Reasoning from examples is threatened by Simpson's paradox

Harry Long and Richard Pettigrew

Abstract

A very common method of reasoning in philosophy involves describing examples, cases, scenarios, or vignettes and asking one's readers to agree with certain intuitive judgements about them. Call this *reasoning from examples*. Certain writers—notably Kagan (1988), Wilson (2016), and Crain (2023)—have raised objections to it. This essay contributes two novel objections. First, unless the reasoner ensures that the reader fixes certain features of the examples before she gives her intuitive judgement, the reasoning is unreliable. This is because the reasoning might be vulnerable to versions of Simpson's paradox. Second, if such reasoning is used to justify policy, then certain formal results in the causal inference literature show that it is only valid if she takes into consideration the causal structure that holds between features of the scenario.

1 Introduction

In philosophy, when we *reason from examples*, we present a collection of cases, elicit certain judgements about them from our reader, and appeal to those judgements to provide support for a hypothesis, H, that concerns the relation that holds between some collection of variables. Some such reasoning involves the attempt to reproduce, in certain ways, the methodology employed in controlled scientific experimentation.[1] For example, one might make the cases in each comparison-class match in their value for all but one variable apparently relevant to the value of the dependent variable of interest, so that any subsequent difference in the value of this variable can be inferred to be due to the one independent variable in which the cases differ (see, for example, the 'contrast arguments' in Kagan (1988)).[2] Call this *philosophico-scientific reasoning from examples*. A number of writers—notably Kagan (1988), Wilson (2016), and Crain (2023)—have raised objections either to the general style of reasoning, or to the claimed analogy with scientific reasoning from data.

[1] Cf. (Tooley 1972; Trammell 1975; Thomson 1985; McMahan 2002; Boonin 2008).
[2] Kamm calls contrast argumentation the 'technique of equalizing cases' (Kamm 2007, 427-8).

This essay contributes two novel objections to reasoning from examples in general. First, unless the reasoner ensures that the reader fixes certain features of the examples, cases, scenarios and vignettes before she gives her intuitive judgement, the reasoning is unreliable. This is because, without such an assurance, the reasoning might be vulnerable to versions of Simpson's paradox. We will argue that this first objection is a special case of the objection originally posed by Bernard Williams (1973) in 'The Self and the Future'. Second, if such reasoning is used to justify policy or action, then certain formal results in the causal inference literature—exposited clearly in Pearl (2009) and Spirtes et al. (2000), for example—show that it is only valid if she takes into consideration the causal structure that holds between features of the scenario. If the reasoner endorses the analogy to controlled scientific experimentation, then it is only valid *by her own lights* if she does so.

2 Simpson's paradox and probabilistic reversals

Simpson's paradox occurs when the direction of association between two variables overall is reversed or absent for every cell of some partition (Simpson 1951). There might be a negative overall association between a treatment b and recovery a, but a positive association between treatment and recovery for every size of kidney stone c, c', that one might have. Take the following data:

Table 1 Treatment data partitioned by B only

	Treatment (b)	Placebo (b')	Total
Success (a)	130	170	300
Failure (a')	200	160	360
Total	330	330	660

From Table 1, we see that the placebo b' has a higher success rate overall:

$$P(a|b) < P(a|b') \tag{1}$$

But now consider another summary of the data: From Table 2, we see that both for

Table 2 Data partitioned by B and C

	Treatment (b) Small (c)	Placebo (b') Small (c)	Treatment (b) Large (c')	Placebo (b') Large (c')	Total
Success (a)	30	10	100	160	300
Failure (a')	160	80	40	80	360
Total	190	90	140	240	660

small kidney stones c and for large ones c', the treatment has a higher success rate:

$$P(a|bc) > P(a|b'c) \tag{2}$$

$$P(a|bc') > P(a|b'c') \tag{3}$$

This is an instance of *Simpson's Paradox* (we will also call it *the reversal paradox* or simply a *reversal*).[3]

The 'paradox' arises because the data are oddly distributed (Cartwright 1979). In the case at hand, inequalities (1), (2) and (3) hold because the following pair of inequalities also do:

$$P(a|c) < P(a|c') \tag{4}$$

$$P(c|b) > P(c'|b) \tag{5}$$

The problem is exacerbated by the fact that different partitions of the data may *retrieve* the aggregate-level relationship between b and a. For example, the same data could be partitioned by the variable D—with values: d = being under 50, d' = being at least 50—such that, for both jointly exhaustive age groups, there is a negative association again between the treatment and recovery.

Association reversal may obtain across a subjective *probability distribution*, or *in the world*, if the phenomena are so related (or the objective chance distribution such) that the structure described above holds for them. Say there is *worldly* reversal exactly if such reversal is exhibited either by the (short- or long-run) statistical frequencies or the chance distribution. The first problem we raise for reasoning from examples concerns the first of these possibilities, and the second problem the second.

It might seem that Simpson's paradox arises predominantly in non-experimental scenarios, and that in thought experimentation, we have all the control we need to avoid it. We argue this appearance is illusory. If we are not careful with our reasoning, our reader may, in a sense, generate the paradox themselves.

[3]We are grateful to an anonymous reviewer for pointing out that Judea Pearl distinguishes 'Simpson's paradox' from 'the reversal paradox' for principled—albeit contentious—reasons bearing on what the putative source of the paradox is (Pearl 2016). We use the terms interchangeably without commitment to his view.

3 Reasoning from examples

Reasoning from examples looks like a highly prevalent method in philosophy. It is arguably exhibited as early as the Ring of Gyges story in Plato's *Republic* (Book 2, 360 b-d). It also appears throughout canonical twentieth-century analytic philosophy (Gettier 1963; Frankfurt 1969; Dretske 1970; Goldman 1976; Thomson 1985)

How to define or characterise the method more precisely is contested. For instance, Edouard Machery distinguishes five uses to which philosophers put cases: the *provocative*, *illustrative*, *formal*, *material*, and *exploratory* (Chapter 1, Machery 2017). James Wilson draws the lines differently by distinguishing *one-off clear cases*, *one-off problem cases*, and these same two as elements of *interventional sequences* (Wilson 2021). We need not choose among these taxonomies, for there is common ground enough among them for our point to go through.

One important piece of common ground is this: on every standard characterisation of the method, some agent forms a judgement, or some other propositional attitude, about the situation a case describes (Machery 2017, pp.11-12), (Malmgren 2011, p.264), (Horvath 2015, p.386), (Wilson 2016, p.48). Following Machery, we have been calling this agent *the reader* (Machery 2017, p.12). She need not be distinct from the reasoner. Much reasoning from examples is conducted by a philosopher consulting her own responses.[4] For the problems we raise to obtain, though, it is not required that the reader be distinct from the reasoner.[5]

Our problems are independent of other families of objection levelled at reasoning from examples. One family questions the probative force of judgements about examples deemed excessively far-fetched or atypical (Lewis 1986; Elster 2011; Machery 2017); another whether results applying in one context transfer to another (Kagan 1988; Wilson 2016); and yet another whether intuitive judgements are invariant across non-alethic factors such as ethnicity (Weinberg et al. 2001). Our problems do not hinge on their success.

4 The first problem: the data obtained is unreliable

In this section, we raise our first concern about such reasoning. We illustrate by looking at a specific case, inspired by an argument given by Jeff McMahan (2002).

[4] We thank an anonymous reviewer for pressing this point.

