Inference to the Best Explanation in the Catch-22: How much autonomy for Mill's method of difference?

Raphael Scholl*

University of Bern
History and Philosophy of Science
Institute of Philosophy
Sidlerstr. 5
CH-3012 Bern
Switzerland

Draft of September 23, 2014

Forthcoming in the *European Journal for Philosophy of Science*.

Acknowledgements

I benefited from discussing an early version of this material at the "Evidence and Explanation" workshop organized by the Episteme Group at the University of Geneva in April 2012. I thank Adrian Wüthrich, Tim Räz, Michael Baumgartner, the members of the Lake Geneva Biology Interest Group (LG-BIG) and several anonymous referees for helpful comments on earlier drafts of the paper.

*e-mail: raphael.scholl@gmail.com

Abstract

In his seminal *Inference to the Best Explanation*, Peter Lipton adopted a causal view of explanation and a broadly Millian view of how causal knowledge is obtained. This made his account vulnerable to critics who charged that Inference to the Best Explanation is merely a dressed-up version of Mill's methods, which in the critics' view do the real inductive work. Lipton advanced two arguments to protect Inference to the Best Explanation against this line of criticism: the problem of multiple differences and the problem of inferred differences. Lipton claimed that these two problems show Mill's method of difference to be largely unworkable unless it is embedded in an explanationist framework. Here I consider both arguments as well as the best Millian defense against them. Since the existing Millian defense is only partially successful, I will develop a new and improved account. As an integral part of the argument, I show that my solutions to the problems of multiple and inferred differences offer new insight into Lipton's main case study: Ignaz Semmelweis's discovery of the cause of childbed fever. I conclude that the method of difference can overcome Lipton's challenges outside an explanationist framework.

Keywords: Inference to the Best Explanation – Mill's methods – causal inference – Semmelweis – catch-22 – multiple differences – inferred differences – integrated history and philosophy of science

1 Introduction

Mill's methods are among the most attractive methodological proposals in philosophy of science. In particular, the method of difference captures the intuitive notion of "varying one thing at a time while keeping everything else constant" to determine causal roles. The two most prominent modern accounts of confirmation – Inference to the Best Explanation (IBE) and Bayesianism – do not have their roots in Mill's methodology, but IBE nevertheless has a marked Millian character in Peter Lipton's influential formulation: Mill's method of difference is the core of Lipton's account of causal explanation.[1]

However, Lipton recognized that IBE and the method of difference live uneasily together. On the one hand, Lipton wished to make the Millian component of IBE strong, both because the method of difference is a promising description

[1] Lipton (1991, 2004). For Lipton's account of causal explanation, see chapter 3 in Lipton (2004) and especially pp. 41–54.

of much actual scientific practice and because it is already accepted as relevant by many. On the other hand, Lipton needed to avoid aligning IBE too closely with Mill's method. Otherwise, Millians could charge that the real inductive work in IBE is done by the method of difference.[2]

Lipton described the problem as an instance of a "catch-22", where success at one stage of a task precludes success at the next. Lipton's argument for IBE has three stages: identification, matching and guiding. In the identification step, explanatory and confirmatory virtues are identified. In the matching step, it is shown that these virtues are the same, that is, that there is a match between the characteristics of highly explanatory and of well-confirmed hypotheses. Finally, in the guiding step it is shown that we use explanatory power to guide our judgments about the likeliness of hypotheses. However, the matching and the guiding tasks constitute a catch-22. Lipton writes:

> Now I will either convince you that Inference to the Best Explanation is roughly co-extensive with your account [of confirmation] or I will fail in this. If I fail, you will not buy the matching claim; but if I succeed, you will not buy the guiding claim, since you will maintain that it is your account that describes what is doing the real inferential work, without any appeal to explanatory virtues. So either way I lose: that is the catch-22. (2004, p. 125)

By forcefully arguing that judgments of explanatory power are compatible with and even rely on the method of difference, Lipton gives the Millian the ammunition she needs to claim that all the inductive work is *actually* done by the method of difference. The more successful Lipton is at supporting the matching claim, the more forcefully will his opponent reject the guiding claim.

In order to undercut the catch-22, Lipton argues that although much of our inferential practice looks Millian on the surface, the method of difference on its own cannot do substantial inferential work. In his view, the method's scope and usefulness is limited by two problems: the problem of multiple differences and the problem of inferred differences. He argues that the way to overcome these problems is to embed the method of difference in an IBE framework.

The present paper is a defense of the method of difference against Lipton's challenges. I will argue that the Millian has the resources to solve Lipton's two

[2]For the IBE vs. Mill argument, see Rappaport (1996); for a Millian analysis of Lipton's main case study of IBE, see Scholl (2013).

problems, and that the method of difference is more widely and more easily applicable than Lipton thought.

However, the debate about Lipton's challenges to Mill's methods has broader significance. With Lipton, I take the general goal of the history and philosophy of science to be a descriptively adequate and philosophically insightful account of past and present science. We wish to know how science worked and why it was successful (or sometimes failed to be). The currently prominent causal philosophies of science – mainly the mechanistic (Glennan, 1996; Machamer et al., 2000) and the interventionist (Woodward, 2003) – promise interesting answers to both questions, especially where the life sciences are concerned. But whether these approaches can be widely (especially historically) applied and justified should be a key area of interest for the history and philosophy of science community: How much of past science can we understand as the search for causes and underlying mechanisms? How precisely did this work? How did scientists (like Semmelweis, who was Lipton's main case study) gain knowledge of causes and mechanisms? In this context, it becomes a key question whether Mill's methods can do serious work – or whether (as Lipton claims) a proper understanding of past inferential practices requires the explanationist view. An analysis of the power of Mill's methods is moreover interesting because these methods can do historical work free of the charge of anachronism (see also Scholl, 2013): While they may not be the most powerful methods for causal inference *currently* on offer, it is at least plausible that past scientists (like Semmelweis) consciously applied *something like* the methods described by Mill in 1843 to investigate causal relationships.

In section 2, I will introduce Mill's method of difference and sketch Lipton's problems of multiple and inferred differences. I will then discuss the problems in turn in sections 3 and 4. In each case, I will discuss a previous Millian defense by Steven Rappaport (1996). I will argue that Rappaport's solutions of the problems of multiple and inferred differences are inadequate in multiple respects, and I will then develop improved solutions. In the spirit of integrated history and philosophy of science, I will show that my solutions offer new insight into Lipton's own main case study: Semmelweis's discovery of the cause of childbed fever. My improved Millian account will make intelligible previously disregarded aspects of the Semmelweis case.

Before proceeding, two caveats are in order. First, the following discussion is not intended as an attempt to reduce IBE fully to the method of difference. There are likely "hard cases" where the method is not applicable, and where perhaps some

type of IBE is necessary to describe and justify our inductive inferences. The goal here is a rather more modest Millian defense: It is to show that in experimental studies the method of difference on its own can do substantial inferential work independently of an IBE framework, and that it does so in Lipton's own main case study. How much of our inductive practice can be understood purely in terms of the method of difference is a question for another time.

