. Nothing is added to confirmation, because E is screened-off by O, or in formal terms, because Pr(H/O&E) = Pr(H/O). Hence we have the conclusion that 'the explanatoriness of H is evidentially idle, once the truth of O is taken into account' (Roche and Sober, 2013, p. 660). The argument has a deceivingly simple form, but of course some of the most powerful arguments in philosophy owe their appeal precisely to their straightforward exposition (cf. for instance Kripke's argumentation in favour of the rigidity of proper names).

We have seen that, in their first reply, McCain and Poston argue in the main that, even admitting the screening off is in place, the explanation features are still evidentially relevant. That is to say, they argue that the explanatory story does have an influence on confirmation or prediction, without modifying as such the probabilities. What they influence, more precisely, is the *resilience* of the Pr function, by making certain probabilities more stable, or less volatile, given new evidence (McCain and Poston, 2014, p. 148).

And here we enter into the second part of their dispute. Roche and Sober's response, provided in their (2014), is to acknowledge the resilience of probabilities in cases in which the explanatory stories are presented, but to contend that such explanatory stories simply concern *causal facts* that could be integrated into the background knowledge of conditionalization, and need not be viewed as 'explanatory' in a sense relevant *for* confirmation purposes, although they could be viewed as explanatory in themselves (Roche and Sober, 2014, pp. 196-197).[95]That is to say, according to Roche and Sober, the explanatory evidence just adds more grist to the Bayesian mill. *Qua* evidence tout court, the explanatory evidence just amounts to more data to enter into the inferential machine of conditionalization. *Qua* explanatory, such evidence plays no role in theory confirmation and accordingly explanations are still evidentially irrelevant.

Roche and Sober pick upon the example of the x-spheres provided by McCain and Poston,

[94] In the beginning of their (2013) article, Roche and Sober touch briefly on the issue of prior probabilities, dismissing the use of explanatory values for constraining priors with the casual reply that today's priors should be yesterday's posteriors.

[95] For the purposes of the present chapter though, whether the contribution of IBE to the confirmation area is qualified as simply contributing to the background knowledge (as Roche and Sober affirm) or as bringing to the fore a crucial facet of our confirmation practices (as McCain and Poston suggest) is not important. My feeling is that what is considered to the part of 'background knowledge' and what is taken to be as important as to be brought to the forefront, will depend much on whether the theorist in question is an enthusiastic Bayesian proponent or an equally enthusiastic proponent of the Inference to the Best Explanation, respectively.

which we have discussed in the previous chapter. Here is again the example (which I have transposed in a medical context). Suppose that (i) Sally and Tom have a credence of X in proposition H (that smoking is followed by cancer), (ii) Sally's credence is more resilient than Tom's, (iii) Sally but not Tom knows the mechanism of lung cancer and (iv) Sally but not Tom has an explanation of why the probability of getting lung cancer after smoking is X. Given Sally's knowledge as described in (iii), her credence in H should remain at X even if certain data from (biased) population studies shows a very low correlation between smoking and cancer. By contrast, given Tom's lack of knowledge as described in (iii), and given, thus, that all he has to go on is the observed frequency, it follows that if the observed frequency of cases of lung cancer following smoking deviates from X, then his credence in H would *not* remain at X.

Now, what we have here, contend Roche and Sober, is a case in which certain causal information from the background knowledge is relevant for evidential support (Roche and Sober refer to another example with x-spheres being drawn out of an urn, but this should not matter for exposition purposes)

> This is a case where differences in credences are dictated by differences in *background knowledge*. It is true that Sally but not Tom has an explanation of why the probability of a blue x-sphere on a random drawing from the urn is 0.5, but this difference between Sally and Tom is doing no work. It might be countered that (iv) is true because (iii) is true and that, thus, explanatoriness is still in play. We are not denying that explanatoriness is in play. In fact, we are supposing for the sake of argument that explanatoriness is in play in that (iv) is true. Our point is that (ii) is true because (iii) is true*, and (iv) does nothing to make (ii) true once (iii) is taken into account*. (Roche and Sober, 2014, pp. 196, italics added)

Finally, in their last contribution, McCain and Poston's reply is that IBE of course brings about *causal* stories, and bringing forth such causal stories is just how it contributes to resilience (McCain and Poston, *forthcoming, pp.* 3-4). Crucially, the way McCain and Poston articulate the causal side of the explanatory stories provides a way of tying in my argument from the previous sections - based on the difference between ideal and non-ideal scenarios - with their argumentation as to the resilience of probabilities. More precisely, the entry point is their discussion of how knowing the structure of water amounts to the possession of an explanation for the fire-extinguishing behaviour of water. Since this is a crucial juncture in my argumentation, the reader should allow me to give here the relevant quotation in full.

> [C]onsider an argument […] that denies that water has the property of extinguishing fire. Water is H_2O, and H_2O has special chemical properties that make it an excellent chemical to extinguish fire. Because of the high degree of hydrogen bonding between water molecules, H_2O has the second highest heat capacity of all known substances. In virtue of its high heat capacity, its transition from a liquid to a gas requires a significant amount of energy, which enables it to rapidly quench flames. Once we account for these facts about H_2O the fact that water is present is screened off. Thus, water is irrelevant to extinguishing fire because H_2O is doing all the work. The response to the water/ H_2O argument is obvious: Water = H_2O. The properties of H_2O in virtue of which it makes an excellent fire extinguisher are the properties of water. They are one and the same. We might put our point thusly: to the extent that water has the property of extinguishing fires, explanatoriness […] is evidentially relevant. Water extinguishes fire in virtue of its

chemical structure. Explanatoriness is evidentially relevant in virtue of it *specifying certain relations between H and E that get encoded in a Pr function*….if explanations are constituted by such causal facts then conditionalization on causal facts will screen-off explanatoriness. But this is just the sense in which water is screened off from H_2O. In other words, it is not really screened off at all because those facts have already been taken into account. Conditionalizing separately on Sally's knowledge of the causal relation (as expressed in (iii)) and her knowledge of the explanation (as expressed in (iv)) is counting the same facts twice (McCain and Poston, *forthcoming*, pp. 3-4)

Now, the example of water and its microstructure, as paradigmatically employed in the philosophy of language, can, I believe, make us see that what is at stake in the dispute between the four authors described above is the distinction between the ideal (epistemically transparent) and the non-ideal epistemically opaque) scenarios.

Recall that this distinction is not be reduced to the distinction between availability or non-availability of the *total* evidence, being used in a multitude of philosophical discussions, as for instance in the Kripke/Putnam account of reference. When a certain term or expression is analysed as to its capacity to refer to certain state of affairs, the access to the respective state of affairs is framed in terms of the ideal scenario (on the part of the theorist), and our use of the respective term or expression entails the possibility that the non-ideal scenario holds, in the sense at least that one might not be aware of how and to what one's terms refer (*cf.* for instance Quine's distinction between referentially opaque and transparent contexts; Quine, 1960). Does 'water' refer to H_2O? Perhaps, but in order for the very proposal of Kripke and Putnam to make sense in the first place, one needs to assume the ideal scenario on the part of the theorist, just like, on the other hand, intensionalist (or two-dimensionalist) theories of reference are working within the non-ideal scenario.