[5] It seems plausible, though, that a claim defended by reasoning from examples would be taken to be *better* supported when the judgements of others, and not merely the reasoner's own, fall in its favour. This may be why, when the method is described, writers tend to employ the pronouns 'we' and 'our' when referring to the judgements (Kagan 1988, p.5), (Kamm 1993, pp.7-8). The idea that the judgements of others are involved in the method is also a presupposition of the experimental philosophy literature, which would be harder to motivate if the relevant judgements were not expected to be shared (see, for instance, Weinberg et al. 2001).

4.1 McMahan on the ethics of killing

Consider the thesis:

(TRI): *ceteris paribus*, the fundamental determinant of the degree to which a being may morally permissibly be killed is the strength of that being's time-relative interest in continuing to live: in particular, *ceteris paribus*, as the strength of the being's time-relative interest in continuing to live increases, the degree to which that being may morally permissibly be killed decreases (McMahan 2002, pp.194, 203, 235).[6]

One could argue for *TRI* by presenting comparison-classes of cases, as McMahan does when he invites us to imagine that an agent, Raskolnikov, faces four pairs of people, and must kill exactly one person (i.e., one 'case') within each pair. We are told in advance that, somehow, '[t]he background and circumstances of each case are the same', and that '[e]ach case can be stated simply by describing the person killed, each of whom is given a name that indicates his salient characteristic [...]

Bright was a person with exceptionally high cognitive and emotional capacities that made possible for him an unusually high level of well-being.

Dull was the same age as Bright but was constitutionally dim-witted and stolid. There was thus a range of goods—including engagement in rich, complexly and subtly layered personal relations, the experience of intense, refined aesthetic states, and so on—that were accessible to Bright but from which Dull was by nature excluded.

Cheerful was, as economists say, an efficient converter of resources into well-being. Whatever happened, he managed to find the bright side.

Melancholy was perpetually lugubrious and depressive. Whatever happened, he derived little satisfaction from it.

Fortunate's circumstances were highly favorable. He was good-looking, wealthy, socially powerful, and physically robust and healthy.

Unfortunate was plagued by bad luck in all the corresponding areas of his life.

Young was twenty years old.

Old was ninety.' (McMahan 2002, p.234).

McMahan writes of this collection that '[w]hat should be conspicuous is that one

[6]The strength of one's time-relative interests in continuing to live is, in turn, taken to be a function of two variables: how much welfare one's future can reasonably be expected to contain, and the strength of the prudential unity relations in which one stands to oneself at different times. The strength of these prudential unity relations consists in the extent to which the 'physical, functional, and organizational continuity' of one's brain supports one's 'psychological unity over time' (McMahan 2002, p.193). *TRI* strengthens the view McMahan in fact defends, which restricts the claim to beings below a threshold of 'cognitive capacity'.

victim in each pair has a significantly stronger time-relative interest in continuing to live than the other', and hence that 'it may seem that...Bright's possession of higher cognitive and emotional capacities must make his killing by Raskolnikov more seriously objectionable than the killing of Dull' (McMahan 2002, p.235).

Suppose that a reader responds to each case by judging that it would indeed be less morally objectionable to kill the person from the pair with the weaker time-relevant interest: Dull instead of Bright, Melancholy instead of Cheerful, Unfortunate instead of Fortunate, and Old instead of Young. We might then take these four judgements to provide inductive evidence in favour of the extreme claim, *TRI*.

But how does the reader make these judgements? *Prima facie*, it is natural to think she makes them in part because her credence that, say, Bright may morally permissibly be killed, conditional on Bright's having those features that the reader managed to fix, is less than her credence that Dull may morally permissibly be killed, conditional on Dull's having those features that that reader managed to fix. The features that the reader managed to fix may include how much pain the agent is in, the affective relation (such as fearing) in which the agent stands to their own death, whether the agent has any dependents (and the nature of their relationships with those dependents), *etc.* And suppose her judgements in the other three cases are also based, at least in part, on credences in this way. This is a natural suggestion: when asked to make a judgement about a case, we will rarely be fully certain what we think; but we will make a particular judgement providing our credence in it is sufficiently high. This credence will in turn be based, in part, on our past experience of similar such cases, as well as on other non-empirical beliefs about how the variables in question relate to one another.

However, suppose that the reader is, unbeknown to her, the kind of snob you really do not want to meet in a pinch. In particular, suppose she sets her credences in the way she does *only because*, on her experience, people with strong time-relative interests like Bright are *more* likely to have a *high social status*, and those with a *high social status* are *less* likely to be such that one may permissibly kill them. Moreover the reader does not fix the value of this variable (call it *the status variable*) in her cases. Then, once this reader *does* fix the value of the status variable across both cases in the pair, she might well not only *not* rationally judge that (*TRI*) is true; she might rationally judge that things are *exactly opposed* to how (*TRI*) says they are. That is, for both those with high social status, and for those without it, the reader might well rationally judge that the *stronger* one's time-relative interests in continuing to live, the *greater* the degree to which one may morally permissibly be killed.

The reason is that the reader's credences might well exhibit Simpson's paradox. Let A be the response variable, so that a possible killing receives value a if it is morally permissible and a' if not; let B be the treatment variable, so that a possible killing has value b if the victim has strong time-relative interests in continuing to live; and let C be our partition variable, so that a possible killing has value c if the victim

has a high social status, and c' if they do not. Then our reader's credences from the previous paragraph might be exactly like the probabilities in the example of the kidney stones, so that there is an aggregate-level negative association between b and a—that is, inequality (1) holds—and that negative association is reversed for every subpopulation of the variable C—that is, inequalities (2) and (3) hold. And this is so exactly because, across the reader's credences, b is strongly positively associated with c—that is, inequality (5) holds—which is, in turn, strongly negatively associated with a—that is, inequality (4) holds.

In such a case, the data yielded by the reader's judgements that favour (*TRI*) are unreliable. Once the reader considers cases fleshed out with detail on the formerly omitted status variable, her judgements are reversed, and reversed *regardless of the value given to that variable*. The problem facing the reasoner from examples is quite general. If that reasoner also presses the analogy to controlled scientific experimentation, then since the reasoning now lacks reproducibility and objectivity, it is also *by its own lights* deficient.

We need not insist on that interpretation of the variable C for our point to go through. For example, we might think it more plausible that the omitted variable C has the value, c: *not having post-reflectively and freely consenting to die for good reasons*. Call this *the consent variable*. Alternately, let C have the value: c: *being a lovable living thing, of a kind capable of rich experiences and relationships, for which it is repugnant even to consider the question of what the conditions are under which it may morally permissibly be killed in an analytical manner*.

4.2 Conditions for the first problem

When the problem we have described obtains, the data provided by an instance of reasoning from examples become unreliable. Under what conditions does it obtain? Three conditions are individually necessary and jointly sufficient.[7]

Condition I: That the cases are underspecified by the reader with respect to some apparently relevant variable C; that is, described at a level of detail that leaves C without a value.

Condition II: That the reader's judgement on the underspecified case is informed, at least in part, by certain of her background beliefs or credences about how the variables in the case relate.[8]

Condition III: That the omitted, apparently relevant variable C is one that, according to the reader's relevant background beliefs or credences, is associated with both

[7] It might be reasonable to weight our assessment of the evidential force of a given instance of reasoning from examples by the complement of our credence that these conditions are all met for the total group of readers whose judgements furnish the evidence.