Second, I follow Lipton (2004, p. 3) in not adopting any particular account of causation as the basis for my arguments: My hope is that my epistemological claims will be compatible with whatever the best analysis of causation turns out to be (whether it be probabilistic, mechanistic, interventionist, or something else).[3]

2 The problems of multiple and inferred differences

To understand Lipton's challenges, we begin with Mill's statement of the method of difference, his "second canon":

> 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. (1843, III.VIII.§2)

Mill's terminology is initially forbidding. However, it serves an epistemic purpose: Mill needs vocabulary to distinguish between established causes and effects and merely suspected causes and effects ("circumstances" or "phenomena"). By "instances" Mill refers to occurrences and non-occurrences of the effect under investigation which can be individuated in some way so that regularities among them can be ascertained.

In order for Mill's definition to do any work, we must assume that the system we are investigating is causal (no events without causes) and deterministic (same causes, same effects). If the first is violated, we cannot infer a causal role from

[3] As far as modeling our causal intuitions goes, I think a regularity theoretic account based on John Mackie's notion of *INUS*-conditions (Mackie, 1980) has much to recommend itself (for recent defenses, see Graßhoff and May, 2001; Baumgartner, 2008). While my arguments about multiple and inferred differences do not depend on the regularity theoretic account, they are influenced by suggestions in the textbook by Baumgartner and Graßhoff (2004), which defends such an account. Particularly relevant to my arguments concerning multiple and inferred differences are chapters 10 and 11, respectively. The textbook is regrettably not available in English.

a single antecedent difference; if the second is violated, no comparison between instances is possible.[4]

In order for us to infer causal relationships, Mill's definition demands of the instances we are comparing that they fulfill two requirements, the second of which is quite strict:

1. The phenomenon occurs in one but not the other instance, and

2. the circumstances of the two instances differ with respect to only one condition.

If both requirements are met, the method of difference sanctions one of two inferences: Either the sole difference is the cause or part of the cause of the phenomenon, or the phenomenon is the cause or part of the cause of the sole difference.

In Mill's conception, the method of difference is the method of experiment, in which case our own intervention brings about the difference. For him this settles the question of the direction of the arrow of causation, and so only the second inference remains: The sole difference is the cause or part of the cause of the phenomenon. In order to circumnavigate this issue in cases where it is not relevant, many writers (including Mill and Lipton) refer to sole *antecedent* differences between two instances. I will follow their example.

Lipton's arguments for the limitations of the method of difference both concern the strict requirement 2: that the antecedents of the instances we are comparing differ with respect to only one condition. First, how can we ever be in a position to know that there exists only one difference between the antecedent conditions of the instances we are comparing? Isn't it always possible for an additional (perhaps unobserved) difference between the antecedents to exist, which would invalidate our inference from sole difference to causal role? This is what Lipton terms the problem of multiple differences.

Second, Mill's phrasing requires that we already know about the antecedent conditions whose causal roles are to be inferred. This limits the applicability of the method, since we will sometimes wish to infer the causal role of antecedent conditions which are either unobservable or unobserved. This is what Lipton terms the problem of inferred differences.

[4] I here follow Baumgartner and Graßhoff (2004), p. 68. But see also Mill (1843), III.V. And see Hofmann and Baumgartner (2011) for a recent in-depth discussion of the logic underlying the method of difference.

3 The problem of multiple differences

According to Mill, the method of difference licenses inferences from a comparison of two instances which "have every circumstance save one in common". What Lipton termed the "problem of multiple differences" is that we will seldom or never be in a position to know that two instances fulfill this very strict condition. Lipton writes:

> Although Mill's strict statement of the Method of Difference sanctions an inference only when we know that there is a sole difference in the histories of fact and foil, Mill recognizes that this is an idealization. However similar the fact and foil, there will always be more than one difference between their antecedents. Some of these will be causally relevant, but others not. The problem of multiple differences is the problem of making this discrimination. (2004, p. 128)

Mill of course is well aware that the conditions he requires are very strict. He tries to ameliorate the problem in two ways: First, we already know many circumstances to be "immaterial" to the phenomenon, and second, experimental intervention usually gives us a reason for believing that only one antecedent condition has changed (1843, III.VIII.§3).

However, neither suggestion is fully satisfying. First, to assume that we know some circumstances to be causally immaterial is to presuppose causal knowledge, and so this suggestion leads us into circular and perhaps viciously circular reasoning.[5] Second, interventions may give us some psychological reassurance that the circumstance we manipulate is the only causally relevant difference between the instances, but of course even in the time it takes to perform the intervention many background circumstances are constantly changing. Most of these are probably immaterial, but again, how do we know this?

3.1 Rappaport on multiple differences

In his defense of Mill's method of difference against the problem of multiple differences, Rappaport (1996) takes up both of Mill's suggestions and elaborates on

[5]There is room for debate about whether such reasoning would be necessarily vicious. If the causal knowledge we need to presuppose for our causal inferences is *different* from the causal knowledge we wish to establish, then the circularity may be benign (see e.g. Woodward, 2003, pp. 104–107).

them. His proposed solutions are ultimately inadequate, but it is instructive to consider them in some detail.

Rappaport first discusses Mill's notion of disregarding potential causes which are known to be irrelevant. Lipton writes that this is a good idea in principle, but it "does not tell us how we determine irrelevance" (1991, p. 116). Rappaport thinks Mill implicitly relies on principles like the following (p. 74):

> [I]f an antecedent condition C is present in the absence of a phenomenon P, then C is not a cause of P.

To use Rappaport's example, we may wish to conclude that the blowing of the wind is the cause of the movement of the leaves. However, at the very moment the wind starts to blow, someone begins to sing on the street corner. How do we know that the wind, and not the singing, is the relevant antecedent difference? On Rappaport's account, we have often observed people to sing without the leaves moving, and so we already know singing to be an irrelevant difference.

However, if we adopted such a principle we would be committing one of the cardinal sins of causal inference. Rappaport's principle is systematically vulnerable to erroneous inferences because, in general, causes are not by themselves sufficient for their effects. In most cases, causes exert their effect only in conjunction with a set of additional causes. To illustrate, flipping the power switch on my coffee machine is not sufficient for fresh coffee to be brewed: It further requires the appropriate voltage being applied to the circuit, a ground coffee capsule, water in the tank, and so on. Now, if we accept that causes are generally not individually sufficient for effects, it tells us little that a potential cause C occurred in the absence of an effect E. This *may* indicate that C is not, in fact, a cause of E; but it may also indicate that, while C is causally relevant to E, some other part of the sufficient set of causes to which C belongs was not realized. If my flipping the power switch produces no coffee, this does not speak to the causal irrelevance of the switch: The water tank may be empty.

