Reasoning from examples in philosophy faces a tradeoff about granularity

Abstract

A common method in philosophy describes examples, cases, scenarios or vignettes and asks whether readers agree with certain judgements about them. Call this *reasoning from examples*. The *philosophico-scientific* species of that method takes scientific inference as its model, and a prominent variant of that species—the *experimental variant*—balances cases by stipulation, on the model of the controlled experiment. Writers including Kagan (1988) and Wilson (2016) have raised objections to the reasoning on various grounds. This paper situates those objections in a broader context. There is an overarching tradeoff concerning the granularity of the cases involved. Underspecified cases—the philosopher's usual choice—invite among other things failures of internal and external validity, undetected non-monotonicity, and divergent completion by readers. But more specific cases invite disagreement, the exclusion of commonplace phenomena, misleading effect-estimates, and an analogue of the bias-variance tradeoff in statistical learning. The tradeoff is not peculiar to philosophy; it is the philosophical form of a problem facing any case-based method. It can be mitigated but not avoided. Newly aware of the modelling decisions at stake, we ought to continue to reason from examples in a cautionary spirit.

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 defeasible support for a hypothesis concerning the relation that holds between some collection of variables. The hypothesis may be formulated before, during, or after the reasoning is executed. For instance, philosophers might use cases to argue that killing and letting die are morally equivalent, that lying is wrong, or that knowledge is not, as it turns out, justified true belief. [1]

This essay is about an underlying tension in the *genus* of reasoning from examples. There are many types of such reasoning and their categorisation is a matter of active

[1] Throughout, 'variable' is used not in the logician's sense of a symbol ranging over a domain, but in the sense standard in causal modelling: a determinable feature—such as *hair colour*—admitting two or more determinate values—such as *blonde* and *white* (Woodward 2003, p.39).

philosophical debate.[2] Some of my arguments apply to the genus as a whole, but many apply to a *species* of the genus that takes some form of scientific inference as its model: call it the *philosophico-scientific* species.

A prominent *variant* of this species attempts to reproduce the methodology of controlled scientific experimentation (Tooley 1972; Thomson 1985; McMahan 2002). One makes the cases within each comparison-class match in their value for all but one variable apparently relevant to the dependent variable of interest, so that any subsequent difference can be inferred to be due to the independent variable in which the cases differ.[3] Since it balances its cases by stipulation on the model of the controlled experiment, I call this the *experimental variant*. Just as not all scientific reasoning proceeds by controlled experimentation, neither does all philosophico-scientific reasoning from examples. Other variants take other types of scientific inference as their model: *Bayesian* inference, say, or the *observational* study.[4] Not every species of reasoning from examples embraces an analogy to scientific inference. For instance, the *interpretive* species does not (§3.6, §6).

Philosophers have historically tended to appeal to highly unspecific, schematic examples (Wilson 2016). On the spectrum of possible degrees of granularity, they almost uniformly choose the coarse end. But the underspecification of examples leads to a family of threats to the reasoning. This may tempt philosophers to provide greater detail—perhaps much greater detail—in their examples. This, however, leads to an entirely separate family of threats.

The result is a tradeoff. The key aim of this paper is to show that it exists. To my knowledge, this has not yet been done systematically, despite the important implications for philosophical methodology. The tradeoff is not peculiar to philosophy. *Any* case-based inquiry must choose a grain and pay for the choice (§3.3), and I do not aim to show that a method sound in science fails in philosophy. One of my aims, instead, is to trace the costs of different responses to the tradeoff in reasoning from examples.

I will make several general claims in conclusion. First, the tradeoff can be mitigated but not altogether avoided. This is because the appropriate degree of granularity is relative to an inquiry and a context, shifting as they do. The tradeoff poses an epistemological worry—of variable severity—that we have made, or are about to make,

[2] For instance, James Wilson distinguishes *one-off clear cases*, *one-off problem cases*, and these same two as elements of *interventional sequences* (Wilson 2021). Wilson's taxonomy tracks evidential structure. Edouard Machery's crosscuts it, sorting cases by five *uses*: the *provocative*, *illustrative*, *formal*, *material*, and the *exploratory* (Chapter 1, Machery 2017). Reasoning from examples is a method for reaching conclusions in its own right. But it also features in more complex methods, such as *reflective equilibrium*—in which our general theories and particular intuitions are 'revised in the light of the other, by an iterative process, until they are brought into harmony' (Williamson 2022, p.246).

[3] Shelly Kagan calls this a *contrast argument*; Frances Kamm calls it the 'technique of equalizing cases' (Kagan 1988; Kamm 2007).

[4] Take Bayesian inference. Kripke's Gödel–Schmidt case is chosen to be maximally *favourable* to the hypothesis he is *attacking*, Descriptivism (Kripke 1981). So, the reader's judgement about this case is likeliest to conform to Descriptivism *in this case if in any*. Kripke stresses that this judgement about the case does *not* conform to Descriptivism—we *still* think 'Gödel' refers to Gödel, not Schmidt—so we have strong evidence *against* Descriptivism. The notion of the *Bayes factor*—the ratio of the judgement's probability under the competing hypotheses—undergirds this inference. I defend this more fully elsewhere.

the wrong modelling choice. Second, certain defeasible heuristics may *guide* us in our choice, but it is implausible that we can produce a set of heuristics applicable exclusively *pre-inquiry* that will settle the matter. For certain inquiries, we can ascertain the appropriate granularity only after beginning work. The use of reasoning from examples that I recommend is cautious.

This need not entail less frequent use. By 'cautious use' I mean use with certain *qualifications*: a statement of which horn of the tradeoff the modelling decisions have embraced, which threats consequently attend the reasoning, and how, if at all, they have been mitigated (§5.3). Such qualifications have often been elided. Made explicit, they would help philosophers both to evaluate instances of reasoning from examples, and to characterise the defeasible evidence those instances provide.

The essay proceeds as follows. §2 gives the rationale for *underspecification*; §3 sets out the threats it incurs; and §4 those incurred by *overspecification*. §5 asks how far the tradeoff extends, and §6 gives some suggestions on what to do about it. The threats differ in what would meet them. Some yield to *more careful use* of a variant (*use threats*). Some attach to the *constitutive* features of a variant or species, and yield only to a change in variant or species (*constitutive threats*). The use threats can be remedied individually but not jointly, since remedying those of §3 by adding detail invites those of §4. *Genus threats* attach to reasoning from examples as such, and no change of species or variant escapes them. The tradeoff that titles the essay is one such threat, and it can be managed but not eliminated.[5] Threats differ too in evidential standing: some are already extensively *evidenced*, others so far only *theoretical*. I flag both dimensions as each arises.[6]

2 Underspecification: a rationale

Some philosophers take underspecification to be an unavoidable—even desirable—feature of an example. As Borges' short story 'On Exactitude in Science' suggests, a map is useful only to the extent that it is simpler than the territory (Borges 1998). Take Kamm:

> 'Real-life cases often do not contain the relevant—or solely the relevant—characteristics to help in our search for principles. If our aim is to discover the relative weight of, say, two factors, we should consider cases that involve only these two factors, perhaps artificially, rather than distract ourselves with other factors and options' (Kamm 1993, p. 7).

On this account, underspecification is an epistemically virtuous form of Galilean idealisation (McMullin 1985; Weisberg 2013): we reduce a subset of a model's parameters to zero, generating a simpler model with improved tractability and predictive power. In reasoning from examples, this may amount to omitting features that appear

[5] A threat's type is relative to the variant in view. What is a use threat for one variant may be constitutive for another (see, for instance, §4.1). Unless noted, I classify relative to the experimental variant.

[6] 'Extensively' is vague, and in any case, the evidence is typically defeasible and empirical. Some of the support for evidenced threats, though, involves formal proof (for instance, §§4.2, 4.3). Theoretical threats include those of §3.2 (Simpson's paradox) and §4.4 (the bias-variance tradeoff). The evidenced/theoretical distinction crosscuts *severity*. A theoretical threat may prove more severe than any evidenced one, and its articulation may incentivise research that makes it evidenced.

relevant but whose inclusion makes judgements more difficult, or to designing cases whose descriptions distinguish too few situations to represent certain variations.[7] For the reasoner of the experimental variant, underspecification doubles as the balancing device: what the description leaves unstated is thereby to be held equal across the pair.

This idealisation is supported by a metaepistemological view we might call *Epistemology as Modelling* (EM), whose components are found in Greco (2023), Titelbaum (2012), Weisberg (2013), and Williamson (2024). Among the commitments of *EM* are that as soon as we theorise using folk psychological or Bayesian entities we are *already* modelling; that for S to have credence n that p is for the best Bayesian model of S to represent her with credence n that p; and that we are permitted to use different semantic models in different contexts, provided no single framework offers a perfect model of the domain. *EM* enables dissolution of certain problems that otherwise appear to devastate orthodox decision theory and Bayesian epistemology (Mahtani 2017; Mahtani 2021; Mahtani 2024), and nullifies arguments for sceptical conclusions such as Lederman's argument against the existence of common knowledge (Lederman 2018; Greco 2023).

EM faces problems of its own, although here is not the place for them. Moreover, a family of concerns arises for the method of reasoning from underspecified examples. I dedicate considerable space to these concerns, due to the prevalence of this method in philosophy.