[8] As we will explain in more depth below, this requires no commitment to an implausible model of judgement on which the reader judges, for instance, that the probability that the action in question is morally permissible is equal to some value x. Additionally, the relevant background doxastic states need not be transparent to the reader.

the treatment variable B and the response variable A in the reversal-paradoxical way exemplified by (for instance) inequalities (1) to (5).[9]

The first condition constrains the examples the reader considers; the remaining two her background doxastic states. Hence, whether the first problem arises is largely a matter of how the reader *takes* the variables to relate. It is not *directly* a matter of how they in fact relate, nor of the truth-value or modal status of the hypothesis the reasoner defends; although of course these may bear on how the reader takes the variables to relate.

Take Condition I. It has long been known that the results yielded by reasoning from examples in philosophy can depend on the level of detail at which the cases involved are fleshed out. Williams' *Experiment* case in his essay 'The Self and the Future'—in which two people A and B undergo a process that might swap their bodies, before one is given \$100,000 and the other is tortured—provides a classic illustration of this phenomenon (Williams 1973).[10] Furthermore, knowing this, those constructing the cases may deliberately (encourage their reader to) flesh them out to that particular level of detail required to elicit their desired result (Williams 1973, p.62). The threat of association reversal in reasoning from examples—which we illustrated using McMahan's examples and the hypothesis (TRI)—is a special case of the problem Williams identifies, where at least one omitted variable stands, according to the reader, in certain kinds of relations to the variables under explicit discussion by the reasoner.

Say a variable is *omitted* from an example if the example is described at a level of detail that fails to specify a value for that variable, and *included* otherwise. Then Condition I is that the reader omits a particular apparently relevant variable, C.

Take Condition II. For it to be so, it is not required that the reader has credences, nor that her judgements are based on her credences, nor that these credences exhibit the association reversal. The obtaining of association reversal across the reader's credences provides a vivid illustration of our worry. But even if the reader lacks credences, our point stands if the reader's *background beliefs* about the relationships that obtain between the variables involved in the cases exhibit the association reversal structure, and her judgements are made at least partly on the basis of these beliefs. In order for these beliefs to exhibit the association reversal structure, all that's required is that the reader believes—in the traditional, all-or-nothing sense—that the variables in question stand in the set of relations captured by the inequalities (1) through (5).

Moreover, if the doxastic states are reversal-paradoxical, that may be so for a variety of reasons. It need not be so solely because she has observed the (short- or long-run)

[9] On the parenthesis: inequalities (1) to (5) illustrate just one way in which the reversal paradox might manifest. Instead, for example, a *positive* overall association may be reversed for all cells within some relevant partition. That would satisfy Condition III too.

[10] The same phenomenon is also taken by Gendler and Hawthorne (2005) to apply to Goldman's 'fake barn' case (Goldman 1976).

frequencies, or solely because she has discovered that the objective chance distribution is reversal-paradoxical. She may have observed no relevant frequencies, and know little about the objective chance distribution. It is possible that the reader's credences exhibit association reversal in the absence of worldly association reversal, and *vice versa*.[11]

In particular, when the hypothesis defended by the reasoner concerns moral features such as the permissibility of killing, it is plausible that the reader's background doxastic states about the relevant variables will be informed in part by the nature of the concepts the reader possesses.[12] The reader's credences or background beliefs may also be guided by—and guide—her espousal of certain principles, conceptual analysis (rather than merely the concepts) and other normative considerations. Even then, though, those doxastic states could have the reversal-paradoxical shape, regardless of what grounds, or in fact generated, those states. All that's required is that the reader's doxastic states do indeed exhibit the shape, and that they at least partly inform the judgement the reader makes about the examples.

This requirement would not obtain for readers who make judgements about cases on the basis of principles alone.[13] It could obtain, however, for readers who made such judgements on the basis of their *credences about* such principles. It is beyond the scope of this paper to settle which of these is the better model of reasoning in this setting. But there is some relevant empirical evidence. Schwitzgebel and Cushman (2012, 2015) report that the moral judgements of philosophers are shifted by manipulations of order and of framing that the principles those philosophers endorse treat as irrelevant; that these shifts are no smaller than those exhibited by non-philosophers; and that they persist despite training. We should be careful about what this establishes. It does not establish that the relevant background doxastic states are or can be reversal-paradoxical. At best, it provides defeasible empirical evidence against the model on which expert case-judgement is the unmediated application of principle (and so immune by its nature to anything able to exhibit reversal). Set that model aside, and the reader's verdict is plausibly informed, at least in part, by beliefs about how the morally relevant features co-vary. Being guided by principles is no bar to leaning, in the particular case, on beliefs of just that kind. Principle-guidance and reversal-prone background belief are not mutually exclusive.

Relatedly, we take it that the reader's output judgement, when considering the examples raised by the reasoner in favour of *TRI*, is itself *non-probabilistic*: for each case-pair, the reader makes—or does not make—a judgement about the relative degree to which the individuals involved may morally permissibly be killed.

---

[11] For instance, the objective chance distribution may be reversal-paradoxical with respect to three variables without the reader's ever discovering this. Even if worldly reversal obtains because the relevant frequencies exhibit reversal, and the reader observes those frequencies, the reader's credences need not exhibit reversal. One reason this is so is that experiences of frequencies cannot produce credences all on their own without bridge principles.

[12] For instance, Machery notes that '[c]oncepts give rise to judgements that some things, including the situations described by philosophers' cases, fall or do not fall under them. These judgements are, at least in part, determined by the content of the concepts' (Machery 2017, p.233).

[13] We thank an anonymous reviewer for pressing us on this point.

By contrast, we do not take it that the reader judges that there is some *probability* with which the individuals involved may respectively be permissibly killed. The non-probabilistic nature of the reader's output judgement also makes room for the likely scenario that, for some case-pairs, the reader comes to *no* judgement at all about whether or not it is more permissible to kill one individual than the other in the *TRI* case. Our point is also compatible with a picture on which reasoning from examples in philosophy is primarily the provision of valid arguments containing at least one counterfactual conditional as a premiss (Williamson 2022). The reader's judgements may well concern certain counterfactual claims.

Take Condition III. This requires the reader's doxastic states to have the reversal-paradoxical shape. Of course, not all such states may plausibly have this shape in response to a case. For example, we may believe—and be unable not to believe—that all cases with the Gettier-property will also have the lacking-knowledge property, and hence our conditional probabilities will be too extreme to exhibit the inequalities (1) through (5) with respect to those variables and any partition variable.[14] But it is hard to imagine how the reader could come to a judgement about the degree to which different living things may morally permissibly be killed without at least some appeal to suitable background credences or beliefs. These credences and beliefs may be shaped by the reader's past experiences. But they may also be shaped by how she has otherwise come to think, perhaps through *a priori* considerations, about the relations between the variables in question. In the case of *TRI*, the reader's conditional probabilities will be nowhere near as extreme as in the Gettier case.

The scope of our worry here is thus quite general. We illustrated it with reasoning from examples used to justify a *ceteris paribus* claim about metaphysical determination relations, *TRI*. But it also applies to reasoning from examples used to justify other kinds of claim (see §4.3).

The worry also applies to reasoning from examples used merely as a so-called 'intuition pump'. For example, the reasoner may invite the reader to consider *solely* the example of McMahan's *Melancholy* character, so described, and ask her to come to a judgement about how morally permissible it is to kill them. If the reader's credences are reversal-paradoxical, then she will come to the judgement that it is morally permissible to kill them to a relatively high degree. But as soon as the reader considers the same character as having a high social status, or as not having a high social status, she will come to the judgement that it is morally impermissible to kill them.