Thus, it would be a mistake to infer causal irrelevance from the absence of differences. This is a general impediment to the falsification of causal hypotheses (see Nickelsen and Graßhoff, 2011). So if we need knowledge about causal irrelevance in order to apply the method of difference successfully, we must establish it by a method different from Rappaport's principle.

Rappaport next discusses Mill's second suggestion, which concerns the special features of experiments. Mill writes that if we perform an intervention in a system

to produce some phenomenon (a potential cause), "we in general are entitled to feel complete assurance, that the pre-existing state, and the state which we have produced, differ in nothing except in the presence or absence of that phenomenon" (1843, III.VIII.§3). Mill thus believes that it is easy in experimental practice to produce instances such as are required by the method of difference.

Rappaport thinks that Mill's optimism is justified, since experiments have historically been successful in finding causes (p. 75). In addition to the argument from successful application (which by itself is cold comfort), Rappaport suggests that the ability to *control* alternative causes of the effect under investigation is the key philosophical justification of the success of the experimental method. But here I agree with Lipton, who notes that we may well be able to control *known* differences, but we should also consider *unknown* differences:

> [E]ven a careful experiment will seldom if ever leave us with only one possible cause, once we allow for the possibility of unobserved and indeed unobservable causes. (2004, p. 128)

Thus, even in an experiment where all known alternative causes are controlled we have no assurance that there aren't unobserved or unobservable antecedent differences which are *actually* responsible for the effect we observe.

3.2 Multiple differences in the IBE framework

Lipton's solution to the problem of multiple differences appeals to explanatory considerations. We may not know that the two instances we are comparing differ in only one circumstance – but we may be able to determine that the antecedent difference we suspect to be a cause would moreover provide a *lovely explanation* of the difference in the occurrence of the effect. Invoking classical virtues of well-confirmed hypotheses, Lipton proposes that "we prefer those differences that allow explanations that specify a mechanism, that are precise, and that contribute to the unification of our overall explanatory scheme" (1991, p. 119).

To substantiate the argument, let us turn to Lipton's chosen case study. This is the discovery of the cause of childbed fever by Ignaz Semmelweis (1818–1865). Semmelweis has served philosophers of science as a paradigm case of scientific discovery and confirmation ever since Carl G. Hempel used the case in his *Philosophy of Natural Science*.[6]

[6] See among others Hempel (1966, pp. 3–18), Lipton (1991, pp. 79–98), Gillies (2005), Russo and Williamson (2007), Bird (2010) and Scholl (2013). Consider also that the most widely used English

Semmelweis worked at Vienna's maternity hospital from 1846 to 1849 before returning to his native Hungary. The Viennese maternity ward had two divisions with shockingly different mortality rates. In the first division, the average mortality from childbed fever (postpartum sepsis) was around 10%, with peaks reaching as high as 30%. In the second division, by contrast, childbed fever was less severe with an average mortality of only around 3%. Semmelweis eventually hypothesized that the difference in mortality was due to the fact that the first division was used to train medical students, while the second was used to train midwives. The medical students, unlike the midwives, performed autopsies nearby and transferred a disease-causing agent (Semmelweis suspected diseased or decaying organic matter) from the autopsies to the patients. By introducing hand-washing measures for all doctors and medical students, Semmelweis was able to bring the mortality in the first ward down to levels even below those of the second ward.

Lipton takes Semmelweis's demonstration that cadaveric matter from the autopsies is the cause of childbed fever as an application of Mill's method of difference. We can take the diachronic comparison between the patients infected with cadaveric matter (before the institution of hand-washing measures) and the patients not so infected (after the intervention) as the two types of instances which are being compared: The effect under investigation (childbed fever) occurs in one but not in the other, and ideally the removal of cadaveric matter is the sole difference between the histories of the two instances.

It is easy to see how the problem of multiple differences applies here, and how Lipton's explanationist solution can help. After introducing hand-washing measures, Semmelweis may indeed not have known that the only causally relevant difference between the months when childbed fever mortality was high and the months when it was low was the absence of cadaveric matter on the hands of physicians. So Mill's method of difference, taken neat, does not sanction an inference. But if cadaveric matter *were* the relevant difference, it *would* provide a mechanism explaining the fact that mortality declines when hand-washing is mandated. Moreover, it would unify multiple facts: for instance, that childbed fever is more frequent in the first division and that street-births are never accompanied by childbed fever; or Semmelweis's belief that particles from a medullary carcinoma and a carious knee could also cause the disease. So Mill's method of difference is a correct but superficial description of our inductive practice: Its application re-

translation of Semmelweis's *Etiology of Childbed Fever* was produced by the philosopher K. Codell Carter (Semmelweis, 1983).

lies on additional judgments about the explanatory power of the cadaveric matter hypothesis.

Lipton's view is appealing, among other things because he ties together such individually plausible considerations as the method of difference, mechanisms, and unification. However, in the next section I will argue that a more parsimonious solution of the problem of multiple differences is possible. This solution does not depend on an explanationist framework, and there is evidence for its adequacy in the Semmelweis case itself. While some applications of the method of difference may work the way Lipton imagined, Semmelweis's did not.

3.3 Towards a better solution of the problem of multiple differences

Both Lipton and Rappaport understood experimental "controls" in the sense of keeping all else equal or of suppressing possible alternative causes. However, keeping all else equal is unattainable and thus unrealistic. Controlling alternative causes is more feasible, but it cannot deal with unknown alternative causes because we presumably cannot control what we do not know.[7]

To get a better grip on the problem, I suggest we should stop thinking in terms of antecedent and consequent "conditions" and, instead, adopt of thoroughly causal point of view: If an alternative cause is active in our system, we will be able to tell by the fact that it causes the effect under investigation to occur even in the absence of our intervention. We do not need to detect alternative causes themselves. We only need to estimate how frequently they interfere with the effect we are investigating.

Take a relatively simple causal structure as in figure 1. The idea is that our test factor causes the effect to occur if it is instantiated along with suitable co-factors. The effect can also be caused by one known alternative cause and one unknown alternative cause. For the purposes of the present discussion, the alternative causes are not split into their parts (e.g. "alternative cause" and "alternative cause co-factors"), but this should be understood as a shorthand.

In order to determine the causal relevance of the test factor by experiment, we need to intervene on the test factor while the alternative causes are silent. If one of the alternative causes were instantiated just as we intervene on the test factor,

[7]This is not without exception. For instance, a cell biologist's use of the same well-mixed culture medium for both the experiment and the control can certainly be understood as an attempt to control for potential unknown confounders by distributing them evenly. The same holds for randomization in clinical trials. However, the precise function of randomization has been the topic of interesting recent debates (reviewed in Howick, 2011, chapter 5).

