Heuristics and Overfitting in Science, Philosophy, and Ethics

Alexander Bird, University of Cambridge

Introduction

In this paper I examine the case made by Timothy Williamson that (a) our judgments ('intuitions') concerning philosophical examples (vignettes, hypothetical cases, counterexamples, etc.) are often generated by heuristics; (b) they are therefore not perfectly reliable; (c) by assuming our judgments to be correct in all cases, we are led to adopt a false philosophical theory that is overly complicated—this is a case of overfitting the theory to the data; (d) by recognizing that some judgments of cases may be false, we see that the correct theory may be a simpler one.

My approach will be to look at an area of philosophy where Williamson's approach might be thought to be particularly promising: metaethics. As we will see, some utilitarians have held views with striking similarity to Williamson's. Nonetheless, we find that making the case as regards utilitarianism turns out to be more difficult than we might have hoped. The point of this exercise is to articulate some lessons for Williamson's project overall from this particular potential application. One conclusion is that science is in a position to put many more constraints on how we treat our data. The curve fitting problem with its potential for overfitting is much more notional in science than philosophers tend to suppose. As a result, proving Williamson's case is faced with obstacles, which, if not insurmountable, are at least challenging.

Williamson's view of heuristics and overfitting

According to Williamson (2020, 2024) some overly elaborate philosophical theories are the result of *overfitting*. Theorists have taken 'intuitions' to be solid, incontrovertible data that any theory must accommodate, on pain of refutation.[1] In Williamson's view these judgments are generated by heuristics. A heuristic is an easy-to-use rule, mechanism, or process for reaching a conclusion. Rules of thumb are heuristics that are consciously employed to achieve quick answers to a question or problem—for example a doctor may use the rule of thumb 'When a patient can bear weight on a leg, it is not broken' (André *et al.* 2003). The trade-off for speed and ease of use is that the answer may be only approximate or may simply be false in some (but not too many) cases. Clearly some stoical or drugged individuals can bear weight on a broken leg, but they will be a small minority. Heuristics include rules of thumb, but need not be conscious. The heuristics and biases research programme is interested specifically in the

[1] I will continue to use the term 'intuition' although Williamson abjures it, for the judgments that philosophers (and others) make regarding particular hypothetical cases. I do *not* imply any particular view of what generates these judgments (e.g. a faulty of intuition), nor that they have any particular character (that they are 'inuititive' rather that sensitive to reason or experience). I believe that it is sufficiently clear what we are talking about.

unconscious mechanisms of judgment (Tversky and Kahneman 1974; Goldstein and Gigerenzer 1999). One of the best known heuristics is the recognition heuristic:

> if one of the two objects is recognized and the other is not, then infer that the recognized object has the higher value [and choose it] (Goldstein and Gigerenzer 1999: 41).

In a well-known instance of a *heuristic*, the recognition heuristic is at work when people are asked to judge which of two cities is the more populous. When faced with the names of two cities, only one of which they recognize, then the recognition heuristic prompts subjects to guess that the city with the name they recognize is the larger. Subjects in this position can be more reliable than those who recognize both names (or neither). A *bias*, in this context, is also an unconscious influence on judgments, but one which tends to make those judgments less reliable. Confirmation bias, for example, occurs when our attention is drawn selectively to evidence that supports a favoured hypothesis and away from evidence that undermines it, leading to an unreliable estimate of that hypothesis's probability of truth.

The fitting of theories to our judgments is appropriate only if those judgments are reliably true. If, however, our judgments are generated by heuristics, we should expect some of those judgments to be mistaken. If we attempt to account for *all* such judgments perfectly in our theories, then we will be aiming to accommodate false data points. This will lead to a false theory—and an unnecessarily complicated theory. This is like a scientist trying to fit a curve exactly through all her data points, even though some of those points are not accurate reflections of the underlying function—this is overfitting. Furthermore, those data points (the erroneous judgments) will mistakenly be taken to refute the true theory, which may be much simpler than the false and complicated theory. Williamson argues that this occurs in philosophy. His *Suppose and Tell* (Williamson 2020) is an extended discussion of how overfitting has led to the erroneous rejection of the simple theory that the correct semantics for conditionals is given by the material conditional. The latter has long been thought to be false on account of its conflict with our intuitive judgments. More complex semantics for the conditional are required in order to accommodate those judgments. *Overfitting and Heuristics in Philosophy* (Williamson 2024) presents this approach in more general terms, while arguing in particular that the development of hyperintensionalism is a product of overfitting.