3 Problems for Underspecification

The threats below follow the inference involved in reasoning from examples outward: from whether the case can be grasped at all (§3.1), to whether the reasoning measures what it purports to (§3.2), to whether what it measures travels (§3.3; §3.4), to whether different readers are considering the same case (§3.5), and finally to whether the schematic style is fit for the subject-matter at all (§3.6).

3.1 Grasping failures, judgements, and cases

Consider Frances Kamm's case, known as:

Reach: 'I stand in a part of India, but I have very long arms that reach all the way to the other end of India, allowing me to reach a child who is drowning in a pond at a great distance' (Kamm 2007, p. 352).

One of Kamm's aims with this case is to persuade us that there is no salient difference between the child we reach with our enormous arms, and a child reachable with regular human arms. Following Elster, Wilson plausibly objects that to grasp *Reach* intellectually—let alone make a morally significant judgement about it—a great deal

[7] Batterman even argues, of mathematical idealisation in science, that some idealisations 'are explanatorily ineliminable', since a 'full understanding of certain phenomena cannot be obtained through a completely detailed, nonidealized representation' (Batterman 2009, p.428). The same may hold for certain phenomena studied in philosophy. I thank an anonymous reviewer for the reference.

more needs to be said about the nature of such a world (Wilson 2021; Elster 2011).

The reason is that we need to be able to imagine *Reach* to make a judgement about it (Parfit 1984, pp.388-9). But to imagine it, we need some idea of how the world would have to be *adjusted*—morally and non-morally—to accommodate the arms (Wilkes 1988; Häggqvist 1996). And the case is too outlandish to supply that idea.[8] The threat attaches to the genus. Outlandish underspecified cases are a problem for reasoning from examples as such, not only for the philosophico-scientific species.

Some apparatus will help us better articulate the threat. A case is always presented by a *description*, D—the text beside the word 'Reach', say (Lewis 1983). Call the set W_D of all and only those possible worlds at which D is true its *realisations*. A description is *coarse* to the extent that it has many realisations and *fine* to the extent that it has few; in the limit, a true description holds at all the worlds or just one.[9] *Reach* lies at the coarse end. Its coarseness is what the Wilson–Elster objection turns on. The description leaves so much unsettled that the reader has no *purchase* on which of its realisations she is even being invited to judge.

When reasoning from examples for some hypothesis, a plausible criterion on a description of an example is that it ought not *itself* to state that hypothesis (call this *Machery's criterion*, after Machery 2017, pp.11-12).[10] Instead, the reader is to infer that the hypothesis holds from—*inter alia*—that description. The criterion may set one limit on how far refinement can be pushed as a remedy to the threats of this section (see §4.4).

3.2 Internal validity failures

Suppose the examples are non-outlandish, the reader makes judgements about them, and the content of those judgements is clear. Still, the reasoning may fail to exhibit *internal validity*: the requirement that the reasoning 'correctly measures the causal effect of an independent variable or variables on one or more dependent variables' (Wilson 2021, p.49).

The notion derives from experimental research design (Campbell 1957). In a philosophical setting, the effect measured may also (or *instead*) be a metaphysical grounding relation, in the image of causation, à la Schaffer (2016). Wilson provides

[8] Does the giant move? *Carefully*?

[9] The granularity of D bears on a separate debate I do not enter here: what the content of a reader's judgement about a case is, given that D is consistent with more specific 'deviant' realisations in which the case's putative lesson fails to hold. In a Gettier case the lesson is that the agent has a justified true belief without knowledge (Gettier 1963). In one deviant realisation, Smith stands to the proposition that someone in his office owns a Ford *exactly* as the original Gettier description stipulates, but—because he has *additional* good reason to think himself disposed to hallucinate people driving Fords—his justification later lapses (Malmgren 2011, p.279). Every account of the content of the reader's judgement about the case quantifies over the set W_D of realisations of D: over all of them on necessity accounts (Ichikawa and Jarvis 2009; Saint-Germier 2021), some on possibility accounts (Malmgren 2011), the closest on counterfactual accounts (Williamson 2022), and the typical on normalcy accounts (Geddes 2018). The grain of D fixes which realisations W_D contains.

[10] For instance, no description of a Gettier case ought to *state* the sentence that the agent has a justified true belief without knowledge. We are to infer this of the case.

multiple instances of reasoning from underspecified examples that fail internal validity (Wilson 2021, pp.51-3). To these I add a novel one. The worry is a *use threat*—one that more judicious use of the method could in principle meet—and, being so far under-investigated, a *theoretical* one. It nonetheless has significant implications, particularly for the prevalent experimental variant. 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).[11]

One can imagine arguing for *TRI* by presenting comparison-classes of cases—as McMahan does with his pairs Bright/Dull, Cheerful/Melancholy, Fortunate/Unfortunate, and Young/Old—in which the background is stipulated identical and only the strength of time-relative interests varies (McMahan 2002, p.234). A reader forms a judgement about each pair partly on the basis of her background beliefs about how the relevant variables are related.

Now, consider the following instance of a phenomenon known as *Simpson's paradox*: at a given university in a given year, there is a negative association overall between womanhood and admittance, but a positive association between womanhood and admittance for every subject.[12] When evaluating an underspecified case, the reader may omit a variable that stands, according to her, to the variables of interest as the 'subject studied' variable stands to 'gender' and 'admittance' in that instance of the paradox. If the reader omits such a variable—say, social status—then according to her, there will be a negative association overall between strength of time-relative interests and permissibility of killing, but a positive association between those variable-values within *every* social status class. The reader will then agree with *TRI* at the coarser level of granularity, but disagree with it for every value of the omitted variable once that variable is included. In such a situation, the reader's judgement does not correctly measure the relation in question, and the reasoning therefore lacks internal validity.

The worry depends neither on the reader's having credences rather than beliefs, nor on the rejection of a picture like Williamson's (2022) on which reasoning from examples supplies valid arguments with a counterfactual premiss, nor on confinement to the experimental variant, nor on the transparency, to the reader, of those background doxastic states. Provided the reader's output judgement about the counterfactual

[11] The strength of one's time-relative interests in continuing to live is, in turn, taken to be constituted by 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).

[12] This might hold, in turn, because women disproportionately apply to a highly selective subject. This case of the paradox is modified from Bickel *et al* (1975). In general, an association is subject to Simpson's paradox just if it is absent, or reversed, for every value of some partition variable. A variable *partitions* a set of cases exactly when each case takes exactly one of that variable's values; the values thus 'sort' the cases into exhaustive, non-overlapping cells.

scenario or intuition pump rests partly on background doxastic states that may take the reversal-paradoxical shape, the worry goes through. For those states to take that shape in the McMahan case, the reader need only hold—for any reason, *a priori* or *a posteriori* alike—that time-relative interests are typically stronger among those with high social status, whom it is typically less permissible to kill.

The threat thus concerns the possibility that the reader's *doxastic states* have the reversal-paradoxical shape. But the phenomena themselves might be reversal-paradoxical, too or instead, if *the chance distribution* governing those phenomena, or the relevant *long-run frequencies*, take that shape. Call this *worldly reversal*. It may also go undiscovered in reasoning from underspecified examples. Worldly reversal may be one cause of the kind of non-monotonicity effects examined more deeply in §3.4. Not all phenomena studied by philosophers can plausibly admit of worldly reversal, of course. For it to obtain, the relations between the relevant variables must admit of *exceptions*—for example, there must be at least one being with strong time-relative interests who may nonetheless be permissibly killed—and this might not be so.

The threat affects any variant of the philosophico-scientific species, but it is especially pressing for the experimental variant. By its own lights, having enjoined the reader to balance cases, why stop there? Clearly, case-balancing does not buy internal validity on its own. Without further motivation, it seems arbitrary to require case-balancing while leaving other threats to internal validity unchecked.

More generally, the experimental variant models itself on the controlled experiment, and so inherits its characteristic failure modes: confounding, interaction effects (§3.4) and the internal-for-external-validity trade (§3.3).[13]

3.3 External validity and the balancing threat

Even where the reasoning measures the relevant effect, and thereby exhibits internal validity, its results may not obtain elsewhere. An important desideratum threatened by underspecification is *external validity*, which holds just if internal validity does, and additionally 'its results are applicable to a wide variety of other contexts, and in particular… there is reason to think that its intervention will work here' (Wilson 2021, p. 49), (Cartwright and Hardie 2012).

External validity in reasoning from examples broadly asks two things of a set of cases. First, that they *cover* the settings we care about. Second, that the relation they report is *stable* across those settings. Either alone would do. Complete coverage makes stability dispensable, since a different relation in each setting would itself be an important discovery; on the other hand, a relation known not to vary makes coverage

[13] More precisely, a variable Z 1. *confounds* the relationship between a treatment variable X and a dependent variable Y just in case Z influences both X and Y, with its influence on Y not running through X nor *vice versa*; 2. *mediates* the relationship between X and Y just in case X influences Z, which in turn influences Y; and 3. is a *collider* with respect to that relationship just in case X and Y each influence Z. Holding a confounder fixed is *required* to adequately measure the effect of X on Y, but holding a mediator or collider fixed can corrupt that measurement (the mediator by blocking part of the total pathway under study, and the collider by opening a spurious one).

dispensable, since establishing it once establishes it everywhere. Underspecification secures neither. Coverage failures are the business of this section, and stability is the subject of §3.4.