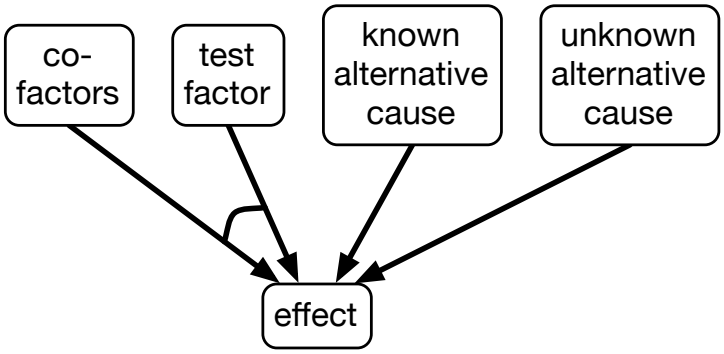

Figure 1. A simple causal structure. The test factor together with suitable co-factors (note the arc joining the arrows of the test- and co-factors) brings about the effect under investigation. In addition, the effect can be caused by a known set of alternative causes and by an unknown set of alternative causes. The aim of our experiment is to investigate the causal relevance of the test factor to the effect. Since our test factor is suppressed in the control experiment, any effects we observe in the control must be due to the known or the unknown alternative cause. Hence, the control experiment permits us to estimate the frequency with which both known *and unknown* alternative causes occur. See section 3.3 for details.

then this might spoof us into thinking that the test factor has caused the effect, when in fact it was the alternative cause. One measure we would certainly take to improve the experiment is to suppress known alternative causes of the effect under investigation. However, we might still be misled by the *unknown* alternative cause being instantiated during the experiment. This is the problem of unknown multiple differences.

Happily, we have the means at our disposal to determine whether the unknown alternative cause is active in our system. This is where the control experiment is key. In the control, we are suppressing the test factor. So we know that *if* the effect occurs, it *must be* because a known or unknown alternative cause of the effect is instantiated (otherwise, the principle of causality – no events without causes – would be violated). So the control experiment allows us to estimate the frequency with which known and unknown alternative causes of the effect under investigation are active in our system.

Once we know the frequency of alternative causes, we must compute whether a given number of occurrences of the effect under experimental intervention is compatible with our null hypothesis: that only the alternative causes are operating in our system. If the effect occurs frequently under intervention but rarely in the control, this gives us reasons to think that the intervention is having an effect – regardless of whether alternative causes that appear in the control are known or unknown. In order to determine causal relevance (as opposed, for example, to effect

size) all we need to know is the probability of at least one effect under investigation *not* being due to the alternative causes.

I submit that this captures the basic logic of the control experiment (presupposing, as before, the principles of causality and determinism: see section 2). By looking for effects *caused by* alternative causes, rather than by looking for alternative causes themselves, we circumvent the problem of unknown differences. Both known and unknown alternative causes will be visible through their effects in the control experiment. We then estimate how frequent such alternative causes are, and calculate how likely it is that the experimental result is due only to the alternative causes.[8]

A toy example will help to clarify the point (but it is important to keep in mind that real-world cases will be far more complicated). Let us imagine that we have a large and somewhat badly wired house where using any light switch in any room will turn on all the lights in the house. Knowing nothing of this, we may now wish to find out whether the light switch in *our* room is causally relevant for the light turning on. We thus flip our light switch and find that the light does turn on. Now this may indicate that our light switch has caused the light to turn on, but of course it may also indicate that somebody in another room of the house has flipped *their* switch. To find out, we would observe the light for some time to determine how frequently it turns on if we don't touch our switch (this is the control experiment). If the light turns on very rarely in the absence of us flipping our switch, then the fact that the light *does* turn on if we flip our switch is a reliable indicator of the causal role of the switch. On the other hand, if the light frequently turns on even if we do nothing, our inference must proceed much more carefully. I take this part of the story to be straightforward and uncontroversial. The important additional point to note, however, is that this procedure does not depend on us having substantial knowledge about the other rooms in the house (although such knowledge would help, of course). Our causal inference can proceed even if we do not know how many light switches or how many rooms there are. All we need to know is how frequently our light is made to turn on by causes, known or unknown, other than our own light switch.

Presumably, among the biggest inductive risks in this procedure is the fact that

[8]This can be made precise using a binomial test. First, we determine the frequency f_C with which alternative causes occur in the control experiment (or more precisely the upper bound of f_C compatible with the control). Then we ask whether the frequency of the occurrence of the effect in the experiment (f_E) is compatible with the null hypothesis that only one or several alternative causes with overall frequency f_C are active in the system.

we need to extrapolate from the frequency with which alternative causes occur in the control to the frequency with which they occur during the experiment. But this risk is fortunately local. We do not need to make any global pronouncements about the frequency with which alternative light switches are used in our house. The frequency of alternative causes may even be quite variable. We may find that the light gets turned on a lot during business hours, but almost never at 3 a.m. So experiments performed at 3 a.m. permit inferences more easily than experiments performed at 3 p.m. We can find out that this is so by means of the control experiment – that is, by simply not flipping our switch.

Another inductive risk in this procedure is that we must assume that our test factor is independently manipulable. If our intervention instantiates the test factor but also causes the effect to occur by some *alternative* causal pathway, then our causal inference would be misleading – and the control experiment would not help us in detecting this state of affairs. We may call this the *problem of systematic multiple differences* or the problem of fat-handedness.[9] Proponents of IBE might use the problem of systematic multiple differences to counter my Millian arguments. After discussing the Semmelweis case and multiple differences in section 3.4, I will consider the problem of systematic multiple differences in section 3.5.

To summarize, what at first appears to be an unachievable requirement – all things equal, or control of all possible alternative causes – turns out to be experimentally tractable: Control experiments tell us something about both known *and unknown* alternative causes. The basic point is that we do not need to know *how* an effect is caused to see in the control experiment that *some* alternative cause is active. Because control experiments control for both known and unknown alternative causes, the method of difference allows us to investigate causal relevance even from a point of relative ignorance.

3.4 Back to Semmelweis

An understanding of the control experiment roughly in agreement with my reconstruction appears to be in the background of several passages in Semmelweis's *Etiology*, particularly in his replies to critics. Remember that Semmelweis observed a relatively high incidence of childbed fever for several years, and then introduced hand-washing measure, at which point the incidence of the disease decreased. In this setup, the early years with high mortality are the "intervention" group and the

[9] See also Scheines (2005), as well as Woodward's characterization of interventions (Woodward, 2003, chapter 3).

later years with low mortality are the "control" group. The role of the control group becomes relevant in Semmelweis's exchange with Professor Levy of Copenhagen, who wrote an early critique of Semmelweis's *Lehre*. Levy writes:

> [A]nyone who has had the opportunity for a number of years to follow the periodic rise and fall of morbidity in the obstetric ward will doubtless have to admit that we cannot credit the results [of Semmelweis's intervention] without knowing whether the wards had equally favorable periods in past years as during the past seven months, for which we would need an exact statistical report about monthly morbidity and mortality.[10]

I take the point to be that the incidence of childbed fever might fluctuate in relatively long intervals, such that a seven-month period during which the incidence is low is to be expected even in the absence of a successful intervention against the disease's cause. Perhaps if we ran the control experiment – where the suspected cause of the disease is suppressed – for a longer period of time, we would see that some (unknown) cause of childbed fever appears and disappears for many months at a time. In other words: Perhaps our control experiment did not run for long enough for us to reliably estimate the true frequency of unknown alternative causes. Perhaps the experimental and control columns are *both* compatible with some unknown cause or causes of childbed fever acting intermittently, and with the intervention having no effect.