Why have we treated intuitive judgments as infallible? One reason is their root in the analytic tradition which took the point of analysis to be the articulation of concepts in other, more basic terms. It was assumed that one could not be mistaken about the nature of our own concepts, at least not about their correct application in cases where no relevant factual information is lacking. There are many things wrong with this argument, but it does explain the practice of treating (sufficiently reflective) intuitive judgments as generating perfectly reliable data points that must be accommodated.

In Williamson's opinion, the aim of philosophy is not to analyse concepts but to generate substantive theories about the phenomena. We do not aim to analyse the concept of knowledge; rather we are interested in understanding knowledge itself. In this respect philosophy is much like the sciences, investigating phenomena 'out there'. Philosophy is continuous with the sciences. So we should not expect our methods to be infallible.

It is important to note that the heuristics claim in philosophy is more difficult to demonstrate than similar claims elsewhere. Elsewhere we can provide strong evidence that we use heuristics or are subject to biases, since we can see that humans make judgments and then we can see whether these judgments are correct or mistaken. We can formulate and test hypotheses about why we see that pattern of correctness and error. In the case of the recognition heuristic we can know whether subjects are making accurate or inaccurate judgments since we have independent information on the sizes of cities. So we know that subjects are more reliable when they recognise one name but not the other, and we know that they typically choose the recognized name when in this position. And so we have good evidence that the hypothesized heuristic is in play and is often reliable. Note that the hypothesis that a heuristic is at work in making some judgments of a certain kind should not necessarily imply that all judgments of that kind are formed by that heuristic. The heuristic may kick in only under particular circumstances, e.g. when matters are uncertain, when the effort required to get things right is too high, or when a swift judgment is required. The recognition heuristic is not needed if asked 'which is larger Beijing or Nagoya?', if one knows that the three largest cities in the world are Chongqing, Beijing, and Shanghai—even if one has never heard of Nagoya.

In philosophy, however, the intuitive judgments are judgments about imaginary scenarios. And they are typically counterfactual or subjunctive judgments. So it is not normally possible to assess, independently, the reliability of these judgments. This means that it is not clear that our intuitive judgments are often erroneous. Which is an additional reason why the heuristics approach was not an obvious one to take—for it is not obvious that we are making errors. This invites two questions. First, why should we think that heuristics are ever in play in philosophy? And secondly, when we do think that they are in play, how should we decide which heuristics are in play?

Regarding the first question, there are two reasons. The first is general. Once we accept that the target of our investigation is not the concept of X but X itself, then we lose the reason given for thinking that our judgments must be highly reliable. In circumstances where we need to make judgments but doing so accurately is difficult, a heuristic, if one is available, will often be in play. The second reason is that we may take the fact that it is impossible to settle on a theory of X, that promising theories turn out to be subject to counterexamples, that the best theories seem to be highly complex and more complex than their predecessors, as evidence that some of our judgments are false and that the complexity and liability to counterexample is the result of overfitting theory to data. It is plausible, but by no means certain, that a heuristic may be generating the false data.

The second question is more demanding. It is one thing to suppose that a heuristic may be guiding our judgments and another thing to identify which that heuristic is. Given that in philosophy we do not have an independent handle on truth and falsity of the relevant judgments, it will be methodologically difficult to decide between different options. Let X be some phenomenon regarding which we make intuitive judgments in philosophy, such that it is difficult to find a theory of X that fits all those judgments. Our best current theories are highly complex. Let us hypothesize therefore that there is some heuristic H that is driving some of those judgments. Is it that X is in fact very simple and H is very unreliable? Or that X is less simple and H is more reliable? Or maybe X is in fact very complicated and our judgments about X are not often wrong?