Consider again *TRI*. The reasoner of the experimental variant wants the reader to hold fixed all morally relevant variables except time-relative interests. Let $\mathbf{\Gamma}$ be the finest morally relevant partition variable.[14] The reader evaluates *TRI* only by considering comparison-classes homogeneous with respect to $\mathbf{\Gamma}$.

This is a tall order. But even if the reader successfully holds $\mathbf{\Gamma}$ fixed within each comparison-class, she may fail to consider the full range of $\mathbf{\Gamma}$'s values. Suppose $\mathbf{\Gamma}$ is the product of two dichotomous variables: whether the individual wants to die, and whether the individual is in some pain. The reader might evaluate *TRI* only for cases in which the individual neither wants to die nor is in pain, entirely omitting the remaining three cells of $\mathbf{\Gamma}$, which may be the most common values $\mathbf{\Gamma}$ takes in actuality. Call this the *truncation problem*. In the terms of our taxonomy from §1, it is a use threat to the experimental and Bayesian variants, and so far a theoretical one.

Worse, suppose every value of $\mathbf{\Gamma}$ *is* somehow represented. At each, the reader judges a pair differing in time-relative interests and agreeing in all else. But actual situations do not hold the background still while the treatment variable alone moves. So a pair balanced by stipulation may be a combination nothing in the world instantiates. Whereas truncation was an omission—repaired by visiting the cells left out—balancing tends to produce *artificial* cases not seen in the actual world. No further sampling within the balanced space may reduce that artificiality.

The internal and external validity problems just canvassed are, in philosophy, analogues of two problems familiar elsewhere. These problems each concern the evidential standing of a particular method. The first method uses historical cases to support theories in the philosophy of science. The second is the randomised controlled trial (or RCT).[15]

Take the first. A historical episode furnishes evidence for a general theory only once it has been *constructed* from the inchoate past, *selected*, and *interpreted* (Bolinska and Martin 2020). But there is reason to think each step is guided by the commitments the analyst brings to it (Kinzel 2015), and that analysts are liable to read a case in support of views already held (Chakravartty 2017). This worry tracks the internal validity concern; a historical case so treated may just not measure what it purports to. The RCT supplies the analogue for external validity. Cartwright (2007) argues that the very design features securing a trial's internal validity on its sample—healthy

[14] For example, suppose just two variables are relevant—whether the individual is in at least some pain, and whether the individual wants to die—each with two values. Then $\mathbf{\Gamma}$ has four values (in pain and wants to die; in pain but does not; and so on). More generally, the values of $\mathbf{\Gamma}$ are exactly the possible combinations of all the other morally relevant variables; that is, $\mathbf{\Gamma}$ is just the *Cartesian product* of all the other morally relevant variables. $\mathbf{\Gamma}$ keeps this meaning throughout the essay. I will also refer to the values of $\mathbf{\Gamma}$ as *the covariate space*.

[15] I am very grateful to an anonymous reviewer for this insight.

young males, say—are what limit its reach beyond that sample.[16] So too with reasoning from examples. The method tends to employ cases in exceptional regions of the covariate space; the trolley problems of Foot (2001, p.147) and Thomson (1985, p.1397) for instance evaluate their principles only where victims are anonymous and bystanders have no institutional role. The balancing that secures internal validity is thus what manufactures the artificiality threatening external validity (see §3.5). Since balancing is *constitutive* of the experimental variant, this is a constitutive threat—the *balancing threat*—that no more careful balancing can meet.

The granularity tradeoff is not a peculiarly philosophical predicament. *Some* level-of-description problem attends any method that reasons from cases, scientific and metascientific inference included. Too fine a description is unusable, too coarse a description uninformative. One feature is nonetheless distinctive to reasoning from examples. Where the analyst constructs her historical case, the reasoner from examples typically underspecifies hers, which leaves the *reader* required to complete the construction; so the reader, too, inherits the granularity choice (§3.1, §3.5).

None of this provides good reason to abandon reasoning from examples. Bolinska and Martin (2020) hold that the analogues of internal and external validity in the historical case are not peculiar to it, and that they admit methodological remedies rather than requiring the method's abandonment. Likewise, certain kinds of reasoning from examples may be constitutively less vulnerable to the threats from internal and external validity. The cases of the *observational variant* of the philosophico-scientific species of reasoning from examples lie in dense regions of the covariate space. Practitioners do not seek to balance cases, and confounding is controlled in the analysis rather than by stipulation. The observational variant may be one way to instantiate the extra care Bolinska and Martin recommend in reasoning from examples, without falling prey to the balancing threat, and without abandoning *some* analogy to scientific inference (§3.6, §6).

3.4 Non-monotonicity

Relations between the variables philosophers care about are frequently *non-monotonic*; that is, such that additional information can weaken, cancel or reverse them. This is partly why the stability §3.3 required can be hard to come by. In general, that P confirms Q does not entail that $P\&R$ confirms Q (Titelbaum 2022), and on the standard semantics for the counterfactual conditional, 'if P were the case, then Q would be the case' may be true even though 'if P and R were the case, then Q would be the case' is false (Lewis 1973). *Detecting* non-monotonicity requires restoring detail, but a coarse case is defined by withholding it. A relation that would not survive the restoration can look, from examination of an underspecified case alone, exactly like one that would. The threat of this section is that reasoning from underspecified

[16] The most general problem here—which afflicts reasoning in many domains—is the *problem of extrapolation*: 'the challenge of transferring causal generalizations from one context to another when homogeneity cannot be presumed' (Steel 2008, p.3). More particularly to the philosophy of science, one faces the problem of *generalisability*: whether—and how—*particular* cases from the history of science can confirm *general* theories in the philosophy of science (Pitt 2001).

examples may obfuscate the fact that the instability exists as well as its source.[17]

An entire research program is founded, in part, on the view that moral reasons are non-monotonic. Moral Particularists such as Jonathan Dancy point to cases such as:

Book: The fact that I borrowed a book from you is a reason to return it to you. But in a context in which you stole the book from the library, it is not clear that this is so. In yet another context in which you stole the book from the library because you sorely need the book, and lack the fixed address required for a library membership, the fact that I borrowed the book from you may again be a reason to return it to you (Horty 2012);

to argue for:

Moral Holism: for every nonnormative feature of an action that gives rise to a reason for (or against) action in one possible context, there is another possible context in which that same feature either gives rise to a reason of opposite valence, or else provides no reason one way or the other.

Moral Particularism is plausibly the conjunction of Moral Holism and the thesis that the combinatorial function for important moral variables of interest—such as rightness, wrongness, moral status, *etc*—is not finitely specifiable (Berker 2007). If the Particularist argument succeeds, reasoning from underspecified examples is dangerous: it may prevent the reasoner or reader from discovering that some feature giving rise to a reason in one context ceases to do so in a more specific one. Such an argument is structurally similar to the use of cases such as:

Prime: the fact that the number I have randomly selected between 1 and 10 inclusive is less than 3 is a reason to believe that that number is prime. Equally, the fact that that number is odd is a reason to believe that that number is prime. But the fact that that number is less than 3, *and* odd, is not a reason to believe that that number is prime, since this fact entails that that number is 1, and so not prime (Titelbaum 2019);

in epistemology to argue for:

Epistemic Holism: for every nonnormative feature of a situation that gives rise to a reason for (or against) belief in one possible context, there is another possible context in which that same feature either gives rise to a reason of opposite valence, or else provides no reason one way or the other.

The same facts in *Prime* that provide epistemic reasons may also provide moral reasons to place (morally important) bets.

[17] This is an evidenced use threat to the experimental and Bayesian variants of the philosophico-scientific species, met by more detailed cases. It does not arise for the observational variant, which uses highly detailed cases by design (§3.6).

Holism need not obtain in order for the threat of this section to go through. In fact, *Book* does not obviously establish Holism. Three readings of *Book* are available. On the Holist reading, the fact that I borrowed the book from you *stops* being a reason to return it to you once you turn out to have stolen it. On Joseph Raz's reading, however, that fact *never* was the whole reason to return it: the reason was always a more inclusive fact, one component of which is that you did not steal the book (Raz 2000). On another reading, the fact that I borrowed it remains a reason to return it to you, but is defeated by a second—that you stole the book—which recommends returning the book to the library instead (Horty 2012). The case alone underdetermines which reading obtains. On each reading, the reader of the underspecified case may miss something: a reason that lapsed, a condition on the reason she has, or a defeater, respectively.[18]

Moreover, underspecified cases seem inapt to provide evidential support for moral principles, regardless of how those principles are construed. Consider four decreasing orders of strength for a principle like 'stealing is wrong, other things equal': (1) for all values of $\mathbf{\Gamma}$, any act of stealing is wrong; (2) for all values of $\mathbf{\Gamma}$, any act of stealing is worse than before (but not necessarily wrong); (3) for some values of $\mathbf{\Gamma}$, any act of stealing is wrong; (4) for some values of $\mathbf{\Gamma}$, any act of stealing is worse than before. All four construals require examining at least some values of $\mathbf{\Gamma}$. But underspecified cases cannot do this: for most important moral questions, the morally relevant partition variables have too many values for a schematic case to hold $\mathbf{\Gamma}$ fixed.