Let us come back to the main points of contention between the four authors above, this time with the distinction between the two scenarios at hand. Roche and Sober's claim that bringing in as evidence explanatory stories, in the form of causal mechanisms, does not contribute explanatorily to the confirmation of hypotheses but is a mere (background) causal information, assumes the ideal scenario, in which the relevance, for our (potentially fallible) epistemic position, of the (causal) facts in question, is substituted with (or by-passed by) a 'God's eye view', retaining from the explanatory *and* causal features of the evidence at hand just the causal features. But the explanatory side does not let itself be effaced so easily.

Mechanisms, under the heading of the explanatory value of individualization, figure prominently in the structure of IBE, and one would have to provide some quite strong reasons (stronger than the ones advanced by Roche and Sober) to strip off from the mechanistic evidence its explanatory dimension. Imagine an analogous example in which one comes to have evidence showing that a certain hypothesis, in contrast to its competitors, has a larger scope (is doing justice to a larger area of the phenomena in question). This evidence, which concerns the scope of a

hypothesis, would have a bearing upon the resilience of the probabilities in question (that is to say, would modify the weight of the evidence, without modifying its balance). We should naturally ask: how come the resilience of the probabilities is raised? The Roche and Sober-type of answer would be to underlie the purely factual aspects of the evidence in question. We are dealing with, the answer would go, with 'observation data', showing how such and such hypothesis is linked with such and such (previously unrelated, scattered) phenomena.

Undeniably, we would be dealing with 'observation *data'* here, but this is not the whole story. Because one would then need to ask: why is it, in the first place, that evidence about the greater scope of a hypothesis would in any way be considered *relevant* for the resilience of probabilities? And here, irrespective of whether one adopts the ideal or the non-ideal scenario, one would have to appeal to the explanatory virtues. On the non-ideal scenario, it would naturally follow that such evidence is relevant because we should be guided by the explanatory virtues, and the greater scope of a hypothesis should have a bearing on the weight of evidence and the resilience of posterior probabilities. On the ideal scenario, the one tacitly adopted by Roche and Sober, one would have to appeal to the ideal description of laws or nomological structure available to an omniscient being, which has (the greatest possible) scope (in optimal combination with other characteristics) as a *qualitative* feature.[96]

Claiming, for the case of mechanisms - that is to say, for the case of the explanatory virtue of individuation - that this is just background (causal) information, which is relevant but which should be relegated to the background knowledge, might do for practical purposes (because tacitly, we are more imbued with the use of explanatory features than we think), but it does not do for theoretical purposes, when the very use of such causal knowledge should be justified.

A proponent Sober and Roche's view might reply that I am burdening their account with features that they have not put forward, and that, when speaking about the background knowledge of certain (causal) facts that should increase the reliability of evidence, they only appeal to knowledge of certain conditionals referring to the microstructure of the analysed macroscopic phenomena (or, in my terms, referring to the *mechanism* of the analysed microscopic phenomena), and that there is not here the slightest hint of an ideal scenario, or of 'God's eye view'. But this would not do either (and the point here is not whether one is a Humean about causation or not). Again, why would such a condition be relevant and increase the resilience of probabilities?

[96] One could also argue here that, as I mentioned in the Introduction to this chapter, theories of explanation in the philosophy of science can be mapped upon the distinction between ideal and non-ideal scenarios, in that the ontic theories of explanation suit the ideal scenarios whereas the epistemic conception of explanation suits the non-ideal scenario. Roche and Sober's stance is, I believe, consistent with an ontic view of explanation, whereas McCain and Poston clearly stick to the epistemic view of explanation. This would complicate however the (already intricate) argument above.

For the explanationist, i.e. the IBE proponent, who works within the epistemically opaque scenario, it is simple. Mechanisms explain, and invoking them amounts to applying the explanatory values of individuation and increasing the weight of evidence, because thereby, our inferences approximate closer the nomological structure expressed by laws (and significantly, the very notion of resilience of probabilities was introduced in Skyrms 1977, in order to account for the nomological character of statistical law-statements). In the case of the epistemically transparent scenario, an omniscient being would know that evidence of mechanisms goes hand in hand with the resilience of probabilities, because the nomological structure he has at disposal just has as a qualitative characteristic the 'individuation', i.e. the fine-grained description of the causal relations underlying the macroscopic phenomena.

Importantly, in both scenarios, the use of explanatory values is already in place – in the epistemically opaque one, as inferential guides towards approximating the nomological structure, and in the epistemically transparent scenario as qualitative characteristics of this nomological structure itself. Sober and Roche mix these scenarios and they should not.[97] That is to say, they explicitly adopt the epistemically opaque scenario (no God's eye view, the conditionals in question are part of the background knowledge of the epistemic subject) but the *justification* for the use of such conditionals for increasing the resilience of posterior probabilities, if the explanationist talk is excluded, can only come from an epistemically transparent scenario.

Hence we have the following double disjunctive: if there is no God's eye view in their arguments about the background knowledge increasing the resilience of probabilities, then either the arguments in question are non-conclusive, or they are conclusive and the evidential relevance of explanations, i.e. the role of explanatory values as inferential guides, is tacitly present. If there is a God's eye view, then either their arguments are non-conclusive, or they are conclusive and the explanatory values are already accepted as objective qualitative characterisations of the nomological structure.

I would like to think that their arguments are after all conclusive. This would be a proper perspective in order to adequately consider the ways in which IBE and could fruitfully be used together.

§4 Back the Clarke *et al* criteria of mechanistic evidence

Before going back to the medical side of discussion, let me briefly take stock of the general

[97] I am not claiming one should not appeal to both these types of scenarios in argumentation; I did it myself in the previous section. However, one should distinguish between them and seek out the justifiable transitions or bridging principles from one to another.

arguments provided in this chapter. I have offered in the previous sections a way to conceptualise the collaboration between IBE and Bayesianism, which explains *how* IBE and Bayesian inference are compatible, while remaining different types of inference. I have argued that these types of inferences are *analogous*, since they are justified by the appeal to ideal scenarios and can be construed as hunting down the ultimate nomological structure. They are *defendable* insofar as, in real-life situations in which the ideal scenarios *do not hold,* their results can be seen as progressively approximating this nomological structure - the more so, the more science progresses. They are *different* insofar as one hunts down for the quantitative aspect of this nomological structure (IBE), whereas the other (the Bayesian inference) hunts down, or has its role of inferential guide underpinned by the qualitative aspect of this nomological structure. The analogy and the difference result in *complementarity* due to the fact that it is the *same* nomological structure that is hunted down (or towards which both methods of inference are guiding). Furthermore, the resulting complementarity is not vulnerable to the type of incompatibilist arguments advanced by Sober and Roche. By accepting the role of causal information in evidential support (as brought about by mechanisms), Sober and Roche already accept the use of explanatory elements in confirmation.