In the *Etiology*, Semmelweis counters this exactly as we would expect if our reconstruction of the logic of the control experiment is roughly correct. He insists, in effect, that the control experiment has now run for a long enough time for such confounders to be found:

> Time has refuted this point, we are no longer concerned with seven months but with twelve years.[11]

[10] Semmelweis (1861), p. 300: "[J]eder nämlich, der durch eine längere Reihe von Jahren dazu Gelegenheit gehabt hat, das periodische Steigen und Fallen der Kränklichkeit in Gebäranstalten zu beobachten, wird ohne Zweifel eingestehen müssen, dass uns zur Würdigung der gewonnen Resultate wesentlich darüber Aufschluss mangelt, ob nicht auch in früheren Jahren die Anstalt ebenso günstige Perioden gehabt hat, als in den letzten sieben Monaten, wozu eine genaue statistische Mittheilung über die monatlichen Krankheits- und Todesfälle erforderlich wäre." K. Codell Carter offers a less faithful but more felicitous translation in Semmelweis (1983), p. 184.

[11] Semmelweis (1861), p. 301: "Diesen Punkt hat die Zeit widerlegt, es handelt such jetzt nicht mehr um sieben Monate, sondern um mehr als zwölf Jahre." See also Semmelweis (1983), p. 184.

Semmelweis is arguing that his control experiment ultimately ran for twelve years during which the incidence of childbed fever never rose again – and thus the frequency of alternative causes of the disease *apart from* the cause Semmelweis is suppressing is extremely low, and cannot account for the high incidence of the disease in earlier years. Cadaveric matter and similar contaminants must be the cause of childbed fever.

Semmelweis has a similar exchange with Eduard Lumpe, who like Semmelweis had served as assistant in the Viennese maternity ward. Lumpe notes that if an effect can be silent for eight months, there is no reason to think that it cannot be silent for three years. So the question again concerns the frequency with which childbed fever appears and disappears irrespective of cadaveric matter contamination. Semmelweis notes wryly that Lumpe may have had a point earlier, but that in the meantime a whole ten years had passed "and the accidental cessation of mortality has repeated itself in several places" (1861, p. 450). Stripped of irony, the claim is that the long-term reduction in childbed fever mortality – Semmelweis's prolonged control experiment – shows that Semmelweis is suppressing the true cause of the disease, since no alternative causes occur with a frequency that would be compatible with the incidences observed in past years.

I hesitate to claim that Semmelweis's understanding of his method is exactly congruent with my reconstruction; it probably is not. However, Semmelweis seems to recognize that he can estimate the frequency of causes of childbed fever *other than cadaveric matter* from observing the "control" experiment; and that the less frequent alternative causes are, the less likely it is that earlier clusters of childbed fever were caused by something other than the cause he is now suppressing. The Semmelweis case may highlight this type of reasoning particularly well, since it was ethically indefensible to reintroduce the suspected cause of the disease in order to create a cleaner experiment – and so Semmelweis had to make the most of what he could learn from the control experiment.

3.5 Systematic multiple differences: A possible explanationist riposte

Proponents of IBE might now object that assessing the frequency of alternative causes in the control experiment does not protect against the problem of systematic multiple differences. Conceivably, the intervention which brings about the test factor is "fat-handed" in that it also brings about some other (unknown) cause of the effect under investigation, and it may be this additional cause – and not the test factor – which explains the intervention's efficacy. Since this additional difference

is systematically caused by our intervention, the control experiment would not help us to detect it. Mill himself recognizes the difficulty but believes that it "generally admits of being conclusively tested by other experiments" (Mill, 1843, III.VIII.§3).

While this is not one of Lipton's arguments, proponents of IBE might argue that only explanationist considerations can provide a way out.[12] Semmelweis could not know that his test factor of cadaveric matter was not accompanied by systematic multiple differences. However, the causal role of cadaveric matter was part of the best explanation of the experimental results, and thus the explanationist framework provides a basis for Semmelweis's judgment.

The problem with this explanationist approach, however, is that it proves too much. For Semmelweis's experiment *was in fact* subject to the problem of multiple differences. Childbed fever is not caused by cadaveric matter, but by bacteria colonizing it. Bacterial infection alone would have caused the disease, while sterilized cadaveric matter would have been innocuous. This was only learned in the course of the subsequent elucidation of the bacterial mechanisms of infectious disease.

More generally, Semmelweis's clinical results were compatible with a range of mechanistic hypotheses concerning the relationship between hygiene and disease. It was possible that cadaveric matter caused the disease, as Semmelweis believed. It was also possible that a concomitant of cadaveric matter – as we later learned, bacteria – caused the disease. But the known facts were compatible with yet a third possibility (incorrect, and never to my knowledge considered): small amounts of chlorinated lime solution might have been an effective *cure* of childbed fever. The clinical experiment did not settle these questions; results form subsequent causal and mechanistic research were required.

Proponents of IBE might now take this admitted requirement to integrate data from multiple avenues of research as an indication that explanationist considerations are paramount after all. However, the need to integrate multiple sources of data only reminds us of what Imre Lakatos taught us long ago: that we should not expect "instant rationality" (Lakatos, 1970). We must be careful to distinguish between (1) the integration of information from multiple sources *into* an explanation and (2) the use of explanatory power to *infer* truth. That we naturally use our best knowledge to explain makes IBE intuitively appealing. But needless to say, it makes a difference for this debate whether we give explanations on the basis of inferred regularities, or whether we infer regularities on the basis of their explana-

[12] I thank an anonymous referee for urging me to discuss this point at greater length.

tory power. That Semmelweis and subsequent researchers did the former does not show that they did the latter.