Whether or not Holism holds, coarse cases can also hide how reasons *combine*. Kagan's cases show this (Kagan 1988). Killing is intuitively worse than letting die in balanced cases. But when I am under attack by an aggressor, it appears to make no difference whether I kill or let die. Being under attack acts as a *zero-multiplier*, and the function determining the moral status of the act looks multiplicative rather than additive. A reasoner working from the unmodified pair alone may misjudge the killing/letting-die differential; she may also attribute an additive structure to moral reasons where none exists. Being under attack cancels the differential. The differential is a fact about the unmodified pair, not an association that carries to other settings.

A final source of non-monotonicity that can be hidden by underspecified cases lies with the reader rather than with the world. We have already seen this source in the case of Simpson's paradox (§3.2). Judgements about cases implicitly invoke a reference class of relevantly similar cases, and the grain of the case fixes which class that is. The point is not confined to the statistical setting. R.M. Hare's account of

[18] I take it that the debate here about what the reason is hinges on the question of whether an agent can have a reason only to the extent that they understand or can articulate that reason. Dancy answers this question in the affirmative: he argues that the enablers—that is, the absence of the defeaters—cannot be part of the reason, because there are 'just too many potential defeaters for the absence of each one to count among our original reasons' (Dancy 1993, p.81). But Raz answers in the negative: Raz holds that Dancy's remark confuses articulation and knowledge, and that an agent can have and know a highly complex reason—and therefore be guided by it—even if they can only ever articulate parts of it (Raz 2000). I remain agnostic. But it seems to me that Raz's position is only plausible if one construes his agents as ideal in a way Dancy's project, from the outset, does not.

moral reasoning as the progressive refinement of principles—the move from 'signal before turning' to 'signal before turning, except in emergencies'—shows a relation that holds at one level of refinement failing at the next (Hare 1952). Detecting this source of non-monotonicity differs from detecting the others. Refining the case only moves the reader to a new reference class. Exposing it requires comparison across grains (see §6's cross-grain check).

3.5 Divergent completions, reproducibility failures, and biased completions

Those sources concerned what a single reader might fail to detect. But readers might yield verdicts incompatible with one another. Recall that in reasoning from examples, the support a case provides comes *via* the judgements it elicits. But two readers may fill in the elided details of an underspecified case in different ways, and inconsistent specifications can generate inconsistent judgements. Williams' *Experiment* case—in which two people undergo a process that might swap their bodies, before one is given $100,000 and the other is tortured—is the classic illustration (Williams 1973). A completion dwelling on the participants' affective attitudes makes the case appear to confirm that the experiment swaps bodies, and that psychological continuity determines identity; a completion dwelling on the incremental psychological changes one person undergoes makes it appear to disconfirm both. This is a concern for reasoning from underspecified examples in general. But it is especially pressing for the experimental variant, for which *reproducibility* is an explicit aspiration (Mayo 2018; Howson and Urbach 2006).

More recently, Crain (2023) argues that underspecified examples enact a *faux anonymisation* of moral agents. Blanks are filled in by the reader in ways that reveal biases confounding the eventual judgement, and this happens unwittingly, even among readers who believe they treat the elided features as indeterminate.

Two further schematic cases have silences that admit divergent, good-faith completions. Foot invites us to imagine 'the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed' (Foot 2001, p.147). She uses the case to motivate a distinction between the duty to refrain from injury and the duty to provide aid. Thomson's bystander variant relocates the choice. A passer-by at a switch can turn the trolley from the five to the one (Thomson 1985, p.1397). Both are minimally specified. Beyond the numbers and the redirection, the reader is told almost nothing: the six are undifferentiated; the cause of the malfunction unspecified; the bystander's institutional role and relationships left open.

On the rationale of §2, underspecification is the experimental variant's balancing device: unstated variables are held equal by stipulation of silence. But silence does not say *how* to hold them equal. Distinguish two good-faith executions of the instruction:

Reading A (*absence*): the reader takes the unmentioned features to be absent.[19] No one on either track has dependents, and the deaths have no further consequences.

Reading B (*openness*): the reader leaves the unmentioned features open. Any of the six may have dependents; the description gives no more reason to ascribe them to one than to another.

Neither is outlandish, and neither favours a track. A treats the six alike by giving each nothing, B by leaving the same questions open about each. Yet they are not equivalent. On A, the choice is between one death and five, *tout court*. On B, whatever might attach to one of the six—dependents, attachments, further consequences—might equally attach to each, so the five's side risks five times whatever the one's does. Losses beyond the deaths themselves then scale with the count; the two are confounded, and a verdict on B cannot, by itself, discriminate which is doing the work. A avoids the confound only by placing the case in an exceptional corner of the covariate space.

A third execution—*sameness*—fills the features in identically across the six: all rail workers in similar circumstances, say. But sameness must be imposed at some *grain*. The choice of grain re-raises the question the instruction was meant to close. These cooperative completions thus reproduce the granularity tradeoff in miniature. Readers need not even settle the silence in the same mode at a finer grain. One may take the six as statistical placeholders, another as determinate strangers about whom nothing further is known—a difference that measurably shifts responses (Small and Loewenstein 2003, pp.9–10, 12).[20]

Nor is declining to fill in an escape. The text does not settle what lies beyond it, and readers screen off deviant realisations—the diversion that provokes a reprisal killing fifty (cf. Malmgren 2011, p.299), say—by their tacit sense of the *intended case*; *not* by the description alone (cf. §3.1). The instruction to hold all else equal thus admits non-equivalent executions.

These considerations generate at least two threats. The first concerns *reproducibility*, and poses a problem particularly to the experimental variant. Readers completing in the mode of A and of B are, in effect, running different experiments. Analogy to controlled experimentation fails at the *ceteris paribus* clause. The second is a generalised version of Crain's concern above and affects any reasoning from examples that attempts to isolate the effect of some variable. Readers appear to settle silence in ways that track their own contingent psychological, demographic or political situation. They do not distribute randomly across modes of completion. But then the resulting verdicts may owe as much to the reader's situation as to the structural features Foot and Thomson take themselves to be isolating.

[19] One might see this as the variables' having 'null' values.

[20] Completions that individuate asymmetrically shift verdicts further. Making the one young, or genetically related to the agent, sharply reduces willingness to divert (Bleske-Rechek et al. 2010, Study 1).

3.6 Adequacy

The threats of §§3.2–3.5 were, for the most part, *use* threats to the experimental variant—failures that it could in principle meet by more careful use. The threats of this section are constitutive. They attach to the experimental variant and to the philosophico-scientific species.

Underspecification may strongly correlate with—if not imply—other properties inimical to good philosophical writing. Nussbaum (1990) and Williams (2006) have argued in different dialectical contexts for claims that entail the same general conclusion: that analytic philosophical argumentation tends to be inadequate to the complexity and significance of the truths supposedly confirmed by it. I cannot settle the dispute here. But to the extent that their arguments have credibility, we have a further worry for reasoning from underspecified examples.

Nussbaum holds that certain important views about how to live can be expressed only in linguistic styles more complex and attentive to particulars than conventional philosophical prose. If she is right, the prose typically used to describe underspecified cases—by design schematic and abstract—may itself be unable to express such views; and it is unclear how one could reliably infer them from deficient descriptions. Williams instead holds that one desideratum on philosophical writing is that it is *reflective*, in a sense that requires engaging the reader's sense of what is important. But an artificially scientistic style abstracts from detail and aims to isolate a treatment variable. It may therefore systematically fail this desideratum.

Two demands on reasoning from examples can be extracted from Nussbaum and Williams' arguments. The first is a demand for *particularity*. The description of a case must be rich enough to express the truths at issue. The second is a demand for *interpretive engagement*. The reader ought to respond to the case as a whole, rather than to a variable putatively isolated within it. They apply at different levels. Particularity is precluded by the schematic style of the experimental variant. Interpretive engagement is precluded by the aspiration of the philosophico-scientific species to isolate a variable. Both demands pose constitutive threats, and they differ in what would meet them.

The particularity demand does not require abandoning the species of philosophico-scientific reasoning from examples. Where the controlled trial is unavailable, medicine turns to observational methods rather than abandoning statistical inference (Rosenbaum 2002). §3.3's observational variant is the analogue in reasoning from examples. Its cases are drawn from the actual record of history, literature, or law, and so have the detail Nussbaum demands, but its analysis still aims to isolate the causal or metaphysical contribution of a particular variable.

Satisfying the demand for interpretive engagement requires abandoning the philosophico-scientific species. To meet it, readers need to engage with the examples more deeply: perhaps in a way analogous to the way readers of literary fiction ought

to engage with those fictions; perhaps, in addition or instead, by attending to the historical origins of the examples or of the concepts involved in them. Likewise, in the philosophy of science, Schickore (2011) objects that the 'confrontation' model—in which theories in the philosophy of science are putatively tested against individual cases—both *misdescribes* how metascientific knowledge is in fact produced, and *misprescribes* how to acquire it. In its place she recommends that theorists attend to the historical origins of the cases or the concepts the invoke, and to the delicate interdependence that obtains between theories and cases (Schickore 2011, p.477). Just as, for Nussbaum and Williams, a scientific reporting style is inadequate to certain truths about how to live, for Schickore it is inadequate to the interpretive character of inquiry. We will take up the choice between the observational variant and interpretive engagement again in §6, once the hazards of richer cases are on the table.