In discussing the distinctness and complementarity between Bayesianism and IBE, I have appealed to the analogy between Russo and Williamson's account of epistemic causation and objective Bayesianism, on the one hand, and my own justificatory account of inferential use of the explanatory values of IBE, on the other. The reason I have drawn this analogy is that their account offers an insightful illustration of the difference between the 'objective' or ideal, and the non-ideal use of key notions at work in the theory of confirmation (causation, probability, explanation, etc.). On the other hand, their account is usefully minimalistic. Epistemic causality, for instance, is a minimalist description of what causation and causal laws should mean. It is minimalistic because it is a Humean account, and it useful because, beyond its heuristic use, it could be accepted by the proponent of a more theory loaded account of causation and laws, a proponent who would have to add some extra-content to this minimalistic framework.

For instance, an anti-Humean theorist of laws *à la* Bird (2007) could add that the omniscient being of the respective ideal scenarios would well be acquainted with the existence of causal powers in the world, a necessitarian theorist *à la* Armstrong (1978) could add that the omniscient being should be aware of the relation of necessitation holding between universals, etc.). In other words, one needs not buy entirely into their account of epistemic causation (and objective Bayesianism)[98] - which is sufficiently neutral and minimalistic in order to be accepted by a wide variety of theorists with various theoretical commitments – in order to profit from the heuristic value of their

[98] See Dragulinescu (2012) which criticizes the Humean aspect of the epistemic account of causation.

arguments, involving, as I have showed, ideal and non-ideal scenarios.

Mutatis mutandis, the same can be said about Lewis' account of laws. Even theorists with anti-Humean commitments should find in Lewis' account a useful description of how laws should look like on a minimalistic perspective (even if they would demand additional modal constraints). I cannot go here into the details of a discussion that I do not wish to simplify; [99] but even when admitting that Lewis' account of laws is not universally accepted, it remains that ***if*** one accepts Lewis' account lf laws as a minimalist description of the ontological structure (and surely this account has more proponents among Bayesian theorists than the thesis of the compatibility between IBE and the Bayesian theory has) and/or ***if*** one accepts Williamson's account of objective Bayesianism (and surely this account has more proponents among Bayesians than the thesis of the compatibility between IBE and the Bayesian theory has), ***then*** one should also accept the compatibility between IBE and the Bayesian theory. It is not a triumphalist conclusion, but it makes a significant headway.

However, Russo and Williamson's account offers even more to the prospective collaboration between IBE and the Bayesian account of confirmation. For one, as noted in the previous chapter, their stand of objective Bayesianism arguably includes already one central explanatory value, namely simplicity (encoded in the requirement of equivocation, expressed in particular by the corresponding demand of epistemic causation; Russo and Williamson, 2007, p. 168). In the medical area, their RWT-based arguments that population studies are unable to rule out spurious correlations in the absence of evidence of mechanisms (Russo and Williamson, 2007, p. 159) can easily be interpreted as indicating the need to integrate the qualitative aspects of evidence in the 'inferential machine' of Bayesianism.

The previous chapter discussed one way to integrate these qualitative aspects - via the criteria of mechanistic evidence, in the sense that these criteria could well be used for an account of the resilience of medical Bayesian probabilities. But Russo and Williamson's strand of objective Bayesianism could well offer a framework to integrate the proposal that these mechanistic criteria could also constrain the priors and likelihoods the prior and/or likelihood probabilities established by the Bayesian theory taking into account the population studies evidence (which corresponds to the point *VII)* of my list of the epistemic advantages of the revised RWT). Such an integration into objective Bayesianism could only be smoothed by my argumentation in the present chapter in favour of the *objectivity* of explanatory values – an argumentation based, among others, on an analogy with the very reasoning that stands behind Russo and Williamson's joint accounts of

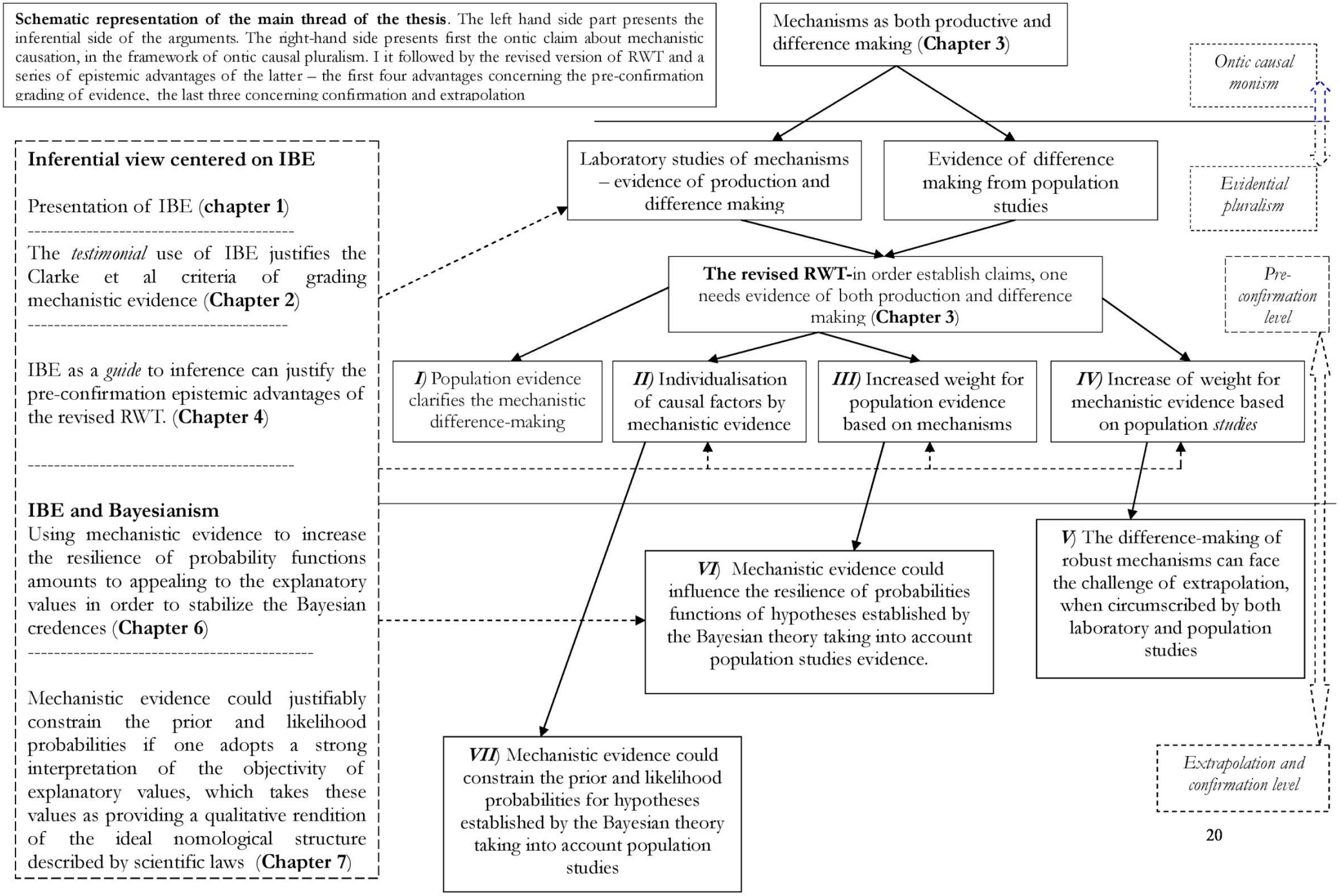
[99] However, see for instance Beebee (2000) who argues that the main difference between Humean and anti-Humean accounts of laws resides in how the issue of the 'governing' of laws is construed. The same point of view is advanced in Bird (2005).