Condition III will obtain only if, according to the relevant background beliefs of the reader, the association between the treatment-value b and the response variable a has at least two exceptions on each side (or: if the agent is represented as having credences, the agent's credences are such that neither $P(a|b)$ nor $P(a|\neg b)$ are zero or

[14] As it turns out, there are plenty of 'deviant realisations' of Gettier cases (Williamson 2022; Ichikawa and Jarvis 2009; Malmgren 2011; Horvath 2015). These are more specific versions of Gettier's original cases in which the agent lacks a justified true belief that isn't knowledge either by turning out to have knowledge after all, or by lacking justification.

one). That is, for instance, unless the reader takes at least two beings with strong time-relative interests to be such that they may permissibly be killed, and at least two beings with weak time-relative interests to be such that they may not, she cannot have reversal-paradoxical beliefs about those variables. This follows from the mathematical properties of the reversal paradox. These properties can suggest mitigating strategies, so it is to them that we now turn.

4.3 Conditions for association reversal, and mitigatory strategies

Consider the following result (Blyth 1972; Lindley and Novick 1981; Mittal 1991):

> *Necessary Condition for Association Reversal*: Suppose that, at the level of the total population, the treatment variable B is associated with the response variable A, and C is some further partition variable. Then association reversal occurs within all cells of C only if C is associated with both A and B.

By taking the contrapositive of this claim, it follows that if any partition variable C is independent of either A or B, or both, association reversal with respect to C cannot occur.

For instance, if the status variable from §4.1 is—within the world—independent of either time-relative interests, the degree to which one may morally permissibly be killed, or both, then worldly association reversal with respect to it cannot occur. Of course, that the status variable is so independent does not entail that the reader *takes it* to be independent. But if she does, then—regardless of whether it is—the first problem is avoided.

Hence, the reasoner need not ensure that her reader fleshes out the examples with detail about the value of a relevant variable C in order to show that no association reversal can occur with respect to C across that reader's judgements. However, on pain of reversed judgements (and unreliable data) for those apparently relevant variables *not* shown to be independent of either the treatment or the response, the reasoner must ensure that the reader fixes the values of those variables when considering the examples. For instance, if the reasoner does not show that the status variable is independent of either time-relative interests, or the permissibility of killing, then the reader's judgements may reverse once the examples are fleshed out with respect to the status variable.

If the reasoner defends a *ceteris paribus* hypothesis, then it is incumbent on *her*, not the reader, to ensure that the reader fixes the values of the right variables in generating the data of the contrast arguments. After all, the reasoner asks the reader to fix characteristics other than those in which (say) *Bright* and *Dull* are *explicitly* said to differ. But on pain of association reversal and unreliable data, these characteristics ought to include every apparently relevant variable for which a reversal could occur; that is, every apparently relevant variable not already shown to be independent of

the treatment or the response variables.[15]

Whether there exists a non-trivial necessary *and sufficient* condition for association reversal depends on which question is asked.

One question asks whether, for a given treatment variable B, response variable A such that B and A are associated at the level of the total population, *and a given* partition variable C, there is a non-trivial set of conditions that obtain iff association reversal obtains within all cells of C. To be sure, the Necessary Condition for Association Reversal above provides solely a necessary condition of this form. But there is no necessary and sufficient condition like this. Yashaswini Mittal observes that although such a condition would be ideal, none that is interpretable has been found, with certain technical biconditionals being easy to write down but resistant to non-trivial interpretations (Mittal 1991, p.167).[16]

Another question asks whether, for a given treatment variable B and a given response variable A such that B and A are associated at the level of the total population, there exists *some* (possibly irrelevant) variable across all of whose cells that association is reversed. Our colleague (forthcoming) shows that, for a finite population of cases, such a reversing variable exists iff at least two of the things bearing the treatment value run counter to the association, and at least two of those lacking it do so as well; equivalently, iff the generalisation defended and its companion each admit at least two exceptions. For instance, for *TRI* to be reversed with respect to *some* variable or other, it is necessary and sufficient that at least two things with strong time-relative interests may permissibly be killed, and at least two things with weak time-relative interests may not be.[17]

This necessary and sufficient condition explains why Gettier's reasoning is plausibly construed as immune to the first problem: the reader will just not *take* there to be any exceptions to the generalisation that any individual with the Gettier-property has the knowledge-lacking property.[18] But of course, the reader may take there to be exceptions even to universal generalisations (or stronger kinds of claim) even if she is wrong, and even if there are none. At least two exceptions either side guarantees association reversal with respect to some variable or other. Provided the reader making

[15] To be sure, a further condition is required if the reader's judgements concern the direction and magnitude of a causal relationship, or a metaphysical grounding relation in the image of causation à la Schaffer (2016): each apparently relevant variable for which detail is provided must also be neither a mediator nor collider according to the reasoner (see §6.2). Otherwise, the reader's judgement concerns only what the reasoner takes to be the *direct*, and not the *total*, influence of the treatment on the response.

[16] For instance, Mittal's own homogeneity result shows that reversal requires the within-cell success rates of the treated and untreated groups to overlap, since row (or column) homogeneity suffices to preclude it; again, a necessary condition, but not sufficient (Definition 3.1, Mittal 1991). Candidate biconditionals fail Mittal's non-triviality requirement. For instance, Sarkar and Bandyopadhyay's Theorem 3 roughly states that reversal holds just in case the odds ratios of the two relevant sub-tables fall on one side of the unity, while that of their amalgamation does not (Sarkar and Bandyopadhyay 2021, p.8). But this is just the kind of uninformative result Mittal had in view in 1991: it is an odds-ratio *restatement* of the definition. The same can be said of the biconditional given in Bandyopadhyay *et al* (2015) in terms of success-rates.

[17] The requirement of two reflects the discreteness of the cases, but the result is extensible to the continuous case.

[18] Excepting deviant realisations for the sake of illustration.

the judgement does so partly on the basis of reversal-paradoxical doxastic states that take the relations between the variables to be much weaker, the worry goes through even for reasoning for true universal generalisations and necessary claims.

Of course, not just *any* reversal across the reader's doxastic states is particularly worrying. Reversal across the reader's doxastic states with respect to *irrelevant* or extraneous variables does not seem to undermine the evidential force of the reasoning. There is a sense in which we need not worry how things look, for the reader, across the cells of variables—such as hair colour—that are irrelevant to the variables at hand.

Thus, the guidance already given concerning the Necessary Condition is our main suggestion for how to mitigate the first problem we raised: ensure the reader considers examples with detail on all apparently relevant variables not shown to be independent of the treatment or the response.

Before turning to our second problem, consider a natural objection. We have shown that reasoning from examples is threatened by Simpson's paradox. We have not shown that a 'typical' reader of a contrast argument such as McMahan's would in fact fall prey to it. The objection is correct, and we defend here only the weaker. Still, there are some reasons in favour of the stronger in the case of the McMahan argument.

Since McMahan's cases do no more than label each victim and stipulate a shared background, some relevant variable is likely to go unfixed for the reader (Condition I). Moreover, as we have argued, the reader's verdict will likely rest on background doxastic states about those features and their relations (Condition II). Condition III requires that the reader's judgements reverse across the unfixed variable's values. Why would this obtain for a *typical* reader?