4 Overspecification

Underspecification faced a family of threats to which a natural remedy is greater detail. This section sets out what that costs. Detail bears on whether readers converge in their verdicts (§4.1); on which phenomena a case can represent at all (§4.2); on what can be estimated from a case (§4.3); and on the proposition a reader ought to infer (§4.4). We will conclude that no choice of granularity, when reasoning from examples, comes for free.

4.1 Disagreement

James Wilson points to a threat to reasoning from more specific cases:

> '[T]he richer and more realistic the case...the less plausible it is to think of the case as akin to a controlled scientific intervention, and the more scope there is for disagreements in judgements about the case' (Wilson 2016, p.138).

Wilson's worry is not that specificity *entails* disagreement. To the extent that granularity is a predictor of consensus, it will only ever be one variable among many. The worry instead is twofold.

First, richly detailed cases strain the experimental variant's analogy. Controlled experimentation attempts to isolate the influence of a single variable. But a collection of rich hypothetical cases bears little resemblance to that model.

Second, even the reasoner not pressing the analogy may face more disagreement than with less specific cases, through several mechanisms. Readers who complete underspecified cases differently may disagree accordingly (§3.5), and richer cases can be the product of more ambitious completions. Detailed cases may also acquire the *narrative texture* of literary texts—characters with implied histories, actions with cultural weight—which different readers may interpret through different frameworks by attending to different aspects of those cases. Moreover, a richer case requires the reader to *integrate* more variables' contributions. As the number of variables grows, so too, at least typically, does the number of ways of integrating them, so readers

may disagree on the verdict even given identical inputs.

These claims are descriptive tendency-claims, not entailments. Richer cases *tend*, other things equal, to elicit more divergent verdicts, because each mechanism has more to operate on. But there is some empirical support for the claims.[21] The more general phenomenon—the sensitivity of the judgements of well-primed populations to normatively irrelevant factors—is well supported in experimental philosophy (for sensitivity to *order-effects*, for instance, see Schwitzgebel and Cushman 2012, pp.145-6).[22]

Some of the agreement enjoyed by coarser cases may accordingly be, adapting Crain's phrase from §3.5, merely *faux agreement*. Two readers may concur on a schematic case *only because* neither has yet rendered it more specific; were each to do so, they would reliably disagree. Increasing specificity seems to have something in its favour: it can reveal otherwise undiscoverable *latent* disagreement.[23]

Latent disagreement is genuinely revealed only if readers of the more detailed version still consider the *same* case, rather than a new one. This is a non-trivial question (see §3.1). Where they are, and where the more granular model is the more explanatorily and predictively fruitful, it does seem plausible that exposure has diagnostic value. It distinguishes substantive consensus from an artefact of coarse description. But if *EM* is the right metaepistemology, the discovery of latent disagreement does not by itself settle what to do. On *EM*, we are permitted, depending on the comparative explanatory and predictive fruitfulness of the available models, to adopt the less granular model, on which there is more consensus, over the more granular one, on which there is less. That we would disagree under finer descriptions does not entail that the disagreement ought to be attended to, examined, and ultimately resolved.[24]

4.2 The preclusion of commonplace phenomena

Some problem-specifications can only be solved at a sufficiently fine level of detail.[25] But maximally fine specifications can bring untenable consequences. There are cases

[21] Nichols and Knobe report that verdicts on hypothetical agents' moral responsibility shifted when additional detail modified the cases' affective content (Nichols and Knobe 2007). Other results suggest that additional detail can increase willingness to help a hypothetical individual (Kogut and Ritov 2005), and that, depending on readers' prior commitments, it can engender disagreement where none was present before (Uhlmann et al. 2009).

[22] Sampling many readers may mitigate some internal validity threats—idiosyncratic completions may wash out, and the Schwitzgebel–Cushman instability can at least be measured—but it leaves the granularity problems untouched. Every vignette the experimentalist administers must be pitched at some granularity, to which the tradeoffs of §§3–4 apply.

[23] I thank an anonymous reviewer for this insight.

[24] The threat of this section is evidenced by the empirical work. Its type varies with the variant. For the experimental and Bayesian variants it is a use threat, met by coarsening. For the observational variant, whose cases are constitutively rich, coarsening is unavailable and the threat is constitutive. For the interpretive species, which does not aim at converging verdicts, it is no threat.

[25] For instance, the right conclusion to draw in the Monty Hall problem is that the probability that the car is behind the door you chose—door A—given that Monty subsequently opens door C to reveal a goat is *half* the probability that the car is behind B given that Monty opens C to reveal a goat, and so that you should switch to B when offered. One can derive this result with basic Bayesian reasoning, but only if one partitions the states of the world with the more granular set {car is behind door A and Monty opens door B, car is behind door A and Monty opens door C, car is behind door B and Monty opens door C, car is behind door C and Monty opens door B}, rather than the coarser set {car is behind door A, car is behind door B, car is behind door C}.

in which they provably stack the deck against putatively commonplace phenomena.

It has been shown that a sufficiently fine description of both possible states of the world and the cognitive or perceptual states of agents in a problem precludes common knowledge in a wide variety of models (Fagin et al. 2003; Lederman 2018), (Greco 2023, §7.2). In the distributed systems framework of Fagin et al., common knowledge can exist only if certain borderline epistemic possibilities are omitted from the models. More generally, Greco shows that for any initially plausible knowledge-ascription, some sufficiently fine re-description of the epistemic possibilities renders that ascription implausible (Greco 2023, p.155). This is not a decisive reason to reject maximally fine problem-specifications: perhaps we were wrong to take the supposedly widespread phenomena for granted. But it shows that maximally fine specifications may be impossible to reconcile with our other ontological commitments. And in a Moorean spirit, we may be more confident in those commitments than in the methodological assumptions that justify the maximally fine specification. *EM* assesses model granularity on the basis of explanatory and predictive power, and provides no guarantee that the maximally fruitful model employs the maximally fine partition of states (Greco 2023).[26]

4.3 Misleading estimates

Recall that, in §3.2, omitting a variable allowed the relation of interest to reverse under refinement. The remedy was to include the variable. But inclusion can have severe costs of its own. Maximally fine partitions of states may be systematically misleading when we are trying to estimate effects, probabilities, or metaphysical relations in the image of causation à la (Schaffer 2016). It is well established in the causal inference literature that conditioning on post-treatment variables—mediators—can block causal pathways and create spurious associations (Spirtes, Glymour, and Scheines 2000; Woodward 2003; Pearl 2009). The canonical illustration is Hesslow's (1976) case: a treatment has a direct harmful effect but also an indirect beneficial effect by inhibiting something itself very harmful. If we condition on the mediator, we wrongly infer that the treatment has only the harmful effect.

Likewise, if a philosopher constructs a case in which she holds fixed a variable causally (or metaphysically) downstream of the treatment variable but which itself affects the dependent variable, she may inadvertently block the very pathway she aims to study. If, for instance, strength of time-relative interests partly determines certain kinds of social relationship, which in turn contribute to the permissibility of killing, then holding those relationships fixed while varying time-relative interests results in no longer studying the total effect of time-relative interests. The key point is that the finest relevant partition variable may be a mediator or collider—and hence precisely not the variable whose value ought to be held fixed when assessing the influence of

[26]The threat is evidenced by formal results. For the experimental and Bayesian variants it is a use threat, met by careful coarsening; for the observational variant, whose cases cannot be significantly coarsened, it is constitutive.

the treatment variable.[27] For the philosophico-scientific reasoner, absent a proposed causal model, it may be quite unclear to the reader which variables are mediators and which should therefore not be held fixed.[28]

4.4 The bias-variance analogue

§§4.2 and 4.3 concerned what a highly specific case can represent and support. This section turns to the proposition the reader infers from such a case. The threat is suggested by a formal analogy.

As the examples become more specific, there may be an analogue to the *bias-variance tradeoff* of statistical learning. In the statistical setting, as model complexity increases, training error decreases—eventually to zero—but the model becomes overfit: sensitive to noise in the training data and liable to fail on testing data (Hastie, Tibshirani, and Friedman 2009, Chapter 7). The tradeoff arises because increasing complexity reduces bias but increases variance.[29]

Analogously in reasoning from examples, increasing case-granularity reduces what we might call *representational bias*—the risk that the case fails to capture morally relevant features of real situations—but increases what we might call *inferential variance*—the risk that the conclusion drawn is so tuned to the specifics of the case that it fails to track anything stable across cases. Two hypotheses should be distinguished:

Bias-variance descriptive (BVD): as the cases to which the reasoner appeals become more specific, the propositions inferred from those cases tend also to become more specific.

Bias-variance normative (BVN): as cases become more specific, the reader ought to infer more specific propositions from them, on pain of epistemic irresponsibility.

BVD is a descriptive tendency-claim about our inferential habits, and, *if* true, admits exceptions. A highly fleshed-out Gettier case *may* still motivate inference to the general proposition that knowledge is not justified true belief (though it need not; see §3.1). More generally, propositions inferred from specific cases may abstract from the particularities.

[27] More precisely, where the total effect is at issue, the reasoner ought to use the finest relevant partition variable that is neither a mediator nor a collider with respect to the causal or metaphysical relationship of interest. Holding fixed a mediator can block the pathway being studied; holding fixed a collider can open spurious pathways.