The case where X is simple and H generates many false judgments is most likely where X is (or is akin to) a natural kind. We often use heuristics and cues to identify natural kinds—typically perceptual appearance, but those can lead us astray, since those heuristics and cues are not perfectly aligned with the underlying essential properties that identify the kind: so we may take fool's gold or Munz metal to be gold, or paste to be diamond, and blood serum or twater to be water. So we might expect the same to occur when the kind in question is open to philosophical investigation.

In my view *knowledge* is a natural kind of mental state (Bird 2024). On the other hand, it is *not* Williamson's view that the failure to find a correct analysis of knowledge that satisfies all intuitive cases is due to those intuitions being generated by a fallible heuristic. Williamson's knowledge-first view is motivated in part by accepting that our intuitions *rightly* refute all the proposed analyses of knowledge. So, one might wonder whether one could defend the tripartite, JTB account of knowledge—as some continue to do, albeit for other reasons (de Grefte 2023)—on precisely those grounds.

In other cases, however, X may be nothing like a natural kind. There may be no unity to all the Xs. Perhaps all that links the Xs is that they fall under the concept 'X'. It seems entirely plausible that there are family resemblance concepts. In which case there is no simple account of what Xs are and so no room for a heuristic to model such an account nor for it to make errors—the rule 'take something to be X if it bears sufficient relevant similarity to accepted exemplars of Xs' is not so much a heuristic as a correct description of the semantics for the concept of X. So a plausible view of the concept 'art' is that it is a family resemblance concept (Kaufman 2007). Furthermore, one might hold that it is a concept that undergoes shifts over time as, for social reasons, society agrees (or not) to allow borderline cases to stand as new exemplars. We know that the brain is good at spotting similarities and patterns, including those with a family resemblance structure. That some concepts should have that character is not surprising. The fact that we cannot find a single analysis that fits all judgments, *e.g.* about whether something is an artwork, should not be taken to show that we are using a fallible heuristic.

We will return to this question of deciding when a heuristic is operating and what that heuristic is.

The prima facie case for a parallel with ethics

The general approach that Williamson recommends is one that recommends itself to ethics in particular. First, ethics is widely not taken to be a matter merely of the conceptual analysis of concepts such as 'good' or 'just'. Rather, it is taken to be aiming at a substantive theory of what is right. The view that philosophy should aim at the provision of substantive theories is contentious, one that runs counter to much of 20th century philosophy. But ethics has long been an exception to the latter, with most theorists rejecting conceptual analysis. As Lazkowski and Finlay (2017: 536) write:

> Despite being considered the traditional and prototypical philosophical method, conceptual analysis is an approach to philosophical inquiry with a chequered past and tarnished reputation. This is particularly true in metaethics . . . In recent decades metaethicists have often consigned it to the trashpile of philosophical mistakes . . .[2]

The second reason concerns the nature of our intuitions in ethics. Our intuitions in ethics do seem to generate judgments that are difficult to reconcile into a single theory. For example, one clear manifestation of this is the difficulty we have in forming a theory that comprehensively captures responses in all trolley cases. More generally, all ethical theories seem to have counterexamples, widespread moral intuitions that are difficult to justify within a larger theory. Notoriously, utilitarianism seems to recommend carving up the healthy man to distribute his organs to several patients in urgent need of transplants, while deontology seems to require us to tell the truth to the would-be murderer. (More subtle counterexamples are available.)

The preceding paragraph noted the evidence for the heuristics hypothesis about moral judgments that comes from the clash between intuitions and theories, in those cases where ethicists have, for the most part, tended to favour the intuitions. We should note that further evidence comes from cases where many people have strong moral intuitions, yet those intuitions seem wrong to many on more considered reflection. Haidt et al. (2000) provide a vignette describing incestuous sex between consenting and caring adults. The case is set up so that it is clear that no harm has come to anyone and indeed that those concerned are better off than they were. Yet the majority of people asked about this case say that the siblings did wrong. Many of us will feel some degree of aversion to what is described. And that aversion will not be phenomenologically different from the aversion we feel at the description of some genuine injustice. In other cases, subjects feel that there is something morally wrong about performing cost-benefit analyses with regard to people's lives and well-being—for example,

[2] As Lazkowski and Finlay (2017) go on to discuss, there are exceptions to this, including Moore and Jackson. Hare would also be an exception.