4 The problem of inferred differences

The second of Lipton's challenges for Mill's method of difference concerns unobserved or unobservable differences. One may grant that the inference from sole antecedent difference to causal role is justified, provided that the problem of multiple differences can be overcome (perhaps along the lines I outlined above). Nevertheless, the method of difference seems to require that we already know about the antecedent difference whose causal role we wish to infer. Lipton however believes that science often involves inferences about previously unobserved or unobservable differences:

> The Method of Difference sanctions the inference that the only difference between the antecedents of a case where the effect occurs and one where it does not marks a cause of the effect. Here the contrastive evidence is not evidence for the existence of the prior difference, but only for its causal role. The method says nothing about the discovery of differences, only about the inference from sole difference to cause. So it does not describe the workings of the many contrastive inferences where the existence of the difference must be inferred, either because it is unobservable or because it is observable but not observed. (2004, p. 127)

I will begin with Lipton's solution to the problem of inferred differences in the explanationist framework. Next, I will consider and critique Rappaport's solution to the problem. Lastly, I will develop my own approach.

4.1 The problem of inferred difference in the IBE framework

Lipton again illustrates the problem of inferred differences using the Semmelweis case. If Semmelweis had observed that only women in the first ward were infected by cadaveric matter, he might have been in a position to infer cadaveric matter's causal role. However, Semmelweis could not make such an observation because infection by cadaveric matter is an unobservable process. Lipton claims that IBE offers a natural solution, since it allows us to infer not only the causal

role of observed differences, but also the explanatory loveliness of unobserved or unobservable differences:

> We are to infer that a difference marks a cause just in case the difference would provide the best explanation of the contrast. Because of this subjunctive process, absent from the Method of Difference, we may judge that the difference that would best explain the evidence is one we do not observe, in which case Inference to the Best Explanation sanctions an inference to the existence of the difference, as well as to its causal role. (2004, p. 127)

As in the abstract, so in the case study: Semmelweis could not observe cadaveric matter infecting women in the first ward, but he was able to judge that if this difference *did* exist, it *would* provide a lovely explanation of the contrast in mortality between the first and second wards, as well as of the contrast before and after the introduction of hygienic measures and between hospital and street births. Hence, Semmelweis was able to infer the causal role of cadaveric matter only by supplementing Mill's method of difference with explanationist considerations.

4.2 Rappaport on the problem of multiple differences

In his defense of Mill's methods against IBE, Rappaport (1996) grants that Lipton has identified a genuine limitation of Mill's method of difference (p. 70–72). He replies, however, that the method of difference is only one tool among several: The method of difference's limitations are compensated for by the method of residues. Rappaport's formulation of the method of residues is the following (p. 71):

> If (a) antecedent conditions C1, C2, … Cn and Cn + 1 are the likeliest candidates for the complex cause of a total phenomenon P, and (b) it is known that part of P is the effect of C1, C2, … Cn, then infer that the residue of P (the other part of P) is the effect of Cn + 1.

Rappaport stresses that "the antecedent condition Cn + 1 may *not* be known to exist prior to the application of the Method of Residues" (p. 71, emphasis in original). By way of illustration, he offers the example of Adams's and Le Verrier's discovery of Neptune: Uranus's position could not be fully accounted for by the gravitational effects of the known seven planets, but it could be accounted for on the supposition that an eighth planet existed beyond Uranus. Hence, the method of residues was

used to infer the *existence* of a previously unknown antecedent condition (p. 71-72).

There are two main flaws to Rappaport's suggested solution. The first and lesser flaw is that Rappaport is proposing something different from what I take to be Mill's notion of the method of residues. Second, regardless of whether we choose Mill's or Rappaport's version of the method, it is impossible to apply it to the Semmelweis case. Since Semmelweis is Lipton's chosen illustration of the problem of inferred differences, we must surely reject any proposed solution that cannot illuminate the Semmelweis case.

If we turn to Mill's formulation of the method of residues (his "fourth canon"), we find something related but distinct from what Rappaport suggests:

> 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. (1843, III.VIII.§5)

Mill is describing a situation where we are observing both a residual effect and a residual antecedent condition, and from this we infer the *causal role* of the residual antecedent condition with respect to the residual effect.[13] The basic idea is analogous to the method of difference, except that we do not have a proper control experiment: We cannot obtain an instance where both cause and effect are absent. The control is replaced by our computation that the known causes account for a certain part of the observed effect, and so any residual effect is attributed to a residual antecedent condition.

That this is Mill's intended procedure is evident from his consideration of the conditions under which the method would fail. First, we have to be sure of the causal roles of the *known* antecedent conditions; only then can we reliably determine whether known causes fully account for the effect, or whether there exists a residual causal influence. Second, when attributing the residual effect to a residual antecedent condition, we need to be certain "that C [the residual antecedent condition] is the *only* antecedent to which the residual phenomenon c [the residual effect] can be referred" (1843, III.VIII.§5). These caveats are only compatible with the method of residues as a method for establishing causal roles, not unknown antecedent conditions.

[13] I take it that Lipton (2004), p. 127, shares my interpretation of Mill.

Rappaport is thinking of something different: His is a diagnostic situation where a number of possible causes of an effect are known – hence, causal roles are already established. We then reason from the occurrence of the effect, *via known causal roles*, to the existence of instances of known types of causes. Consider Rappaport's example of Neptune's discovery (p. 71–72). Adams and Le Verrier worked out that a certain residual effect (a disturbance in the path of Uranus) could be explained through (1) a known type of interaction (gravitational forces) and (2) a plausible constellation of antecedent conditions (the known planets *plus* an additional planet beyond Uranus).

The discovery of Neptune was a discovery made possible by narrow constraints on possible solutions: The relevant causal interactions, as well as the likely distribution of unknown antecedent conditions in orbits around the sun, were well defined. In many ways, the discovery is thus more like a physician's diagnosis of a known disease based on symptoms than Semmelweis's discovery of a novel disease process. It is telling that Rappaport's second example is a medical diagnosis (p. 71). The deviation in Uranus's orbit is a sign permitting the diagnosis that a further accretion of matter must be exerting an effect. It is here quite clear how we are searching for a "residual" cause that is responsible for a "residual" effect, and that our reasoning depends on already understanding causal roles.[14]

Let us for now grant that two variants of Mill's method of residues exist.[15] We may have (1) Mill's version where we have a residual effect and a residual antecedent condition and infer from this the causal role of the residual condition. Or we may have (2) Rappaport's "diagnostic" version where we have a residual effect as well as known causal interactions and deduce from this the constellation of antecedent conditions that must exist to bring about the effect.

The problem with both versions is that they can at best handle *half* of Lipton's challenge. The two versions of the method of residues may be able to deal with cases where the relevant antecedent difference is previously *unobserved* as in the case of Neptune. Here the method of residues may give us considerable guidance in the context of discovery, something which Mill already recognized:

[14] An anonymous referee has raised an interesting worry about modularity: If A causes Y and B causes Z, then the conjunction of $A\&B$ may not simply cause the conjunction of $Y\&Z$ but rather some qualitatively different phenomenon. In this case, the method of residues would not yield good results. However, since Semmelweis's inferences cannot be reconstructed as an instance of the method of residues in any case (as discussed below), the method's problems remain an intriguing topic for another occasion.