[28] The threat is evidenced, and a use threat for the experimental, observational and Bayesian variants alike, though what it demands of each differs. The experimental reasoner must screen a variable before holding it fixed by stipulation, the observational reasoner before admitting it to her adjustment set, and the Bayesian reasoner before conditioning on it. It does not arise for the interpretive species, which is not in the business of estimating the effects of isolated variables.

[29] This essay was begun before the publication of Williamson's text *Overfitting and Heuristics in Philosophy* (Williamson 2024). Williamson's sense of overfitting in philosophy is different to mine. His concern is that philosophers can overfit theories to data (comprising verdicts on cases), but that these data may be corrupted by the use of heuristics that, in certain contexts, generate errors. My concern is different; it concerns the granularity of the cases themselves and the propositions inferred from them, not whether the data are reliable.

Nothing as strong as *BVD* is needed though. The analogue to the bias-variance tradeoff in reasoning from examples may still obtain provided only that, in circumstances of reasonable inferential responsiveness, more specific cases tend to elicit more specific inferences. The inferred proposition may take the form of a conditional whose antecedent encodes all the detail specified in the case. Even if the conditional is true, it is unlikely that any unseen future situation will satisfy its antecedent. There is a tradeoff: too many coarse principles— such as 'don't steal' and 'look after your loved ones'—may apply to present situations while yielding incompatible verdicts; too few finely qualified principles may apply to future ones.

BVN is an independent normative hypothesis. It captures the intuition that a reader who infers only a coarse-grained proposition from a case whose details bear on the question at hand has failed to be appropriately responsive to the evidence before her. It may appear supported by the principle of total evidence (*PTE*), which requires that 'in evaluating a hypothesis, you should take account of all the evidence you have' (Barrett and Sober 2020, p. 191). But recall the evidence when reasoning from examples consists in the content of the reader's judgements about cases rather than in the cases themselves. So *PTE* requires neither constructing cases at maximal granularity, nor inferring the maximally specific proposition from a richly detailed one. One can attend to all the detail and still infer a relatively general proposition, provided that that is indeed what the totality of the evidence most supports. Still, case detail may exert a normatively legitimate demand on attention. A conclusion insensitive to details that bear on its truth is, to that extent, inadequately supported by the case.

To see this, and the analogue of the tradeoff in a philosophical setting, consider an extremely detailed specification of a moral problem:[30]

Emma: Emma Bovary is a young bourgeois woman living in the provincial 19th century town of Yonville-L'Abbaye in Normandy. She has one child. Partly out of the desire to realise an idea of the virtuous life inherited from her exposure to Catholicism, partly out of the pressures of patriarchy, and partly out of boredom, she has married a man—Charles Bovary—who hugely frustrates her. She has consumed a variety of Romantic novels, including those by Jacques-Henri Bernardin de Saint-Pierre and Walter Scott. This has had profound psychological consequences for her: the rich and evasive language of those novels has come to represent for her an apparently mythical realm into which she longs to break, but from which she feels perpetually excluded. [...] Rodolphe Boulanger, an ostentatious, 'brutal' and 'intelligent' manager of a nearby estate several years older than Emma, exploits the language of Romanticism to seduce and sleep with her before abruptly abandoning her.

[30] Assume the ellipsis in the case is vast, containing details on the surrounding world described by Flaubert in his novel (Flaubert 2001/1856). In using *Emma* as an example of a highly specific case, I am not claiming, implausibly, that the hypothetical cases to which philosophers appeal when reasoning from examples are or ought to be construed as themselves pieces of literary fiction. To see why this construal is indeed implausible, see (Egan 2016). I use *Emma* because it is a good example of how a precise description of a highly developed fictional world might look.

One could appeal to this case to support a wide range of propositions in moral philosophy and moral psychology, including that lying is wrong, the means formulation of the categorical imperative, and the proposition that Romantic novels can cause in their readers the perspective they go on accurately to describe, and that this fact can be exploited by cynical agents in ways that can have disastrous consequences for others.[31]

In an important sense, using *Emma* only to support the proposition that lying is wrong fails to do justice to the case. A reader who attends properly to the detail will feel pulled toward richer conclusions. But those richer conclusions may be so specific to Emma's situation that they are unlikely to apply straightforwardly to the next moral problem the reader encounters.

The analogue to the bias-variance tradeoff cannot be solved, but it can be mitigated, by ensuring that the proposition inferred from a highly specific case does justice to the details while remaining coarse-grained enough to have explanatory and predictive power in future situations. Intuitively, too, granularity ought not to be increased to such an extent that Machery's criterion (§3.1) is violated.

Three contextual parameters bear on the appropriate granularity of the inferred proposition. First, context fixes the *contrast class*: which alternatives to the inferred proposition are live determines which of the case's details bear on the answer (van Fraassen 1980).[32] Second, the intended *domain of extrapolation*—that is, the future situations to which the conclusion is supposed to apply—fixes how much specificity the conclusion can afford; the denser that domain, the finer the principle tolerated. Third, contexts differ in the relative costs of the two errors. In some inquiries a false general verdict is costlier than a specific verdict that never again applies, and in others the reverse. On *EM*, these parameters jointly constrain which granularity maximises explanatory and predictive fruitfulness. There is no context-free answer, though in a given context, there may be a fact of the matter; and old cases may need revisiting so that propositions of the right granularity for the present context are inferred from them. The check on inferential variance is, moreover, *predictive*. An inferred principle earns credibility by succeeding on cases other than those that generated it, which is defeasible evidence against overfitting. This instantiates the general point of Hitchcock and Sober (2004), that prediction is epistemically superior to accommodation wherever overfitting is a live risk.

Highly granular cases will also tend to change the *kind*, not just the granularity, of proposition we infer. We typically want to infer principles—characterised by Väyrynen (2011) by their ability to support modal claims, their explanatory power,

[31] Flaubert's novel does seem to support this proposition. For instance, the narrator tells us that Emma 'cherchait à savoir ce que l'on entendait au juste dans la vie par les mots de *félicité*, de *passion* et d'*ivresse*, qui lui avaient paru si beaux dans les livres' (italics original, Flaubert 2001/1856, p. 84), and, elsewhere, that she 'voulut se donner de l'amour. Au clair de lune, elle récitait tout ce qu'elle savait par coeur de rimes passionnées et lui chantait en soupirant des adagios mélancoliques.'(Flaubert 2001/1856, p. 95).

[32] van Fraassen's account is of *why*-questions, on which the contrast class is a constituent of the question itself (van Fraassen 1980, p.143); I extend his point to questions of other forms.

and their exceptionlessness. A proposition specific enough to do justice to the *Emma* case will likely exhibit none of these hallmarks, while a proposition coarse enough to be a genuine principle may fail to do justice to the case.[33]

5 Scope

Three questions about the scope of the tradeoff remain. First, the tradeoff might be dissolved by the adoption of an objective Bayesian modelling framework in which granularity choices are not free. Second, the kinds of hypothesis to which the tradeoff applies may not be immediately clear. Third, the tradeoff may appear to warrant a stronger conclusion than the one espoused.

5.1 An objective Bayesian dissolution

One might object that the granularity tradeoff dissolves on a Bayesian picture. Suppose, with the objective Bayesian, that there is a single correct probability function—call it P—defined over the most granular set of propositions we could distinguish. To reason at a coarser grain is then just to *marginalise*.[34] Verdicts at every grain cohere, since each is entailed by the finest. There is no choice to be made about granularity. The tradeoff is illusory.

The objection is helpful but unsuccessful. The tradeoff is epistemological, not metaphysical. It concerns a reasoner who must *choose* a granularity while uncertain which variables are relevant and what the causal or metaphysical structure among them is. A unique P dissolves that uncertainty no more than a unique future dissolves the problem of induction. It tells the reasoner what the right answer *would* be *if* she had it, not how to choose among models when she does not. Model selection is a stubborn problem within Bayesian statistics (McElreath 2016). The granularity tradeoff is its analogue in reasoning from examples.

Moreover, even an objective Bayesian must first specify the propositions over which P is defined. For many moral and metaphysical hypotheses, this is part of what is in dispute. The tradeoff is a feature of our *epistemic access* to the relevant probabilities, dependencies, and structures. It is not a feature of those probabilities and structures themselves.

[33] One might object that the argument of this essay recoils: the reasoning from examples—*Emma*, say—employed in favour of the conclusion that reasoning from examples is vulnerable to the tradeoff is *itself* vulnerable to the tradeoff. Three replies. First, that conclusion is established chiefly by formal results (§§4.2–4.3), empirical findings about how readers treat cases (Small and Loewenstein 2003; Nichols and Knobe 2007) and arguments already in the literature (Kagan 1988); *not* by (the content of) judgements about the kinds of examples the conclusion concerns. Second, where I do reason from such examples, I instantiate the use the essay recommends, and claim only the defeasible support the essay attributes to any well-run instance. This is self-inclusion, not self-refutation. Third, since the tradeoff is a genus threat (§1), *any* essay reasoning from examples must face it. That this one does is entailed by the thesis, not a counterexample to it.

[34] That is, to recover the probability of a coarse-grained proposition by summing P over the distinctions formerly ignored, as one recovers the probability that a fair die lands even, by adding the probabilities it lands two, four, and six.