epistemic causation/objective Bayesianism.

Conclusion chapter 7

In this chapter, I have offered way to conceptualise the collaboration between IBE and Bayesianism, which explains *how* IBE and Bayesian inference are compatible, while remaining different types of inference. I have argued that these types of inferences analogous as being justified by the appeal to ideal scenarios and as hunting down the ultimate nomological structure. They are defendable insofar as, in real-life situations in which the ideal scenarios *do not hold,* their results can be seen as progressively approximating this nomological structure - the more so, the more science progresses. They are different insofar as one hunts down for the quantitative aspect of this nomological structure, whereas the other hunts down (or has its role of inferential guide underpinned by) the qualitative aspect of this nomological structure. The analogy and the difference result in complementarity due to the fact that it is the *same* nomological structure that is hunted down (or towards which both methods of inference are guiding). The resulting complementarity is not vulnerable to the type of incompatibilist arguments advanced by Sober and Roche, and can accommodate the proposal that in medicine, the criteria of grading mechanistic evidence of Clarke *et al* could constrain, as explanatory, the prior and likelihood probabilities established by the Bayesian theory taking into account the population studies evidence.

Schematic representation of the main thread of the thesis. The left hand side part presents the inferential side of the arguments. The right-hand side presents first the ontic claim about mechanistic causation, in the framework of ontic causal pluralism. I it followed by the revised version of RWT and a series of epistemic advantages of the latter – the first four advantages concerning the pre-confirmation grading of evidence, the last three concerning confirmation and extrapolation



Conclusion thesis

The present thesis has taken its cue from the list of criteria for assessing the quality of mechanistic evidence, i.e. for grading mechanistic evidence in medicine at a pre-confirmation level, as provided in Clarke et al. (2014). Having as case studies the history of atherosclerosis and the medical treatments of hypertension and heart failure, I have sought to extend the work done by Clarke et al. in three (interconnected) directions, with the aim of

i) first, on a general level, providing an epistemological argumentation to justify these criteria for grading evidence.

i) second, putting more flesh onto the bones of these criteria (in particular on the criterion of robustness) in terms of the ontology of mechanisms and causal relations in medicine, and enquiring into how these criteria work in the context of the entire medical evidence, i.e. when taking into account also the evidence of population studies.

iii) third, setting out a plausible way in which the quality mechanistic evidence – hierarchized and graded using these criteria at a pre-confirmation stage – could make a contribution at the confirmation stage itself.

The results of my research have been that

i) I have provided an epistemic justification for the Clarke et al. criteria, using as an inferential method the Inference to the Best Explanation (IBE), which I have shown to be a useful epistemic framework for dealing with mechanistic evidence in medicine.

ii) I have defended the view that mechanisms in medicine should be viewed as entailing both production and difference-making. I have shown moreover that, on this construal of mechanisms, we can grasp a hold on the reciprocal increasing of quality between the evidence of mechanisms and the evidence of population studies, which is obtained when the two types of evidence are graded together. Finally, I have shown that, on this construal of mechanisms, one can define the robustness of mechanisms as their capacity to maintain their difference making across varying contexts, and that, in turn, this approach to robustness allows us to better conceptualise the problem of extrapolation in medicine and how mechanisms can contribute to solve it.

iii) I have argued that the quality of mechanistic evidence could make a contribution at the confirmation stage itself - in the framework of the use of IBE and Bayesianism - following the

thread of how the quality of mechanistic evidence, interpreted as evidential weight in Joyce (2015)'s sense, could supplement the balance of evidence assessed quantitatively by the Bayesian theory.

More precisely, I have proposed - along the lines of a proposal made by McCain and Poston in their (2014) - that the contribution of explanatory features to the Bayesian confirmation of medical causal claims amounts to the increase of the resilience of probability functions corresponding to population level assessments that are backed up by mechanistic evidence. I have also suggested, after adopting a strong defense of the objectivity of explanatory values, that such explanatory values employed to grade the evidence of medical mechanisms, could subsequently be used to constrain the priors and/or likelihoods of the Bayesian confirmation stage.

The importance of all the above research lies I believe in i) bringing to the fore the role played by the qualitative aspect of evidence when assessing medical claims, ii) drawing attention to the close relationship between the ontic end epistemic approaches to medical mechanisms and iii) pressing for a fruitful utilisation of both IBE and Bayesianism, which would integrate the qualitative aspect of evidence and a sharper look at the ontology of mechanisms.

These points can lend themselves to further research. I would hope, in particular, that the approach to mechanistic robustness in terms of different making could be useful for the current research into systems medicine, and that the discussion of the compatibility between IBE and Bayesianism could spur interest in integrating the Clarke et al. criteria of mechanistic evidence into the objective Bayesian evaluation of medical claims and hypotheses.

References

Achinstein, P. 2001. *The Book of Evidence*. Oxford: Oxford University Press.

Adler, J. 2012. Epistemological Problems of Testimony. *Stanford Encyclopedia of Philosophy*, *ed.* Edward N. Zalta. http://plato.stanford.edu/entries/testimony-episprob/ (last accessed January 29, 2018)

Audi, R. 1997. The Place of Testimony in the Fabric of Knowledge and Justification. *American Philosophical Quarterly* 34: 405–22.

Anscombe. G.E.M. 2001. *Causality and Determination: An Inaugural Lecture*. Cambridge: Cambridge University Press.