Since *TRI* is a *ceteris paribus* claim, she will likely take it to admit exceptions on each side—some strong-interest beings she takes to be permissibly killable, some weak-interest beings she takes not to be—and given at least two on each side, the necessary and sufficient condition above guarantees that her credences reverse with respect to *some* variable or other. Whether that variable is one she counts as relevant or extraneous turns partly on her more general beliefs about the structure of that part of the moral domain concerning the permissibility of killing. McMahan asks her to vary time-relative interests while keeping every other relevant variable fixed, but Wilson argues that the domain of ethics, like the complex systems in which we are 'biologically and socially embedded', cannot always be analysed one level at a time (Wilson 2021, pp.74-5). A reader who sees the features morally relevant to the permissibility of killing in this way—as interacting, rather than as separately assessable—is one whose credences tie *those features* (and not extraneous ones such as hair colour) to both time-relative interest and permissibility. Any reversal in her credences is correspondingly unlikely to fall on an extraneous variable. Whether a 'typical' reader

does believe the factors determining the permissibility of killing to be a complex system is beyond the scope of this paper. But on these assumptions, the three conditions are jointly likely to be met for such a reader of the McMahan argument.

5 The second problem: how to intervene does not follow from the probabilities

Certain philosophers would like their conclusions to justify certain courses of *action*—particular interventions, policies, *etc.*—over others (Wilson 2021). This is surely a laudable aim. But association reversal may play out *in the world*, rather than merely across the reader's doxastic states. Consequently, Simpson's paradox poses a threat to practical decisions justified by reasoning from examples.

5.1 Causal structure and Simpson's paradox

It has been shown that one and the same reversal-paradoxical data set—such as the one corresponding to associations (1) through (5) in §2—may be generated by different causal structures. But depending on the causal structure, one and the same intervention on the treatment variable B may have different effects on the response variable (see Chapter 3, Spirtes et al. 2000).[19]

In light of these results, we argue for the following claim. Even a reasoner who has evaded our first problem must consider how the variables under consideration are causally related to one another if she wishes to use the reasoning to support action. This injunction is especially pressing for philosophico-scientific reasoners, for this is exactly what is done by those engaging in the very kinds of experimentation to which they take their reasoning to be analogous. But the claim holds even for reasoners from examples who do not uphold that analogy.

To see why, consider that in the *confounded* causal structure, while being treated (b) has a positive causal influence on recovery (a), having small kidney stones (c) has a positive causal influence on whether the patient is treated (b), and a negative causal influence on whether they survive (a); and no other causal relations hold between these variables. Assuming the confounded structure holds, the partition-level (positive) direction of association will be the one to trust for interventions, and evaluating the effect of the treatment in groups homogeneous with respect to kidney stone size will reveal the effect of the treatment.

In the *mediated* causal structure, by contrast, being treated (b) directly positively influences recovery (a), but indirectly negatively influences recovery by directly causing something—namely, stubborn, hard to remove small kidney stones (c)—which in turn are so counter-conducive to recovery (a) that the overall negative association seen in inequality (1) above results; and no other causal relations hold between these

[19] An *intervention* on the treatment variable sets its value in a way that is not accompanied by a change in the value of any variable on which that treatment variable depends, such as confounders (Woodward 2003, p.47).

variables (see, too, the thrombosis case in Hesslow 1976).[20] Assuming the mediated structure holds, the aggregate-level (negative) direction of association will be the one to trust for interventions, and evaluating the effect of treating in groups homogeneous with respect to kidney stone size *masks* the total effect of the treatment by revealing only the direct positive effect.

Importantly, some reversal-paradoxical data sets may lack a causal diagnosis.[21] We need not take a stance on this claim here. For there is general agreement in the literature that the right practical response to Simpson's paradox is *typically* the identification of, first, the causal structure that generated the reversal-paradoxical data set, and second, the set of variables, if any, for which one ought to *adjust* when calculating the effect of the treatment on the response (Bandyopadhyay et al. 2015, pp.66-7), (Pearl 2009).[22]

5.2 Boonin on the non-identity problem

David Boonin argues that the non-identity problem can be solved by accepting a certain initially counterintuitive claim: that any agent who performs a certain kind of *generative* act—an act that brings at least one living thing into existence—does something morally permissible (Boonin 2008). Say that an act is a *Boonin act* just if it brings into existence a 'disabled' living thing who goes on to have a life that's worth living, rather than brings into existence a non-disabled living thing who goes on to have an even better life; and it is performed by the agent at least partly out of convenience. Boonin argues for the following universal generalisation:

(BH): All Boonin acts are morally permissible.

In defending (BH), Boonin reasons from examples, including the examples of individuals he calls Wilma, Fred and Barney (Boonin 2008, pp.130, 150). For instance, suppose one is trying for a child, but knows that unless one refrains from smoking one's favourite cigarettes until conception, one will engender a disabled child who nonetheless will lead a life that's worth living; whereas if one refrains from smoking until conception, one will engender a non-disabled child with a better life. The Boonin act is, in this case, the act of continuing to smoke, and giving birth to the disabled child.

5.3 Justifying action

We take it that Boonin's reasoning is vulnerable to worldly association reversal that could impugn its ability to justify action.

---

[20] There are other possible structures, such as the *collided* structure, in which the partition variable is causally influenced by both the treatment and the response, and no other causal relations hold.

[21] Suppose you have two bags of marbles that are either big or small and either red or blue, such that in each bag separately, big marbles have a greater red-to-blue ratio than small ones, but once you pour both bags into one box, small marbles have a greater red-to-blue ratio than big ones. Yet size doesn't cause colour, nor *vice versa* (Bandyopadhyay et al. 2011).

[22] One adjusts for C iff one calculates the effect of b on a in each of the subpopulations of C, before weighting those effects by the share of the total population that those subpopulations comprise. The effect of the treatment on the response within each cell is standardly equated to the result of subtracting the proportion of the untreated within that cell who succeed from the proportion of the treated within that cell who succeed. The adjustment computation yields a single number whose sign and magnitude respectively indicate the direction, and the strength, of the effect of the treatment on the response when the data is partitioned by C.

Take one way in which the reversal might play out. Suppose an agent is faced with the choice of whether or not to perform a Boonin act, like the smoker trying to conceive. One motive she may have for performing it is that bringing into existence the disabled child is relatively convenient, and that bringing into existence a disabled living thing is at least no worse than bringing into existence a non-disabled living thing. Call this motive the *disability-tolerant* motive.

But suppose that the phenomena are arranged reversal-paradoxically with respect to the variables B (= whether the act is a Boonin act), A (= whether the act is morally permissible), and C (= whether the act is performed by someone with the disability-tolerant motive); in particular: while there is a *positive* association, overall, between performing a Boonin act b' and moral permissibility a (reflected in inequality (1)) there is a negative association between these variable-values both for those agents *with* the disability-tolerant motive c', and for those agents *without* it, c (reflected in inequalities (2), (3)). All that's required, for worldly association-reversal, is that generative actions done by those with the disability-tolerant motive are other things equal more likely to be morally permissible than relevant alternatives done by those with different motives—inequality (4)—and generative actions that are also Boonin acts are other things equal done by those with disability-tolerant motives—a consequence of inequality (5).[23]

Two qualifications. First, it is not *(BH)*—which is a universal generalisation—that would be subject to association reversal if inequalities (4) and (5) obtained, but rather the weaker claim that, *ceteris paribus*, an act's being a Boonin act is strongly positively associated with its being morally permissible. As we have seen, a generalisation can be subject to association reversal only if it admits of at least two exceptions either side; and no true universal generalisation will admit of exceptions. Still, it is a live possibility (and may turn out) that, unbeknown both to reasoner and reader, (BH) is *false*, and the strongest true claim in its vicinity is the weaker *ceteris paribus* claim. That makes worldly association reversal a risk even for reasoning from examples in favour of universal generalisations.