deciding whether a road improvement scheme will save enough lives to be worth implementing or whether a new drug will prolong the lives of cancer patients for long enough to be recommended for use by a national health service. Although the majority of people feel very negatively about such calculations, further reflection should make it clear that we have no alternative than to make such calculations. Indeed reflection tells us that it would be wrong not to consider such facts, other things being equal, since to fail to do so runs a high risk of making decisions that will lead to more death and more suffering. So it looks as if these kinds of aversive feelings or negative affect are a source of moral judgment that is at least sometimes unreliable. Which is what we would expect if they are the product of heuristics. Indeed Sunstein (2005: 356–7) argues that there must be heuristics operating in ethics for these reasons. Of course, it should be noted that anti-consequentialists argue that they have responses to Haidt's cases. Nonetheless, one does not have to be a committed consequentialist to think that a cost-benefit analysis should inform, even if not always determine, decision-making.

Ethics is, of course, like other areas in philosophy, such as epistemology and metaphysics, where there is no independent source of verdicts on our judgments that can be used to tell us which are accurate and which are not.[3] Correspondingly one might think that similar methods for overcoming this problem would operate in ethics and in other areas, including those discussed by Williamson.

Utilitarianism and heuristics

Of the ethical theories, utilitarianism looks to be one whose defence might most benefit from this approach, since it seems to suffer from the most glaring gaps between its recommendations and the moral intuitions of many: framing the innocent to prevent a deadly riot, harvesting organs from the healthy man to save several sick people, and so on.

One problem for utilitarianism is that it is difficult to know what the utilitarian calculus in fact demands. Even if one can see the immediate effects of one's action it can be difficult to know what effects it will have further into the future. Strictly, all effects determine the rightness of one's actions. Perhaps the motions involved in saving the drowning child will set off a butterfly effect whose resulting tornado will kill dozens of people. In which case it would be a wrong action.[4]

So utilitarianism looks ripe for the following form of defence along the lines of Williamson's strategy. Implementing the utilitarian calculus is difficult and indeed prone to errors of miscalculation and ignorance. So a heuristic would be helpful here, one that delivers the correct utilitarian verdict in enough cases. One can hypothesize that our intuitions are the products of one or more heuristics. Like all heuristics, these heuristics get things wrong in

[3] One might argue that outright logical inconsistency is an independent source of refutation of some claims—for example Frege's Axiom V. But examples such as this are somewhat rare.
[4] There are subjectivist versions of utilitarianism that avoid this problem by tying rightness to expected utility.

some cases.[5] These are the instances where intuition diverges from the requirements of utilitarianism.

(An important question that this hypothesis would need to address concerns why it is that we possess such a heuristic—that it would be helpful does not guarantee that it would exist. How one answers this question rather depends upon one's moral epistemology. One possible answer assumes that there are non-heuristic routes to moral knowledge. The mind's pattern-recognition capacity observes that many moral truths fit a simple pattern, or, more likely, one of a set of simple patterns, and this pattern, or these patterns, generate the heuristic(s).)

Note that we are not bound to follow our intuitions in our final judgments, but they inevitably intrude upon our thinking. Stephen Jay Gould (1991: 469) likens such intuitions to the voice of a homunculus in his head—difficult though possible to ignore. A heuristic is plausibly at work in cases such as Linda the feminist bank teller, where Gould says 'I know [the right answer], yet a little homunculus in my head continues to jump up and down, shouting at me—"but she can't just be a bank teller; read the description".' This allows a response to Williams's famous discussion of Jim and the Indians (Williams 1973). Jim eventually comes to what Williams agrees might well be the right decision, to kill the one villager to save the other nineteen. Williams points out that Jim will feel reluctance, coming to this decision and indeed implementing it requires a lot of effort (1994). Whereas the demands of the utilitarian calculus (if we consider near term consequences only) are clear. So the fact that Jim finds the right decision difficult to reach and to act on is itself evidence against utilitarianism. But if in fact we are not guided by the utilitarian calculus itself but instead by some heuristic that approximates to the utilitarian calculus, then we can see why Jim is reluctant. For this may be one of the occasions when the heuristic may be directing Jim in the wrong direction. As in the case of Gould it may be true of Jim that even though he knows what the right thing to do is, the heuristic/homunculus is jumping up and down in his head shouting 'Don't kill the poor guy! Just look at him! What kind of person are you—a brutal killer? How could you even think of doing such a thing?'