[15] However, my inclination would be to confine the term to Mill's notion and to regard what Rappaport describes as "diagnostic causal reasoning".

> Of all of the methods of investigating laws of nature, this is the most fertile in unexpected results [...] The agent C may be an obscure circumstance, not likely to have been perceived unless sought for, nor likely to have been sought for until attention had been awakened by the insufficiency of the obvious causes to account for the whole of the effect. (1843, III.VIII.§5)

However, Lipton's challenge concerns not only unobserved but also *unobservable* differences, as in the case of Semmelweis's investigation. And here it is not obvious that either version of the method of residues could be sufficient.

Semmelweis was not engaged in a diagnostic procedure of the type Rappaport describes as the method of residues. Such a procedure would have to take roughly the form of Semmelweis observing a residual effect (increased childbed fever in the first ward), knowing how the effect can come about (infection by cadaveric matter), and inferring from this that cadaveric matter must have infected the women who fell ill. While this would permit Semmelweis to infer an unobservable difference in cadaveric matter between the wards, it cannot be his actual procedure – for he did not know with any certainty that cadaveric matter causes childbed fever. This was just the causal role that needed to be inferred by clinical and animal experiments.

Alternatively, we may adopt Mill's own method of residues and ask whether it helps us to understand how Semmelweis inferred unobservable antecedent differences. The procedure would have to work along the following lines: Semmelweis observed a residual effect (increased childbed fever in the first ward) which he could only attribute to one particular unobservable residual antecedent difference between the wards (cadaveric matter being transferred during examination). If the method of residues is to be of any help with unobservables, we must be in a situation where the candidate antecedent conditions are clearly and exhaustively defined. We must be able to infer, from the fact that none of the *observable* differences cause the residual effect, that the cause can only be one particular *unobservable* difference, cadaveric matter infection. While we may perhaps get to the point where none of the observable differences remain as candidate causes, it is difficult to see how we could ever narrow down the range of unobservable candidate antecedent differences to just one.

The conclusion is therefore that the method of residues does not solve the problem of inferred differences for the case of unobservables – neither on Rappaport's nor on Mill's conception of the method. Both conceptions have merit as part of a methodology of science: Diagnostic reasoning as described by Rappaport is in-

dispensable, and Mill's notion that unexplained residual effects lead us to investigate additional possible causes is a plausible guide in the context of discovery. Nevertheless, the problem of inferred unobservable differences requires a different solution.

4.3 Towards a better solution of the problem of inferred differences

How then can we approach the problem of inferred unobservable differences? In discussing the problem of multiple differences above, I suggested that we need to adopt a thoroughly causal point of view. The same approach can also help with the problem of inferred differences. All our interactions with the system under investigation are causal: In experiments, we cause certain antecedent differences whose effects we wish to study, and when we observe an effect to occur or not to occur this will be through some causal interaction, be it a straightforward visual inspection or an instrumental measurement. We can call these processes "intervention processes" and "measurement processes", respectively.

It would be a mistake to think that the intervention on the test factor and the measurement of the effect under investigation are the *only* causal interactions that enter into an application of the method of difference. Let us assume that we wish to study whether a particular unobservable condition is the cause of an also unobservable effect. We will need some means by which we can produce the condition, and we will need some means by which we can measure that the effect has occurred. In addition, however, in any realistic experimental scenario we will require some way to measure whether our intervention has *successfully* produced the condition whose effects we wish to study. Let us assume, as a simple example, that we wish to study the effect of a restricted DNA fragment within a type of cell. We will certainly not fly blind without instruments by creating a particular DNA fragment using restriction enzymes, inserting the fragment into cells, and measuring whether the effect in question appears. After using the restriction enzymes, we will make sure that the *correct* DNA fragment has been produced – for instance, by running a polymerase chain reaction on the product of the restriction reaction and sequencing the product. It is this further causal interaction with the system that assures us that we are in fact studying the correct unobservable antecedent difference. Figure 2 gives a schema of the various interactions that are involved in an experiment where the causal role of inferred differences is tested.

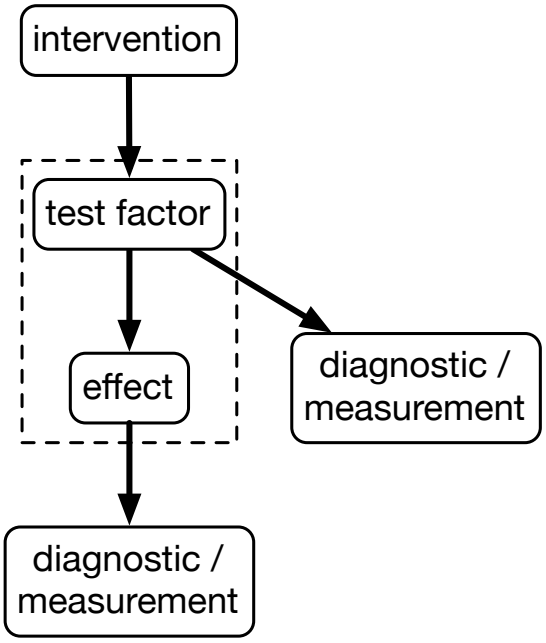

Figure 2. In a standard experimental application of the method of difference, we intervene on a test factor whose causal relevance to an effect we wish to ascertain (the dashed box indicates the causal relationship under investigation). In the simplest case, we have direct visual access to the test factor and the effect. Even if the test factor and the effect are unobservable, however, the method can be applied provided we have some sort of causal access. We call such access either a "diagnostic" causal interaction, a process of "detection", or simply a "measurement". Note that we would typically not only measure whether the effect has occurred, but also whether our intervention successfully produced the test factor we are interested in. This is the solution to the problem of inferred differences which is developed in the main text. See section 4.3 for details.

4.4 Back to Semmelweis

These considerations are applicable to Semmelweis's clinical investigation. The cadaveric matter whose causal role was in question could not be directly observed: Even before Semmelweis instituted hygienic measures, physicians *did* clean their hands after autopsies – certainly to the point where no visible residue remained on their hands. This is why Lipton writes that Semmelweis was investigating the causal role of an inferred difference. In fact, however, Semmelweis did have causal access to the "inferred" difference. While he could not *see* minute residues of cadaveric matter, he could *smell* them. This is how Semmelweis "diagnosed" or "measured" the presence of minute amounts of cadaveric matter whose causal role he put to experimental test.