Bechtel, W. and Abrahamsen, A. 2005. Explanation: A Mechanist Alternative, *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*. 36(2): 421—41.

Bechtel, W. and Abrahamsen, A. 2011. Complex biological mechanisms: Cyclic, oscillatory, and autonomous. In C. A. Hooker (Ed.), *Philosophy of complex systems. Handbook of the philosophy of science*, Volume 10. New York: Elsevier.

Berkovitz, J. 2007. Action at a Distance in Quantum Mechanics. *Stanford Encyclopedia of Philosophy*, *ed.* Edward N. Zalta. http://plato.stanford.edu/entries/qm-action-distance/ (last accessed January 29, 2018)

Bird, A. 2005. Laws and Essences. *Ratio* 18 (4): 437–461

Bird, A. 2005. Laws and essences. *Ratio* 18(4): 437–461.

Bird, A. 2007. Inference to the Only Explanation. *Philosophy and Phenomenological Research* 74: 424–32.

Bird, A. 2007b. *Nature's Metaphysics: Laws and Properties*. Oxford: Oxford University Press.

Bird, A. 2010. Eliminative abduction—examples from medicine. *Studies in History and Philosophy of Science* 4: 345-352.

Bird, A. 2011. The epistemological function of Hill's criteria. *Preventive Medicine* 53: 85-96.

Braunwald, E. and C. A. Chidsey. 1965. The adrenergic nervous system in the control of the normal and failing heart. *Proc R Soc. Med.* 58(12): 1063–1066.

Broadbent, A. 2007. Reversing the counterfactual analysis of causation. *International Journal of Philosophical Studies* 15 (2):169 – 189.

Bovens, L. and S. Hartmann. 2003. *Bayesian Epistemology*. Oxford: Clarendon Press.

Broadbent, A. 2011. Inferring causation in epidemiology: Mechanisms, black boxes, and contrasts. In P.

McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences*. Oxford: Oxford University Press.

Bohm, D.1980.*Wholeness and the Implicate Order*. New York: Routledge.

Bohm, D., and Hiley, B. J., 1993. *The Undivided Universe: An Ontological Interpretation of Quantum Theory*, London: Routledge & Kegan Paul.

Brown, S., and J.L. Goldstein. 2009. History of discovery: The LDL receptor. *Arteriosclerosis, Thrombosis, and Vascular Biology* 29(4): 431-438.

Jimenez-Buedo, M. 2011. Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology*. 3 (18)

Campaner, R. 2011. Understanding mechanisms in the health sciences. *Theoretical Medicine and Bioethics* 32 (1):5-17

Campaner, R, Galavotti. 2012. M. Evidence and the Assessment of Causal Relations in the Health Sciences. *International Studies in the Philosophy of Science. 26(1):27-45*.

Cartwright, R. 1983. How the Laws of Physics Lie. Oxford: Oxford University Press.

Cartwright, N. 1989. *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

Cartwright, N. 1999. *The Dappled World*. Cambridge: Cambridge University Press.

Cartwright, N. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Cartwright, N., & Munro, E. 2010. The limitations of RCTs in predicting effectiveness. *Journal of Experimental Child Psychology, 16*(2), 260–266.

Casini, L, McKay Illari, P., Russo, F., Williamson, J. 2011. *Theoria* 26(1):495-4548.

Casini, L., McKay Illari, P., Russo, F., & Williamson, J. 2011. Recursive Bayesian nets for prediction, explanation and control in cancer science. *Theoria, 26*(1), 495–4548.

Casini, L. 2012. Causation: Many Words, One Thing? *Theoria*, 27, pp. 203–19

Cassini, L. 2015. Can Interventions Rescue Glennan's Mechanistic Account of Causality?. *British Journal for the Philosophy of Science*. Advanced access, published online March 2015.

Chowdhury, R., S. Warnakula, S. Kunutsor, F. Crowe, H. Ward, L. Johnson. 2014. Association of dietary, circulating, and supplement fatty acids with coronary risk: A systematic review and meta-analysis. *Annals of Internal Medicine* 160: 398-406.

Chrysant GS. Bakir S. Oparil S. 1999. Dietary salt reduction in hypertension--what is the evidence and why is it still controversial? *Prog Cardiovasc Dis.* 42(1):23-38.

Clarke, B., Gillies, D., Illari, P., Russo, F. and Williamson, J. 2014. Mechanisms and the Evidence Hierarchy. *Topoi,* 33 (2): 339-360.

Clarke, B. Leuridan B., Williamson, J. 2014***b***. Modelling mechanisms with causal cycles, *Synthese* 191(8): 1651-1681.

Coady, C. A. J. 1992. *Testimony*. Oxford: Oxford University Press.

Cohen, J. 1986. Twelve questions about Keynes's concept of weight. *British Journal for the Philosophy of Science* 37(3): 263-