5.2 A problem for everyone?

The granularity tradeoff affects reasoning from examples deployed in favour of a wide variety of kinds of hypothesis. We have already seen it affect reasoning in favour of *ceteris paribus* claims (*TRI*), existential claims (that common knowledge exists), and modal claims (that switching bodies is possible, or that justified true belief without knowledge is possible). It also affects reasoning for default rules, as regimented in Horty's non-monotonic default logic (Horty 2012). Default rules stipulate that whenever X is established, we ought to conclude Y by default, unless the default is excluded or an undefeated conflicting reason supports a competing conclusion. Such rules are subject to a priority ranking determined in part by specificity: more specific defaults take priority over less specific ones. Case-granularity affects which default rules are discovered, and the specificity of those rules in turn influences their ranking. The tradeoff is thus operative directly and indirectly.

It may initially appear that reasoning for universal generalisations is not affected. This appearance is mistaken for at least two reasons. First, the defended generalisation may be false, the nearest true claim in its vicinity a *ceteris paribus* claim—and reasoning from examples in favour of such claims is, as we have seen, vulnerable to the tradeoff. Second, whether a case constitutes a confirming instance of a universal generalisation can depend on the level of detail at which it is specified. The fact that there is a black raven confirms that all ravens are black; the fact that there is a black raven in the Hall of Atypically Coloured Birds does not (Titelbaum 2022).

5.3 Appropriate granularity

I said the appropriate granularity is relative to an inquiry and a context, and shifts from one ⟨inquiry, context⟩ pair to another. That commits me to there being, for at least some such pairs, an appropriate degree of granularity or set of such degrees. Two questions follow. If the threats of under- and overspecification are as significant as I have suggested, what makes it the case that there is *ever* a fact of the matter? And if there is one, can the reasoner find it?

Recall the metaepistemological picture *EM* from §2, adopted here as a working hypothesis.[35] On *EM*, the facts about (say) an agent's credences are fixed by what the best Bayesian model of her would say about her in the present context. Likewise for granularity. As soon as we theorise at *some* granularity rather than another, we are already modelling; and the appropriate granularities for an ⟨inquiry, context⟩ pair are those of the models in fact best serving the inquiry's aims in that context. Appropriateness is not global: a partition appropriate for one inquiry may be barren for another.

That the tradeoff is unavoidable means only that no granularity is costless. Each pays on one horn or the other. That there is an appropriate granularity means only

[35] The argument is conditional on *EM*. Those who reject it may substitute their preferred metaepistemology. The relativisation of appropriateness to inquiry and context may survive, though its ground will of course differ.

that, for a given ⟨inquiry, context⟩ pair, one grain—or a set of them—carries costs that it is better, by the inquiry's own lights, to bear than any other does. Appropriateness is the best position within the tradeoff. The parameters of §4.4 are how a context prices the two errors: how much coverage the inquiry needs, and how much instability it can tolerate in what it recovers. Moreover, the comparative fruitfulness of candidate granularities is a broadly empirical matter. It is constrained and given content by those parameters. Whether one granularity serves an inquiry better than another is thus discoverable; not always easily, but in the ordinary, fallible way of model selection elsewhere.

Unfortunately, that there is a fact of the matter does not guarantee the reasoner can always attain it. Her identification of the appropriate granularity is defeasible. The sensitivity analyses of §6 are how she canvasses the candidate grains, and the predictive check of §4.4 is a useful corrective on the result. But the predicament of the reasoner from examples is not peculiar to her. Fallible identification is the ordinary problem of model selection; it is no argument for abandoning models, and none for abandoning reasoning from examples. Where no granularity available to the experimental variant serves the inquiry well, the conclusion is that the variant should change (§3.6), not that that method should be given up. Cautious use, then, means use with qualifications—explicit attention to the parameters, predictive checking, readiness to change variant—not rarer use.

6 Conclusion

The granularity tradeoff is unavoidable. I have tried to make clear what getting it wrong in either direction costs, for each main kind of hypothesis one might defend by reasoning from examples. As §5.3 argued, no set of pre-inquiry heuristics settles the appropriate granularity in advance. Just as it may be fruitful to treat cases as scientific experiments in some contexts and as literary narratives in others, coarse-grained methods suit some inquiries and finer-grained ones others (Wilson 2026). For certain inquiries, the right degree emerges only once work is under way.

At first sight, though, three general defeasible heuristics merit further development. First, reasoners from examples plausibly ought to employ two kinds of *sensitivity analysis* to evaluate the stability of reader judgements. The first kind is a *within-grain* check that judgements yielded at a particular grain obtain for all (or at least the most directly relevant) settings *at that grain*.[36] The second kind is a *cross-grain* check that judgements yielded at a particular grain obtain across other candidate reasonable granularities.[37] If those judgements remain unchanged, they should be accorded additional defeasible credibility; if they differ, reasons for this can be investigated and recorded. Second, other things equal, the reasoner might sensibly opt for granularities readers can at least *epistemically access* and keep track of. Third, credible results in

[36] In other words, let $\mathbf{\Gamma}*$ be what the reader *takes* the finest relevant partition variable $\mathbf{\Gamma}$ to be (see fn. 14). The within-grain check is that her judgement is stable across all (or the most relevant) values of $\mathbf{\Gamma}*$.

[37] In other words, the check is that her judgement is stable as she reasonably refines or coarsens $\mathbf{\Gamma}*$ itself; *i.e.*, as she makes $\mathbf{\Gamma}*$ contain more or fewer values.

other fields may inform judgements about the appropriate granularity for a given inquiry.

Most of the threats disproportionately jeopardise the experimental variant, so it is worth asking whether they warrant abandoning it. Its failures fall into the two kinds already distinguished. Some are *use failures*. These are met by more judicious use of the variant. Some use failures arise because it does not realise the controlled-experiment analogy it aspires to (§3.2, the truncation threat of §3.3, §3.4, §3.5). Other failures are *constitutive failures*, entailed by adopting that analogy, or the philosophico-scientific species more broadly, in the first place (the balancing threat of §3.3, the adequacy threat of §3.6).

Take the use failures. These may move the reasoner of the experimental variant to *redouble* her discipline over the method. Three further heuristics, specific to the experimental variant, might aid her. First, if the project is to estimate causal effects, probabilities, or metaphysical relations in the image of causation, the reasoner might screen for confounders, mediators or colliders before adopting finer-grained representations (Schaffer 2016). Second, she may compare hypotheses by fit under a *regularising prior*—one that encodes a preference for simpler models by penalising extreme parameter values (McElreath 2016)—rather than reading a single contrast off a balanced pair. This might be implemented, in the context of reasoning from examples, by adding detail *only* when a good argument that it is doing significant work has been provided. Finally, for the reproducibility and biased-completion worries of §3.5, support may be sought from experimental philosophy, which elicits case-judgements empirically rather than from the armchair.

Take the constitutive failures. Unlike use failures, constitutive failures may rationally motivate abandonment of that analogy. There are two natural destinations of different kinds.

The first is the *observational* variant of the philosophico-scientific species. The reasoner stops constructing artificially balanced pairs, attends instead to richer and more naturalistic cases, and handles confounding by observational means (such as by comparing cases only within cells of the confounder). She thereby trades some internal validity for external. Developing this variant further is not within this essay's scope, but would, I think, be highly fruitful future work.

Still, if the observational variant relocates the genus threat, it does not escape it. Its cases are the rich ones whose hazards §4 documented: they invite divergent attention and divergent integration (§4.1), and exert the pull towards conclusions too finely tuned to generalise (§4.4). It is not free of granularity choices. To control confounding analytically, the reasoner must select an adjustment set—decide which variables to condition on—and conditioning on mediators or colliders corrupts the very effect the analysis was meant to isolate (§4.3). Selecting an adjustment set is a granularity choice by another name. What the variant does secure is greater *transparency*. The

experimental variant leaves confounders to be supplied tacitly by each reader; the observational variant requires them to be stated, and so opened to scrutiny and contest. Where agreement over schematic cases is faux agreement (§4.1), exposing the disagreement is a gain.

The second destination to which the constitutive failures of the experimental variant may move reasoners is the interpretive or hermeneutic engagement urged by Nussbaum, Williams, and—in the philosophy of science—Schickore: abandon the variable-isolating aspiration altogether, and encourage readers to engage deeply and holistically with the historical context of the case or the relevant concepts. This move does away with the philosophico-scientific species. It is beyond the scope of this paper to evaluate comprehensively whether this response succeeds, but it would be a fruitful topic for further research.

Heuristics put forward to reasoners from examples of all stripes should be subject to the same stress tests we apply to any other methodological assumption. Newly aware of the modelling decisions at stake, we ought to continue to reason from examples in a cautionary spirit.

References

Barrett, M. and E. Sober (2020). "The Requirement of Total Evidence: A Reply to Epstein's Critique". In: *Philosophy of Science* 87.1, pp. 191–203. DOI: http://dx.doi.org/10.1086/706086.

Batterman, R. W. (2009). "Idealization and modeling". In: *Synthese* 169.3, pp. 427–446.

Berker, S. (2007). "Particular Reasons". In: *Ethics* 118.1, pp. 109–139.