There is ample evidence in Semmelweis's writings that detection of cadaveric matter by smell played a major role in his reasoning (and the philosophical litera-

ture has not so far acknowledged this). Semmelweis's first public presentation of his findings on May 15, 1850, is only preserved as a report (we do not have Semmelweis's actual text), but the minute writer of the k.-k. Akademie der Aerzte zu Wien, Dr. Heinrich Herzfelder, duly notes:

> Following this idea [that cadaveric matter is the cause of childbed fever], Dr. Semmelweis instituted that any students or other examiners of women who were pregnant or in childbed had to clean their hands carefully in a solution of chlorinated lime, such that any possible putrid organic atom adhering to the fingers was thoroughly extinguished, *even down to its smell*...[16]

Similarly, in the *Etiology* Semmelweis takes the smell of cadaveric matter residues as proof that hand-washing with normal soap is insufficient:

> That after the normal way of hand washing with soap not all cadaveric matter adhering to the hands is removed, is proved by the cadaveric smell which the hands retain for a longer or shorter period of time.[17]

The point resurfaces several times in the *Etiology*. Hence, Semmelweis did not need to regard cadaveric matter as merely an "inferred difference": Through the sense of smell, he was able to detect its presence.

It may be objected that Semmelweis's causal access to the cadaveric matter residues was quite tenuous. I would agree: Semmelweis was only at the beginning of the investigation of infectious diseases, and in time it was necessary to understand in detail what it was about cadaveric matter that could cause disease. The role of microorganisms needed to be elucidated, and reliable methods for their detection needed to be developed (growth cultures, stainings, PCR). With this additional knowledge, better experiments involving unobservables became possible. Causal access to unobservables comes in degrees. But Semmelweis already had *some* access, and so cadaveric matter was not a difference inferred on explanatory grounds.

[16] Herzfelder (1850), p. CXXXVIII: "Dieser Idee nun folgend, führte Herr Dr. Semmelweiss [sic] ein, dass jedweder der Schüler oder sonst Untersuchenden vor jeder Exploration einer Schwangeren, Kreissenden oder Wöchnerin seine Hände in einer Chlorkalklösung sorgfältig wasche, um so jedes möglicher Weise an den Fingern haftende, faulende organische Atom, selbst *bis auf den Geruch desselben* vollends zu tilgen...". The emphasis is mine, as is the translation.

[17] Semmelweis (1861, p. 54): "Dass nach der gewöhnlichen Art des Waschens der Hände mit Seife die an der Hand klebenden Cadavertheile nicht sämmtlich entfernt werden, beweist der cadaveröse Geruch, welchen die Hand für längere oder kürzere Zeit behält."

5 Conclusions

I have considered Lipton's two main arguments for thinking that Mill's method of difference cannot do substantial inferential work unless embedded in an IBE framework. These are the problems of multiple and inferred differences.

Mill's method of difference allows us to infer causal roles if we are in possession of two instances where the effect occurs in one but not in the other, and where a sole difference exists in the antecedents. However, the problem of multiple differences is that we are rarely or never able to say with any certainty that the antecedents of two instances differ in only one circumstance. This difficulty is compounded if we consider not only unobserved or unconsidered but also unobservable antecedent differences. I have argued that this problem can be overcome through a proper account of the role of control experiments. The solution requires that we think in terms of causation rather than differences, and that we recognize how the control experiment gives us information about the frequency with which alternative causes (both known and unknown) exert their effects in our system. Finally, I showed that this type of thinking is compatible with Semmelweis's exchanges with critics who doubted that he had identified the true cause of childbed fever.

The problem of inferred differences is that we are often dealing with unobserved or unobservable antecedent differences. Lipton argues that only IBE allows us to infer the causal roles of such differences. Rappaport successfully showed that the method of residues offers guidance for the discovery of previously *unobserved* antecedent differences (such as Neptune in the case of Adams and Le Verrier). However, he failed to offer a solution which is applicable to Semmelweis's *unobservable* difference in cadaveric matter. I have shown that Lipton is incorrect to describe Semmelweis's cadaveric matter as inferred on explanatory grounds, since Semmelweis had causal access to minute residues of it through the sense of smell.

As I wrote in the introduction, the goal is not a reduction of IBE to Mill's methods. There likely exist "hard cases" where Mill's methods cannot account for our inductive practices, and in those cases IBE may be a useful framework for confirmation theory. How often a purely Millian approach suffices to describe and justify our inductive practices must be investigated by careful historical and philosophical scholarship. For now, my task has been to show that Mill's method of difference can be successfully defended against Lipton's two challenges. At

least in so far as the Semmelweis case is concerned, the method has substantial autonomy outside an IBE framework.

References

Baumgartner, M. (2008). Regularity theories reassessed. *Philosophia*, 36(3):327–354.

Baumgartner, M. and Graßhoff, G. (2004). *Kausalität und kausales Schliessen: eine Einführung mit interaktiven Übungen*. Bern Studies in the History and Philosophy of Science, Bern.

Bird, A. (2010). Eliminative abduction: examples from medicine. *Studies in History and Philosophy of Science*, 41:345–352.

Gillies, D. (2005). Hempelian and Kuhnian approaches in the philosophy of medicine: the Semmelweis case. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(1):159–181.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, 44(1):49–71.

Graßhoff, G. and May, M. (2001). Causal regularities. In Spohn, W., Ledwig, M., and Esfeld, M., editors, *Current issues in causation*, pages 85–114. Mentis, Paderborn.

Hempel, C. G. (1966). *Philosophy of Natural Science*. Prentice Hall.

Herzfelder, H. (1850). Protokoll der allgemeinen Versammlung der k. k. Gesellschaft der Ärzte, vom 15. Mai 1850. *Zeitschrift der k. k. Gesellschaft der Aerzte zu Wien*, 6(1):CXXXVI–CXLI.

Hofmann, U. and Baumgartner, M. (2011). Determinism and the method of difference. *Theoria*, 26(2):155–176.

Howick, J. H. (2011). *The philosophy of evidence-based medicine*. Wiley-Blackwell.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos, I. and Musgrave, A., editors, *Criticism and the Growth of Knowledge*, pages 91–196. Cambridge University Press.

Lipton, P. (1991). *Inference to the Best Explanation*. Routledge, London and New York.

Lipton, P. (2004). *Inference to the Best Explanation*. Routledge, London and New York.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, pages 1–25.

Mackie, J. (1980). *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford.

Mill, J. S. (1843). *A System of Logic*. John W. Parker, London.

Nickelsen, K. and Graßhoff, G. (2011). In pursuit of formaldehyde: Causally explanatory models and falsification. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(3):297–305.

Rappaport, S. (1996). Inference to the Best Explanation: Is It Really Different from Mill's Methods? *Philosophy of Science*, 63(1):65–80.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science*, 72(5):927–940.

Scholl, R. (2013). Causal inference, mechanisms, and the Semmelweis case. *Studies in History and Philosophy of Science*, 44(1):66–76.

Semmelweis, I. P. (1861). *Die Aetiologie, der Begriff und die Prophylaxis des Kindbettfiebers*. C. A. Hartleben, Pest, Wien und Leipzig.

Semmelweis, I. P. (1983). *The etiology, concept, and prophylaxis of childbed fever (Carter translation)*. University of Wisconsin Press, Madison.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.