278.
Craver, C. Tabery, J. 2015. Mechanisms in Science. *Stanford Encyclopedia of Philosophy*, *ed.* Edward N. Zalta. http://plato.stanford.edu/entries/science-mechanisms/#ProUndMai. (last accessed January 29, 2018)
Davies, M. 2004. Epistemic Entitlement, Warrant Transmission, and Easy Knowledge. *Aristotelian Society Supplementary* 78: 213-45
Darby, G. Williamson, J. 2011. Imaging technology and the philosophy of causality. *Philosophy & Technology*, 24(2):115–136.
Darsee, J. 1983. A Retraction of Two Papers on Cardiomyopathy. *N Engl J Med* 308:1419.
Debru, A. 1996. *Le corps respirant: La pensée physiologique chez Galien*. Leiden: Brill.
Devitt, M. 2010. *Putting Metaphysics First*. Oxford: Oxford University Press.
de Vreese, L. 2006. Causal pluralism and scientific knowledge: an underexposed problem. *Philosophica, 77*, 125–150.
Douven, I. 2014. Abduction. In *Stanford Encyclopedia of Philosophy*, ed. E. Zalta.
www. plato.stanford.edu/entries/abduction/
Douven, I., and S. Wenmackers. 2015. Inference to the Best Explanation versus Bayes's Rule in a Social Setting. *British Journal for the Philosophy of Science*, Online first.
Douven, I., and J. Schupbach. 2015. Probabilistic alternatives to Bayesianism: the case of explanationism. *Front Psychol* 6: 459.
Douven, I., and J. Schupbach. 2015b. The role of explanatory considerations in updating. *Cognition* 142:299-311.
Douven, I. 2016. Inference to the Best Explanation: What Is It? And Why Should We Care? In *Best Explanations: New Essays on Inference to the Best Explanation*. K. McCain and T. Poston (eds.), Oxford: Oxford University Press.
DuBroff, R., and M. de Lorgeril. 2015. Cholesterol confusion and statin controversy. *World Journal of Cardiology* 26(7): 404–409.
Dragulinescu, S. 2012. On 'Stabilising' medical mechanisms, truth-makers and epistemic causality: a critique to Williamson and Russo's approach. *Synthese 187*(2):785–800.
Earman, J. 1992. *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge: MIT Press.
Epstein, S. Robinson, B.F. Kahler, R.L. and E. Braunwald. 1965. Effects of beta-adrenergic blockade on the cardiac response to maximal and submaximal exercise in man. *J Clin Invest.* 44(11): 1745–1753.
Feduzi, A. 2010. On Keynes's conception of the weight of evidence. *Journal of Economic Behavior & Organization* 76(2): 338-351.
Feldman, MD. Copelas, L. Gwathmey, JK. Phillips, P. Warren, SE. Schoen, F., Grossman, W. Morgan, JP. 1987. Deficient production of cyclic AMP: pharmacologic evidence of an important cause of contractile dysfunction in patients with end-stage heart failure. *Circulation* 75: 331.
Franco V., Oparil S. 2006. Salt sensitivity, a determinant of blood pressure, cardiovascular disease and survival. *J Am Coll Nutr.* 25(3 Suppl):247S-255S.
Furie, M., and R. Mitchell. 2012. Plaque attack: One hundred years of atherosclerosis. *American Journal of Pathology* 180(6): 2184-2187.
Gabbay, D. and J. Woods. 2005. *The Reach of Abduction*. Amsterdam: North Holland.
Gaffney, T. and E. Braunwald. 1963. Importance of the adrenergic nervous system in the support of circulatory function in patients with congestive heart failure. *Am J Med.* 34:320-4.
Gelfert, A. 2010. Reconsidering the role of inference to the best explanation in the epistemology of testimony. *Studies in History and Philosophy of Science* Part A 41 (4): 386-396.
Glass D. 2007. Coherence measures and inference to the best explanation. *Synthese* 157: 275-296.
Glass, D. 2012. Inference to the best explanation: does it track truth? *Synthese* 185: 411-427.
Glennan, S. 1996. Mechanisms and the Nature of Causation. *Erkenntnis, 44,* 49–71.
Glennan, S. 2002. Rethinking Mechanistic Explanation. *Philosophy of Science*. 69(S3): 342—53.
Glennan, S. 2005.Modeling Mechanisms. *Studies in History and Philosophy of Science Part C 36* (2):443-464.
Glynn, L. 2013 Causal foundationalism, physical causation, and difference-making *Synthese* 190:1017–1037
Godfrey-Smith, P. 2008. Causal pluralism. In *The Oxford handbook of causation,* Beebee, H., Hitchcock, C., and Menzies, P., *eds*. Oxford University Press, Oxford.
Good, I. 1985. *Weight of evidence: A brief survey*. http://www.swrcb.ca.gov/water_issues/programs/tmdl/docs/303d_policydocs/207.pdf. (last accessed January 29, 2018)
Graham, P. J. 2006. Testimonial Justification: Inferential or Non-inferential. *Philosophical Quarterly*, 56: 84–95.
Greco, J., 2012. Recent work on Testimonial Knowledge. *American Philosophical Quarterly,* 49: 15–28.
Guala, F. 2010. Extrapolation, Analogy and Comparative Process Tracing. *Philosophy of Science* 77 (5): 1070-1082.
Hall, N. 2004. Two Concepts of Causation. In J. Collins, N. Hall, and L. A. Paul, eds., *Causation and Counterfactuals*. pp. 181-204. Massachusetts: The M. I. T. Press
Hall, N. 2012. Comments on Strevens' Depth. *Philosophy and Phenomenological Research* 84 (2):474-482
Harman, G. 1965. The Inference to the Best Explanation. *The Philosophical Review* 74: 88-95.
Hempel, C. G. and P. Oppenheim, 1948. Studies in the Logic of Explanation. *Philosophy of Science* 15: 135-175.
Hempel, C. G. 1970. On the 'Standard Conception' of Scientific Theories. In M. Radner and S. Winokur (eds.), 142–163. *Minnesota Studies in the Philosophy of Science,* Vol. IV, Minneapolis, MN: University of Minnesota Press.