Second, we are not claiming there is overwhelming reason to think that inequalities (4) and (5), with the interpretations above, are *true*. We are claiming that they are possible, that in the world in which they are so, association reversal is guaranteed, and that as a consequence, reasoners ought to attend to the risk of reversal when using their reasoning to support action. Equally, this is not to say that the present interpretations of inequalities (4) and (5) have nothing in their favour. One reason in favour of the former is that agents performing generative acts out of the disability-tolerant motive are unlikely to subscribe to repugnant disability-intolerant eugenics programs. One reason in favour of the latter is that, as we saw, it explains why an agent faced with such a choice would opt for the Boonin act. Moreover, the motive variable is merely one among many apparently relevant variables for which the problem could arise.

[23] Since we are now focussing on worldly association reversal, the sense of 'likely' employed in this paragraph expresses *objective* probability, and we will assume—for simplicity—that the statistical frequencies perfectly reflect the objective probabilities; although nothing hinges on this. The relevant instance of association reversal here is not (just) across the reader's doxastic states.

Now, the reasoner need not realise that the phenomena in fact exhibit the reversal. Epistemologically speaking, Boonin's omission of the motive variable in his cases—such as the Wilma case—may preclude discovery that such a reversal obtains in the world.

But suppose she does: the doxastic states of the reasoner reflect these facts, and in reasoning from examples with respect to (BH), the reasoner *does* ensure that the reader considers cases at a level of granularity that includes detail on the values of all apparently relevant variables—including C—that have not been shown to be independent of the treatment or the response. The first problem we expressed in §4.1 about unreliable data thus does not apply. Suppose further that the judgements to which both reasoner and reader arrive *disconfirm* BH, running instead in favour of the hypothesis—confirmed by the more granular associations captured in inequalities (2), (3)—that Boonin acts are, *ceteris paribus*, *not* morally permissible. Call the evidence generated by these judgements E.

The problem is that what the reasoner ought to *do* on the basis of the evidence E yielded by the consideration of cases depends on the causal structure that holds between the variables. Since E contains no hypotheses about this structure, it provides no guidance on what she should do.

Assuming the reasoner's preferences over acts track her credences in the natural way, then on the basis of E, she will know that:

> *Refrain Come What May*: refraining from performing the Boonin act is preferable to performing it, regardless of whether one has, or doesn't have, the disability-tolerant motive.

But *Refrain Come What May* supports the preferability of refraining from performing the Boonin act if and only if whether one has the disability-tolerant motive C is not causally influenced by whether or not one performs the Boonin act (and hence, only if C is not a mediator or collider with respect to B and A).

For example, suppose that whether one has the disability-tolerant motive c' *confounds* the effect of b' on a (and hence C is not causally influenced by B). That is, having the disability-tolerant motive c', other things equal, makes one *both* more likely to to perform Boonin acts b', and to act morally permissibly a. Then one can validly infer from *Refrain Come What May* that one indeed ought to refrain from performing the Boonin act: since C is a confounder, it is the within-cell associations of inequalities (2) and (3) that ought to be trusted, and within each cell, refraining is the more permissible course.

But suppose instead that whether one performs Boonin acts *influences* whether or not one has the motive, which in turn influences whether one acts morally permissibly. Then *Refrain Come What May* does *not* license the inference that one ought to refrain from performing the Boonin act. Here C is a mediator, so it is the aggregate association of inequality (1)—which captures the total effect of the act on its permissibility—that is to be trusted; and that association favours performing the Boonin act. Refraining could thus be morally harmful, even though *Refrain Come What May*, and inequalities (2) and (3), are all true.

In general, Pearl has shown that the fact that one action is preferable to another in

each of two mutually exclusive and jointly exhaustive states of the world warrants the inference to the claim that that first action is preferable to the second, *tout court*, only if which state of the world turns out to obtain is *not* causally affected by what one does:

> *Causal Sure-Thing Principle* (Causal STP): Suppose one is considering two acts b, b', and the probability distribution of the partition variable C is not changed by whether one intervenes to choose b', or intervenes to choose b. Then if one prefers b to b' either knowing that the value c of C obtains, or knowing that c does not obtain, then one prefers b to b' (Pearl 2009).

Causal STP licences the inference from, on the one hand, the preferability of refraining b to performing b' for each subpopulation of C, to, on the other, the preferability of b over b', *tout court*, only if C is *not* causally downwind of B (and hence is neither a collider nor a mediator).

To sum up, even if the reasoner 1. possesses the total evidence E output by the consideration of the cases, 2. knows that the phenomena exemplify association reversal, and 3. assents to *Refrain Come What May*, she *still* needs to execute further causal investigation in order to know whether to perform the Boonin act.

The remedy is this. Flesh out one's cases with detail on all apparently relevant variables that have not already been shown to be independent of the treatment or the response. If association reversal appears, nonetheless, to obtain in the world, consult the part of one's evidence—if it exists—bearing on the causal structure that holds between the variables. This causal structure will determine which intervention on the treatment will lead to the desired outcome.

While this remedy applies to all reasoners from examples, it is especially pressing for the philosophico-scientific. Since these philosophers hope their reasoning is analogous to controlled scientific experimentation that leverages Pearl's findings in its methods, if those philosophers would like to use their results as a basis for action, it seems incumbent upon them to abide by those findings.[24]

5.4 Conditions for the second problem

For completeness, here are the conditions for the second problem.

First, the reasoning is put to practical use. Second, in the world—the chances or the

[24] One may worry that the causal inference framework cannot apply to moral reasoning, because moral claims seem predominantly to posit something like metaphysical grounding or determination relations (rather than causal relations). However, a popular family of accounts of the (otherwise quite obscure) notion of metaphysical grounding characterise it as 'metaphysical causation' (Sider 2011, p.145). Schaffer, for example, argues that grounding and causation share multiple important features, are both 'best formalized via structural equation models', and even asks 'whether there is any more than a nominal distinction between grounding and causation' (Schaffer 2016, p.59). If this is the right account, applying Pearl's machinery when deciding whether to perform (for instance) a Boonin act will be entirely appropriate. But even if it is not the right account, many hypotheses defended by appeal to reasoning from examples in philosophy are plausibly interpreted as *involving* the attribution of a causal relation, even if they also attribute other kinds of relations. And this is all it takes for the causal inference framework to be a suitable remedy to the problem of this section.

frequencies—the treatment-value and the response-value stand in an association that reverses across the values of some apparently relevant partition variable C. Third, the reversal underdetermines the causal structure that could have generated it.

The three conditions are jointly the more likely to be met the closer the reasoning lies to practical philosophy. Take them in turn.

The first is met wherever the reasoning underwrites a verdict that itself licenses or forbids conduct: in applied ethics, in bioethics, in the ethics of killing and of procreation, in any domain in which a judgement of permissibility is a judgement about what one may do.