Bickel, P.J., E.A Hammel, and J.W O'Connell (1975). "Sex Bias in Graduate Admissions: Data from Berkeley". In: *Science* 187.4175, pp. 398–404. DOI: http://dx.doi.org/10.1126/science.187.4175.398.

Bleske-Rechek, A. et al. (2010). "Evolution and the Trolley Problem: People Save Five over One unless the One Is Young, Genetically Related, or a Romantic Partner". In: *Journal of Social, Evolutionary, and Cultural Psychology* 4.3, pp. 115–127.

Bolinska, Agnes and Joseph D. Martin (2020). "Negotiating History: Contingency, Canonicity, and Case Studies". In: *Studies in History and Philosophy of Science Part A* 80, pp. 37–46.

Borges, J. L. (1998). "On Exactitude in Science". In: *Collected Fictions, translated by Andrew Hurley*. New York: Penguin.

Campbell, D. T. (1957). "Factors relevant to the validity of experiments in social settings". In: *Psychological Bulletin* 54.4, pp. 297–312.

Cartwright, N. (2007). "Are RCTs the Gold Standard?" In: *BioSocieties* 2.1, pp. 11–20.

Cartwright, N. and J. Hardie (2012). *Evidence-based policy : a practical guide to doing it better*. Oxford: Oxford University Press.

Chakravartty, Anjan (2017). *Scientific Ontology: Integrating Naturalized Metaphysics and Voluntarist Epistemology*. New York: Oxford University Press.

Crain, G. (2023). "Three Shortcomings of The Trolley Method of Moral Philosophy". In: *Journal of Ethics and Social Philosophy* 26.2, pp. 420–443.

Dancy, J. (1993). *Moral Reasons*. Oxford: Blackwell.

Egan, D. (2016). "Literature and Thought Experiments". In: *The Journal of aesthetics and art criticism* 74.2, pp. 139–150.

Elster, J. (2011). "How Outlandish Can Imaginary Cases Be?" In: *Journal of Applied Philosophy* 28.3, pp. 241–258.

Fagin, R. et al. (2003). *Reasoning about Knowledge*. MIT Press.

Flaubert, G. (2001/1856). *Madame Bovary: mœurs de province; édition présentée, établie et annotée par Thierry Laget*. Paris: Gallimard.

Foot, P. (2001). "The Problem of Abortion and the Doctrine of the Double Effect". In: *The Doctrine of Double Effect*. Ed. by P. A. Woodward. University of Notre Dame Press.

Geddes, A. (2018). "Judgements about Thought Experiments". In: *Mind* 127.505, pp. 35–67.

Gettier, E. (1963). "Is Justified True Belief Knowledge?" In: *Analysis* 23.6, pp. 121–123.

Greco, D. (2023). *Idealization in Epistemology: A Modest Modeling Approach*. Oxford: Oxford University Press.

Häggqvist, S. (1996). *Thought Experiments in Philosophy*. Stockholm: Almqvist Wiksell International.
Hare, R. M. (1952). *The Language of Morals*. Oxford: Oxford University Press.
Hastie, T., R. Tibshirani, and J. Friedman (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
Hesslow, G. (1976). "Two Notes on the Probabilistic Approach to Causality". In: *Philosophy of Science* 43.2, pp. 290–292. DOI: http://dx.doi.org/10.1086/288684.
Hitchcock, C. and E. Sober (2004). "Prediction versus Accommodation and the Risk of Overfitting". In: *The British Journal for the Philosophy of Science* 55.1, pp. 1–34.
Horty, J. F. (2012). *Reasons as Defaults*. New York: Oxford University Press.
Howson, C. and P. Urbach (2006). *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court.
Ichikawa, J. and B. Jarvis (2009). "Thought-Experiment Intuitions and Truth in Fiction". In: *Philosophical Studies* 142.2, pp. 221–246.
Kagan, S. (1988). "The Additive Fallacy". In: *Ethics* 99.1, pp. 5–31.
Kamm, F. M. (1993). *Morality, Mortality, Volume I: Death and Whom to Save from It*. Oxford: Oxford University Press.
— (2007). *Intricate Ethics — Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.
Kinzel, Katherina (2015). "Narrative and Evidence: How Can Case Studies from the History of Science Support Claims in the Philosophy of Science?" In: *Studies in History and Philosophy of Science Part A* 49, pp. 48–57.
Kogut, T. and I. Ritov (2005). "The "Identified Victim" Effect: An Identified Group, or Just a Single Individual?" In: *Journal of Behavioral Decision Making* 18, pp. 157–167.
Kripke, S. (1981). *Naming and Necessity*. Oxford: Blackwell.
Lederman, H. (2018). "Uncommon Knowledge". In: *Mind* 127, pp. 1069–1105.
Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.
— (1983). "Truth in Fiction". In: *Philosophical papers. Vol. 1*. New York: Oxford University Press, pp. 261–280.
Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press.
Mahtani, A. (2017). "The Ex Ante Pareto Principle". In: *Journal of Philosophy* 114.6, pp. 303–323. DOI: http://dx.doi.org/10.5840/jphil2017114622.
— (2021). "Frege's puzzle and the ex ante Pareto principle". In: *Philosophical Studies* 178.6, pp. 2077–2100.
— (2024). *The Objects of Credence*. Oxford: Oxford University Press.
Malmgren, A-S. (2011). "Rationalism and the Content of Intuitive Judgements". In: *Mind* 120.478, pp. 263–327.
Mayo, D. G. (2018). *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.
McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: CRC Press.
McMahan, J. (2002). *The Ethics of Killing*. Oxford: Oxford University Press.

McMullin, E. (1985). “Galilean Idealization”. In: *Studies in History and Philosophy of Science Part A* 16.3, pp. 247–273.
Nichols, S. and J. Knobe (2007). “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions”. In: *Noûs* 41.4, pp. 663–685.
Nussbaum, M. (1990). *Love's knowledge : essays on philosophy and literature*. New York: Oxford University Press.
Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
Pearl, J. (2009). *Causality: Models, Reasoning, and Inference: 2nd Edition*. Cambridge: Cambridge University Press.
Pitt, Joseph C. (2001). “The Dilemma of Case Studies: Toward a Heraclitian Philosophy of Science”. In: *Perspectives on Science* 9.4, pp. 373–382.
Raz, J. (2000). “The Truth in Particularism”. In: *Moral particularism*. Ed. by B. Hooker and M. O. Little. New York: Oxford University Press, pp. 179–204.
Rosenbaum, P. R. (2002). *Observational studies, 2nd edition*. New York: Springer-Verlag.
Saint-Germier, P. (2021). “Getting Gettier Straight”. In: *Synthese* 198.2, pp. 1783–1806.
Schaffer, J. (2016). “Grounding in the image of causation”. In: *Philosophical Studies* 173.1, pp. 49–100.
Schickore, J. (2011). “More Thoughts on HPS: Another 20 Years Later”. In: *Perspectives on Science* 19.4, pp. 453–481.
Schwitzgebel, E. and F. Cushman (2012). “Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers”. In: *Mind and Language* 27.2, pp. 135–153.
Small, D. A. and G. Loewenstein (2003). “Helping a Victim or Helping the Victim: Altruism and Identifiability”. In: *Journal of Risk and Uncertainty* 26.1, pp. 5–16.
Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction and Search, 2nd Edition*. Cambridge: MIT Press.
Steel, D. (2008). *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.
Thomson, J. J. (1985). “The Trolley Problem”. In: *The Yale Law Journal* 94.6, pp. 1395–1415.
Titelbaum, M. (2012). *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford University Press.
— (2019). “Reasons without Reasons For”. In: *Oxford Studies in Metaethics Volume 14*. Ed. by R. Shafer-Landau. Oxford: Oxford University Press, pp. 189–215.
— (2022). *Fundamentals of Bayesian epistemology. 2, Arguments, challenges, alternatives*. Oxford: Oxford University Press.
Tooley, M. (1972). “Abortion and Infanticide”. In: *Philosophy and Public Affairs* 2.1, pp. 37–65.
Uhlmann, E. L. et al. (2009). “The Motivated Use of Moral Principles”. In: *Judgment and Decision Making* 4.6, pp. 476–491.
Van Fraassen, B. (1980). *The Scientific Image*. Oxford: Clarendon Press.
Väyrynen, P. (2011). “Moral Particularism”. In: *Continuum Companion to Ethics*. Ed. by C. Miller. Continuum, pp. 247–260.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.
Wilkes, K. V. (1988). *Real people : personal identity without thought experiments*. Oxford: Clarendon.
Williams, B. (1973). *Problems of the self : philosophical papers, 1956-1972*. Cambridge: Cambridge University Press.
— (2006). *Philosophy as a humanistic discipline*. New Jersey: Princeton University Press.
Williamson, T. (2022). *The philosophy of philosophy*. Hoboken, NJ: Wiley Blackwell.
— (2024). *Overfitting and Heuristics in Philosophy*. New York: Oxford University Press.
Wilson, J. (2016). "Internal and External Validity in Thought Experiments". In: *Proceedings of the Aristotelian Society* 116.2, pp. 127–152.
— (2021). *Philosophy for Public Health and Public Policy: Beyond the Neglectful State*. Oxford: Oxford University Press.
— (2026). "Williams, Dewey, and the nature of value inquiry". In: *Philosophy* 101.2, pp. 511–538.
Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.