Hernandez et.al. 2009. Clinical Effectiveness of Beta-Blockers in Heart Failure. *J Am Coll Cardiol.* 53(2): 184-192.
Hintikka, J. 1998. What is abduction? The fundamental problem of contemporary epistemology. *Transactions of the Charles S. Pierce Society,* 34: 503-533.
Howick, J. 2011. Exposing the Vanities—and a Qualified Defense—of Mechanistic Reasoning in Health Care Decision Making. *Philosophy of Science* 78 (5):926-940.
Howick, J. Glasziou, P. Aronson, K. 2013. Problems with Using Mechanisms to Solve the Problem of Extrapolation. *Theoretical Medicine and Bioethics* 34 (4):275-291.
Hume, D. 1977 [1748]. *An Enquiry Concerning Human Understanding*. Eric Steinberg (ed.). Indianapolis: Hackett Publishing Company.
Illari, P. 2011. Mechanistic Evidence: Disambiguating the Russo–Williamson Thesis. *International Studies in the Philosophy of Science* 25 (2):139 – 157.
Illari, P. and Williamson, J. 2012. What is a mechanism: thinking about mechanisms across the sciences, *European Journal for Philosophy of Science* 2:119-135.
Iranzo, V. 2008. Bayesianism and inference to the best explanation. *Theoria* 23 (1): 89-106
Jimenez-Buedo, M. 2011. Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology.* 3 (18): 271-282.
Joffe, M. 2013. The Concept of Causation in Biology. *Erkenntnis* 78 (2):179-197.
Joyce, J. 2005. How probabilities reflect evidence. *Philosophical Perspectives* 19(1): 153-178.
Jouanna, J. 1992. *Hippocrate.* Paris: Fayard
Kelly, T. 2008. Evidence: Fundamental concepts and the phenomenal conception. *Philosophy Compass* 3(5): 933-955.
Kelly, T. 2014. Evidence. In *The Stanford Encyclopedia of Philosophy*, *ed.* Edward N. Zalta. http://plato.stanford.edu/entries/evidence/. (last accessed January 29, 2018)
Keynes, M. 1921. *A treatise on probability*. London: Macmillan & Co.
Khan, N. McAlister, F.A. 2006. Re-examining the efficacy of β-blockers for the treatment of hypertension: a meta-analysis. *CMAJ* 174: 1737–1742.
Kline, D. and Matheson, C. 1986.How the Laws of Physics Don't Even Fib. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association. Volume One: Contributed Papers, pp. 33-41.
Kritchevsky, D. 1995. Dietary protein, cholesterol and atherosclerosis: A review of the early history. *Journal of Nutrition* 125: 589S-593S.
Lackey, J. 1999. Testimonial knowledge and transmission. *The Philosophical Quarterly* 49: 471–490.
Lackey, J. 2003. Non-reductionism in the Epistemology of Testimony. *Nous* 37: 706–735.
Lackey, J. 2006. It Takes Two to Tango: Beyond Reductionism and Non-Reductionism in the Epistemology of Testimony. In Jennifer Lackey & Ernest Sosa (eds.), *The Epistemology of Testimony*. Oxford University Press.
Lackey, J. 2008. *Learning from Words*. Oxford: Oxford University Press.
Lee, T. 2013. *Eugene Braunwald and the Rise of Modern Medicine*. Harvard University Press
Landau, M. 2012. An Interview with Dr. Eugene Braunwald. *Clinical Chemistry* 58 (1): 11-20.
Lipton, P. 1993. Making a Difference. *Philosophica* 51: 39-54.
Lipton, P. 1994. Truth, existence, and the best explanation. In A. Derksen (ed.), *The Scientific Realism of Rom Harré*. Tilburg University Press
Lipton, P. 1998. The Epistemology of Testimony. *Studies in History and Philosophy of Science*. 29:1-31.
Lipton, P. 2001. Is Explanation a Guide to Inference? A reply to Wesley C. Salmon. In *Explanation, Theoretical Approaches and Applications,* ed. G. Hon and S. Rakover, 93-120. Dordrecht: Springer.
Lipton. P. 2004. *Inference to the Best Explanation* (second edition). London: Routledge.
Lipton, P. 2007. Alien abduction: Inference to the best explanation and the management of testimony *Episteme* 4 (3):238-251.
Lewis, D. 1973. Causation. *Journal of Philosophy* 70: 556-67.
Lewis, D. 1986. Postscript to 'Causation'. *Philosophical Papers*, Volume 2. New York: Oxford University Press, 172-213.
Lewis, D. 2000. Causation as Influence. *Journal of Philosophy* 97: 182-97.
Longworth, F. 2006. Causation, Pluralism and Responsibility. *Philosophica* 77 pp. 45-68.
Machamer, P., Darden, L. and Craver, C.F. 2000. Thinking about Mechanisms, *Philosophy of Science*. 67(1): 1—25.
Mann, S. 2012. *Hypertension and You.* Rowman & Littlefield.
Mann, D. and M. Bristow. 2005. Mechanisms and Models in Heart Failure. The Biomechanical Model and Beyond. Circulation 111 (21): 2837-2849.
Malmgren, A-S., 2006. Is There A Priori Knowledge By Testimony. *The Philosophical Review* 115: 199–241.
McCain, K. and T, Poston. Why Explanatoriness is Evidentially Relevant. *Thought* 3(2): 145-53.
McMullin, E., 1992. *The Inference that Makes Science.* Milwaukee WI: Marquette University Press.
Mcauliffe, W. 2015. How did abduction get confused with inference to the best explanation? *Transactions of the Charles S. Pierce Society* 51: 300-319.
Menzies, P. 2014. Counterfactual Theories of Causation. In *Stanford Encyclopedia of Philosophy*, *ed.* Edward N. Zalta. http://plato.stanford.edu/entries/causation-counterfactual/ (last accessed January 29, 2018)
Messerli, F., Grossman, E., & Goldbourt, U. (1998). Are beta-blockers efficacious as first-line therapy for hypertension in the elderly? A systematic review. *JAMA, 279*, 1903–1907.

Messerli F, Grossman E, Goldbourt U. 1998. Are beta-blockers efficacious as first-line therapy for hypertension in the elderly? A systematic review. *JAMA* 279:1903–7.
Mill, J., S. [1843] 2002. *A System of Logic*. Honolulu: University Press of the Pacific.
Minnameier, G. 2004. Pierce-suit of truth – Why inference to the best explanation and abduction ought not to be confused. *Erkenntnis*, 60:75-105.
Niiniluoto, I. 1999. Defending abduction. *Philosophy of Science* 66: S436-S451.
Niiniluoto, I. 2007. Abduction and Scientific Realism. In Ferda Keskin (ed.), *The Proceedings of the Twenty-First World Congress of Philosophy, vol. 12: Philosophical Trends in the XXth Century*, pp. 137-142, Philosophical Society of Turkey, Ankara.
Nutton, V. 2013. *Ancient medicin*e. London: Routledge.
O'Donnell, R. 1992. Keynes's weight of argument and Popper's paradox of ideal evidence. *Philosophy of Science* 59(1): 44-52.
Okasha, S. 2000. Van Fraassen's Critique of Inference to the Best Explanation. *Stud. Hist. Phil. Sci.*, 31 (4): 691–710
Parkkinen, V-P., Strand, A. 2015. Causation in evidence-based medicine: in reply to Kerry *et al.Journal of Evaluation in Clinical Practice*, online first.
Perrine, T. 2004. In Defense of Non-Reductionism in the Epistemology of Testimony. *Synthese* 191 (14):3227-3237.
Pippin, R. B. 1979. Negation and not-being in Wittgenstein's tractatus and Plato's sophist. *Kant-Studien, 70*(1–4), 179–196.
Pizzi, C. 2013. Counterfactuals and modus tollens in abductive arguments, *Logic Journal of the IGPL*, Volume 21, Issue 6, pp. 962–979
Psaty *et al.* 1995. The risk of myocardial infarction associated with antihypertensive drug therapies. JAMA. 274(8):620-5.
Psillos. S. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.
Psillos, S. 2000. Abduction: Between Conceptual Richness and Computational Complexity'. In *Abduction and Induction: Essays in their Relation and Integration*, ed. A. C. Kakas and P. Flach, 59-74. Dordrecht: Kluwer.
Psillos, S. 2002. Simply the Best: A Case for Abduction. In *Computational Logic: From Logic Programming into the Future*, ed. A. C. Kakas and F. Sadri, 605-625. Dordrecht: Springer.
Psillos, S. 2004. A Glimpse of the Secret Connexion: Harmonizing Mechanisms with Counterfactuals. *Perspectives on Science* 12 (3):288-319.
Psillos, S. 2007. The Fine Structure of Inference to the Best Explanation. *Philosophy and Phenomenological Research* 74: 441-448.
Ravnskov, U. 1992. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ*. Jul 4; 305(6844):15-9.
Ravnskov, U. 2006. The International Network of Cholesterol Skeptics (THINCS) http://www.ravnskov.nu/references/ (last accessed January 29, 2018)
Ravnskov, U. 2013. High cholesterol may protect against infections and atherosclerosis. *QJM* 96(12): 927-934.
Reiss, J. 2007. Do We Need Mechanisms in the Social Sciences? *Philosophy of the Social Sciences* 37(2), 163-184.
Reiss, J. 2009. Causation in the Social Sciences: Evidence, Inference, Purpose. P*hilosophy of the Social Sciences* 39(1): 20-40.
Reiss, J. 2010. Review of Across the Boundaries: Extrapolation in Biology and Social Science, Daniel P. Steel. Oxford University Press, 2007. *Economics and Philosophy* 26(03): 382-90.
Relman, A. 1983. Lessons from the Darsee affair. *The New England Journal of Medicine* 308 (23): 1415–7.
Roche, W. and Sober, E. 2013. Explanatoriness is evidentially irrelevant, or inference to the best explanation meets bayesian confirmation theory. *Analysis* 73: 659– 68.
Romeijn, J-M. 2013. Abducted by Bayesians? Journal of Applied Logic. 11 (4): 430-439.
Russo, F. Williamson, J. 2007. Interpreting causality in the health sciences. *International Studies in the Philosophy of Science* 21(2): 157-170.
Russo, F. Williamson, J. 2011. Epistemic causality and evidence-based medicine. *History and Philosophy of the Life Sciences* 33(4):563-582.
Ryle, G. 1966. *Plato's progress*. Cambridge: Cambridge University Press.
Salmon, W. 2001. Reflections of a bashful Bayesian: A reply to Peter Lipton. In *Explanation, Theoretical Approaches and Applications,* ed. G. Hon and S. Rakover, 121-136. Dordrecht: Springer.
Saunders E. 1988. Drug treatment considerations for the hypertensive black patient. *J Fam Pract*. Jun; 26(6):659-64.
Schurz, G. 2008. Patterns of Abduction. *Synthese*. 164 (2):201-234.
Shapin, S. 1994. *A Social History of Truth*. Chicago: University of Chicago Press.
Skipper, R. A., & Millstein, R. L. (2005). Thinking about evolutionary mechanisms: natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 327–347.
Skyrms, B. 1977. Resiliency, probability, and causal necessity. *Journal of Philosophy* 74: 704-713.
Starko, K. M. 2009. Salicylates and pandemic influenza mortality, 1918–1919 pharmacology, pathology, and historic evidence. *clinical Infectious Diseases, 49*(9), 1405.
Stehbens WE (2001).Coronary heart disease, hypercholesterolemia, and atherosclerosis I. False premises. *Exp Mol Pathol.* 70 (2): 103–119.
Steel, D. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford University Press.
Steinberg, D. 2004. An interpretive history of the cholesterol controversy: part I. *Journal of Lipid Research, 45*: 1583–1593.