The second is the more likely the more heterogeneous the domain over which the generalisation ranges, and the more the treatment-value co-varies, across that domain, with other features bearing on the response. Where bearing the feature under test is systematically bound up with much else (as strong time-relative interests are bound up with cognitive sophistication, social embedding, and the rest) there is the more room for some such concomitant to be a variable across whose values the aggregate verdict reverses. Since philosophical questions often concern complex and heterogeneous phenomena, our prior that the second condition is met, for many questions, ought to be considerable.

The third condition is more likely met in situations in which the causal mechanisms involving the variables in question are as yet poorly understood. But importantly, philosophers may possess defeasible inductive evidence in favour of certain causal structures over others, and indeed further reasoning by examples may provide strong but defeasible evidence in favour of certain such structures.

6 Two responses to our problems

We use the remainder of the essay to consider two responses to the problems we have raised.

6.1 Response 1: Just go fine?

Consider the practical problem of how to act, considered in §5.

It might seem that the right response to association reversal is to find the finest relevant partition variable $\mathbf{C}$—generated by taking the Cartesian product of all the other relevant partition variables—examine the associations that hold across its cells, and perform the action they support. In the kidney-stone case, this means attending to the cells given by every combination of values of C and D (those under fifty with small stones, and so on). The finest relevant partition variable need neither exhibit reversal nor have all its cells agree in direction with the aggregate; but the adjustment measure, as defined above, quantifies the effect of the treatment on the response across all of its cells. Perhaps, then, the right response is always: find the

finest relevant partition variable; adjust one's estimate of the effect of the treatment on the response by appeal to it; and perform the action that adjustment supports. Call this the *finest relevant partition strategy*. There are three worries with it.

The first is epistemic. We will have reached the finest relevant partition variable only if it captures, in its values, the values of all other relevant partition variables. But we will rarely know that a candidate variable does so. Unless one is confident that one's initial set of relevant variables is complete, one cannot be confident that their Cartesian product yields it.

The second is an aspect of the reference class problem, and is independent of the first (Hájek 2007). Other things equal, finer partition variables are more relevant to the effect estimated. But the finer the variable C, the fewer things fall into each reference class into which it divides the population, and the less the success-frequencies in each class resemble, or can inform, probabilities of success. As Wallmann (2017) shows, when only extremely small numbers fall into a class, the frequencies lack a feature canonical of probabilities: the ability to guide our actions and expectations. In the limit you are alone in your cell, where the success rate is either 0 or 1. This is too extreme to guide us, and admitting two, three, or four others does not relieve the problem, for the same reason.

We have already met the third, and most severe. Pearl has shown that the right response to association reversal typically requires identifying the causal structure underlying the data, and the set of variables— if any—for which one ought to adjust. But this set need not include the finest relevant partition variable, let alone consist of it alone. Hence the finest relevant partition strategy fails.

6.2 Response 2: *ceteris paribus* to the rescue?

Consider the first problem, of unreliable data, from §4. It might seem that including a *ceteris paribus* clause in the hypothesis automatically blocks the threat of association reversal.

To see how the response goes, recall *TRI*. *TRI* entails that, all other (apparently relevant) things equal, an increase in the strength of a being's time-relative interests leads to a decrease in the degree to which it may morally permissibly be killed. Now, there is a single partition variable $\mathbf{C}$ capturing exactly all the possible combinations of values for all the other apparently relevant variables C_1,...,C_n. So *TRI* makes a claim only about the associations holding within groups homogeneous in $\mathbf{C}$, and none about associations across any other groups. If *TRI* is an exceptionless cp-law, it says the association holds for every such group; if non-exceptionless, it says it holds for some but not all. Either way, *TRI* postulates no aggregate-level association, nor any across the cells of a variable of different granularity to $\mathbf{C}$.

A defence now emerges. Either association reversal obtains across the reader's credences with respect to $\mathbf{C}$, or it does not. If it does, there is no problem: it is the

direction of association the reader judges across the cells of **C**—not the aggregate, nor any other—that should carry evidential weight, since only those case-pairs hold everything equal as the clause instructs. If it does not, there is no problem. Therefore there is no problem. To illustrate: suppose the status variable is the only apparently relevant one, so that **C** is the status variable, and *TRI* claims a negative association both for those of high status and for those of low. Reversal may obtain across the reader's credences with respect to status, making her judgements volatile depending on whether the cases are fleshed out with respect to **C**. But this does not matter, since only her judgements about cases fleshed out with respect to **C** bear on *TRI*, and these, *ex hypothesi*, are not volatile. If they support *TRI*, so much the better for its proponent; if not, so much the worse; but no problem arises from reversal across her credences.

The defence fails, for three reasons which, interestingly, approximate those facing the finest relevant partition strategy.

First, the reader evaluates *TRI*, rather than some other hypothesis, only if she fleshes out the cases with respect to exactly the finest apparently relevant partition variable **C**; so if they are not fleshed out with respect to **C**, *TRI* is not even tested by her reasoning, let alone supported. And we have seen that cases fleshed out with respect to a given **C** may elicit judgements that shift in direction and magnitude once fleshed out with respect to a variable of only slightly different granularity. It is therefore not enough to invoke a *ceteris paribus* clause without attempting to spell out the apparently relevant variables and ensure the reader fleshes out the cases with respect to them. Crucially, though, the reasoner's uncertainty about whether she has succeeded need not entail that her hypothesis goes wholly unsupported. If she is unsure whether she has exhaustively spelt out the apparently relevant variables, or whether her reader has fleshed out the cases on them, she ought to weight the support the judgements provide by her credence that both tasks have succeeded. Inductive support is continuous.

Second, suppose this worry resolved. The subject-matter philosophers model is highly complex. Wilson defends modelling the subject-matter of ethics as a complex system, so that any ethical question of genuine interest involves a vast multitude of apparently relevant variables (Chapter 4, Wilson 2021). Since **C** is their Cartesian product, its values undergo a combinatorial explosion as that multitude grows, making the case-pairs the reader must consider—each homogeneous in **C**—intractably large. In practice the proponent will retreat to a non-exceptionless law, and will instead hold the association for some but not all values of **C**. But the onus is then on her to specify or approximate the values for which it holds, and to say why it holds for those and fails for the rest. Moreover, the more values **C** has, the fewer comparable cases the reader will have experienced resembling those in each pair; and since such experience partly informs the credences and beliefs grounding her judgements, her intuitions about the response will tend to weaken accordingly. Required to hold fixed, say, social status, post-reflective consent, the affective relation to one's death, the presence of

dependents, and many other variables, the reader may have only very weak intuitions about the resulting cases, having met so few of their like.

The third worry is fatal to the strategy in its present form. Even the finest apparently relevant partition variable **C** may be a mediator or a collider; and if it is, even judgements that correctly track the direction of association across its cells may be radically misleading. Suppose **C** is just the consent variable, which mediates (in the sense of §5.1) the influence of time-relative interests on the permissibility of killing in the following sense: an increase in time-relative interests directly decreases permissibility, but also tends to induce appropriate consent, which sharply increases permissibility. This may yield a positive association overall yet, in accordance with *TRI*, a negative one for each value of **C** (both for those who have consented and for those who have not). Here **C** is a mediator, and the direction of association across its cells does not indicate the direction of the treatment's influence on the response.[25]

7 Conclusion

Philosophers very often argue for their conclusions by reasoning from examples. They present examples that they describe with a certain level of detail, elicit judgements about these examples on the basis of the descriptions, and claim that these judgements support their conclusions.