Steinberg, D. 2005a. An interpretive history of the cholesterol controversy, part II: the early evidence linking hypercholesterolemia to coronary disease in humans. *Journal of Lipid Research, 46*: 179–190.
Steinberg, D. 2005b. An interpretive history of the cholesterol controversy, part III: mechanistically defining the role of hyperlipidemia. *Journal of Lipid Research, 46*, 2037–2051.
Steinberg. D. 2007. *The cholesterol wars*. New York: Academic Press.
Strevens, M. 2004. The Causal and Unification Accounts of Explanation Unified – Causally. *Noûs* 38:154–179.
Strevens, M. 2007. Mackie Remixed. In *Causation and Explanation, Topics in Contemporary Philosophy*, J. Keim Campbell, M. O'Rourke, and H. S. Silverstein (eds.), vol. 4, Cambridge: MIT Press.
Strevens, M. 2011. *Depth: An account of scientific explanation*. Cambridge: Harvard University Press.
Strevens, M. 2012a. Precis of Depth. *Philosophy and Phenomenological Research* 84,447–505
Strevens, M. 2012b. Replies to Weatherson, Hall, and Lange. *Philosophy and Phenomenological Research* 84, 447–505
Strevens, M. 2013. Causality reunified. *Erkenntnis, 78*(2), 299–320.
Sutter, M. 1994. Blood cholesterol is not causally related to atherosclerosis. *Cardiovascular Research* 28: 575.
Swedberg, K. 1993. Initial Experience with Beta Blockers in Dilated Cardiomyopathy. *Am J Cardiol.* 71(9):30C-38C
Swedberg, K. 2009. b-Blockers in worsening heart failure: good or bad? *European Heart Journal* 30: 2177–2179.
Thagard, P. 1978. The Best Explanation : Criteria for Theory Choice. *Journal of Philosophy* 75 (2): 76-92.
Thagard, P. 2005. Testimony, Credibility and Explanatory Coherence. *Erkenntnis*, 63 (3): 295-316.
Vandenbroucke, J. 1996. Evidence-Based Medicine and "Medecine d'Observation. *J Clin Epidemiol.* 49 (12): 1335-1338.
Van Fraassen, B. C. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.
Waagstein F, Hjalmarson A, Varnauskas E, Wallentin I. 1975. Effect of chronic beta-adrenergic receptor blockade in congestive cardiomyopathy. *Br Heart J* 37(10):1022-36.
Waagstein, F. 2002. Beta-Blockers in Congestive Heart Failure: the Evolution of a New Treatment Concept – Mechanisms of Action and Clinical Implications, *J Clin Basic Cardiol* 5 (3): 215-223.
Waters, K. 2007 Causes that make a difference, *The Journal of Philosophy*, 104 (11): 551-579
Weisberg, J. 2009. Locating IBE in the Bayesian Framework. *Synthese* 167:125-143.
Weir MR, Chrysant SG, McCarron DA, *et al.* 1998. Influence of race and dietary salt on the antihypertensive efficacy of an angiotensin-converting enzyme inhibitor or a calcium channel antagonist in salt-sensitive hypertensives. *Hypertension* 31:1088–96.
Wilde, M and J. Williamson. 2016. Evidence and Epistemic Causality. In *Statistics and Causality: Methods for Applied Empirical Research*, Wolfgang Wiedermann and Alexander von Eye *eds.,* Wiley.
Williamson, J. 2006. Causal pluralism versus epistemic causality, *Philosophica* 77: 69-96.
Williamson, J. 2011. Mechanistic Theories of Causality. *Philosophy Compass* 6 (6): 421–432.
Wilmshurst, P. 2007. Dishonesty in Medical Research. *Med Leg J.* 75(1): 3-12.
Witztum, J., and D. Steinberg. 2010. History of discovery oxidized low-density lipoprotein and atherosclerosis. *Arteriosclerosis, Thrombosis, and Vascular Biology* 30: 2311-2316.
Woodward, J. 2002. What is a Mechanism? A Counterfactual Account. *Philosophy of Science* 69(S3): S366—77.
Worrall, J. 2007. Why there's no cause to randomize. *British Journal for the Philosophy of Science, 58*: 451–488.
Worrall, J. 2008. Evidence and ethics and medicine. *Perspectives in Biology and Medicine, 51*: 418–431.
Worall, J. 2010. Evidence: Philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice* 16(2): 356-362.
Yablo, S. 2002. De facto dependence. *Journal of Philosophy*, 99: 130–148.