However, when we consider what leads to our judgements in these cases, we see that, unless certain conditions are met, they might be unreliable. After all, we make these judgements partly by appeal to our credences or background beliefs. Certain of these doxastic states relate the variables at hand on *a priori* grounds; others may be grounded in past experiences. Yet others may condition on the kinds of details given in the descriptions of the cases. Regardless of their grounding, these states might be subject to association reversal. That is, there might be a relevant feature of these examples that isn't specified in their description such that, when it is not specified, our doxastic states lead us to judge in one way, but when it is filled in, they lead us to judge in the opposite way, regardless of the way it is filled in. We have argued that this renders these judgements unreliable, and we have explained how a reasoner might render them more reliable. What is required to make them more reliable depends on whether we wish to use the conclusions to guide action.

[25] One might try to salvage the clause by reinterpreting it. A comparative cp-hypothesis holds that an increase or decrease in B leads, or is liable to lead, to one in A exactly when all the other apparently relevant variables $\mathbf{C}$ $(= C_1, ..., C_n)$ that are not causally influenced by B—though they may influence it—remain fixed (Schurz 2002). So construed, the hypothesis concerns not the cells of the finest apparently relevant partition variable, but those of the finest such variable that is neither mediator nor collider, though it may be a confounder. Even then, however, the reasoner can test the claim only by ascertaining which variables to hold constant, which she can do only by thinking causally about how they interrelate; so on this interpretation too, a naïve consideration of arbitrarily specific cases will not do.

References

Bandyopadhyay, P., D. Nelson, M. Greenwood, G. Brittan, and J. Berwald. 2011. The logic of simpson's paradox. *Synthese 181*(2): 185–208 .

Bandyopadhyay, P., R.V. Raghavan, D.W. Dcruz, and G. Britten. 2015. Truths about simpson's paradox: Saving the paradox from falsity, In *Logic and Its Applications*, eds. Banerjee, M. and S.N. Krishna, Volume 8923, 58–73. Berlin / Heidelberg: Springer.

Blyth, C.R. 1972. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association 67*(338): 364–366. https://doi.org/http://dx.doi.org/10.1080/01621459.1972.10482387 .

Boonin, D. 2008. How to solve the non-identity problem. *Public affairs quarterly 22*(2): 129–159 .

Cartwright, N. 1979. Causal laws and effective strategies. *Noûs 13*(4): 419–437 .

Crain, G. 2023. Three shortcomings of the trolley method of moral philosophy. *Journal of Ethics and Social Philosophy 26*(2): 420–443 .

Dretske, F.I. 1970. Epistemic operators. *The Journal of Philosophy* 67: 1007–23 .

Elster, J. 2011. How outlandish can imaginary cases be? *Journal of Applied Philosophy 28*(3): 241–258 .

Frankfurt, H.G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66: 829–839 .

Gendler, T.S. and J. Hawthorne. 2005. The real guide to fake barns: A catalogue of gifts for your epistemic enemies. *Philosophical Studies* 124: 331–352 .

Gettier, E. 1963. Is justified true belief knowledge? *Analysis 23*(6): 121–123 .

Goldman, A.I. 1976. Discrimination and perceptual knowledge. *The Journal of Philosophy* 73: 771–91 .

Hesslow, G. 1976. Two notes on the probabilistic approach to causality. *Philosophy of Science 43*(2): 290–292. https://doi.org/http://dx.doi.org/10.1086/288684 .

Horvath, J. 2015. Thought experiments and experimental philosophy, In *The Palgrave Handbook of Philosophical Methods*, ed. Daly, C., 386–418. London: Palgrave Macmillan.

Hájek, A. 2007. The reference class problem is your problem too. *Synthese 156*(3): 563–585. https://doi.org/http://dx.doi.org/10.1007/s11229-006-9138-5 .

Ichikawa, J. and B. Jarvis. 2009. Thought-experiment intuitions and truth in fiction. *Philosophical Studies 142*(2): 221–246 .

Kagan, S. 1988. The additive fallacy. *Ethics 99*(1): 5–31 .

Kamm, F.M. 1993. *Morality, Mortality, Volume I: Death and Whom to Save from It*. Oxford: Oxford University Press.

Kamm, F.M. 2007. *Intricate Ethics — Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

Lewis, D. 1986. Causation, *Philosophical papers, vol. 2*, 159–213. Oxford: Oxford University Press.

Lindley, D.V. and M.R. Novick. 1981. The role of exchangeability in inference. *The Annals of Statistics 9*(1): 45–58. https://doi.org/http://dx.doi.org/10.1214/aos/1176345331 .

Machery, E. 2017. *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press.

Malmgren, A. 2011. Rationalism and the content of intuitive judgements. *Mind 120*(478): 263–327 .

McMahan, J. 2002. *The Ethics of Killing*. Oxford: Oxford University Press.

Mittal, Y. 1991. Homogeneity of subpopulations and simpson's paradox. *Journal of the American Statistical Association 86*(413): 167–172. https://doi.org/http://dx.doi.org/10.1080/01621459.1991.10475016 .

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference: 2nd Edition*. Cambridge: Cambridge University Press.

Pearl, J. 2016, August. Simpson's paradox: The riddle that would not die. (comments on four recent papers). Blog Post. Accessed: 2026-06-09.

Plato. 2004. *Republic*. Indianapolis: Hackett Publishing.

Sarkar, P. and P. Bandyopadhyay 2021. Simpson's paradox: A singularity of statistical and inductive inference. Unpublished manuscript. Available online via arXiv: arXiv:2103.16860.

Schaffer, J. 2016. Grounding in the image of causation. *Philosophical Studies 173*(1): 49–100 .

Schurz, G. 2002. Ceteris paribus laws: Classification and deconstruction. *Erkenntnis 57*(3): 351–372 .

Schwitzgebel, E. and F. Cushman. 2012. Expertise in moral reasoning? order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language 27*(2): 135–153 .

Schwitzgebel, E. and F. Cushman. 2015. Philosophers' biased judgements persist despite training, expertise and reflection. *Cognition* 141: 127–137 .

Sider, T. 2011. *Writing the book of the world*. Oxford: Oxford University Press.

Simpson, E.H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological) 13*(2): 238–241. https://doi.org/http://dx.doi.org/10.1111/j.2517-6161.1951.tb00088.x .

Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, Prediction and Search, 2nd Edition*. Cambridge: MIT Press.

Thomson, J.J. 1985. The trolley problem. *The Yale Law Journal 94*(6): 1395–1415 .

Tooley, M. 1972. Abortion and infanticide. *Philosophy and Public Affairs 2*(1): 37–65 .

Trammell, R. 1975. Saving life and taking life. *Journal of Philosophy 72*(5): 131–37 .

Wallmann, C. 2017. A bayesian solution to the conflict of narrowness and precision in direct inference. *Journal for general philosophy of science 48*(3): 485–500 .

Weinberg, J.M., S. Nichols, and S. Stich. 2001. Normativity and epistemic intuitions. *Philosophical topics 29*(1/2): 429–460 .

Williams, B. 1973. *Problems of the self : philosophical papers, 1956-1972*. Cambridge: Cambridge University Press.

Williamson, T. 2022. *The philosophy of philosophy*. Hoboken, NJ: Wiley Blackwell.

Wilson, J. 2016. Internal and external validity in thought experiments. *Proceedings of the Aristotelian Society 116*(2): 127–152 .

Wilson, J. 2021. *Philosophy for Public Health and Public Policy: Beyond the Neglectful State*. Oxford: Oxford University Press.

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