That there is a divergence between the demands of utilitarianism and our ethical intuitions has long been recognised, indeed the founding fathers of utilitarianism, Mill and Sidgwick, were well aware of it. Sidgwick (1874: 238–9) in particular noted that the utilitarian 'code' is difficult to implement and erring attempts to do so may in fact lead to harm. So it might be best, from the utilitarian perspective, that most people are not utilitarians. They should follow common-sense morality, maybe backed up by a community's moral codes that are not utilitarian, since doing so will in fact lead to greater utility than inadequate adherence to utilitarianism. That is so because, he says, 'common-sense morality roughly coincides with the utilitarian code'. While it would be anachronistic to attribute to Sidgwick the view that common-sense morality is a heuristic, it is clear that he thinks that common-sense morality does what a heuristic does: namely it is an approximation that is much easier to use than the strict rule and which is therefore used because it achieves the aims of the strict rule more

[5] For an argument that our moral intuitions are not always reliable, see Sinnott-Armstrong 2006.

closely than attempting to follow the strict rule directly would. Sidgwick goes on to argue that the virtues should be seen in this way, as should the idea, which has no place in strict utilitarianism, that some acts are supererogatory. More recently, but in the spirit of Sidgwick and Mill, ethicists have distinguished between the *criterion of right* and a *decision procedure*. The latter, how we decide what is right, might not simply be the direct, conscious application of the criterion of right, the rule that determines what is right. It is plausible, then, to see the decision procedure corresponding to a criterion of right as a heuristic that aims to approximates to the criterion of right.

This view, it should be noted, will require that some heuristics are conscious. The heuristics and biases studied by psychologists are typically unconscious. But they need not be. The ten commandments and the Golden Rule along with other moral precepts are conscious rules that people do use to guide decisions. Of course, that they are conscious heuristics does not entail that the people using them think of them as heuristics.

Is the heuristics and overfitting defence of utilitarianism plausible?

Peter Singer (2005) is one utilitarian who agrees that our moral intuitions are not guided by the utilitarian calculus. He endorses the view of Greene (2002) that one central motivator in moral decision making is whether or not the subject has a personal connection with those who are affected by the subject's actions. So a well-known pair of trolley cases contrasts the case where the subject must divert the trolley by moving the points/switch so that it kills one railway worker rather than the five who will die if he does nothing, with the case where the subject can save the five only by shoving a very large person off the footbridge ahead of the oncoming trolley. People tend to find the former acceptable and the latter unacceptable. Greene's view is that the physical proximity of the large person on the footbridge creates a personal connection with him that is lacking with the distant lone railway worker. Personal actions with harmful outcomes elicit in us a strong aversive emotion; impersonal harmful actions, much less so. Call this 'G'. Could G be a moral heuristic for utilitarianism?

Prima facie it might be. It would explain this trolley case. It would also explain why people can plan administrative actions, whether with good intent or ill, that will harm many individuals, but would not be able to bring themselves to harm a person standing in front of them. It might explain the Jim and the Indians case. For in selecting the single villager, Pedro forges a connection between Jim and the villager, and if Jim were to shoot him that would be a very personal interaction. These connections are weaker or absent in the relationship between Jim and the other villagers who would be shot by Pedro if Jim does nothing. It might also explain the intuitively felt difference between acts and omissions. Directly causing a harm feels worse than failing to act to prevent a harm. The former is more likely to include a closer personal connection lacking in the latter.

Furthermore, one might think that this heuristic might manage some approximation to the utilitarian rule. First, those with whom one has a personal connection, including physical

proximity, are those whose actions are most likely to affect. And these people are those with whom it is most likely that one's actions will have certain kinds of complex consequences, for example the promotion or destruction of social relations. So a disposition that focuses on personal, proximal connections and makes one averse to harming people with whom one has such relations is a rule that is likely to go some way towards maximising utility.

It might be objected that this disposition operates through the emotions, which does not make it look like a rule of thumb. First, one might respond that the outcomes are indeed beliefs: the belief, for example, that it is wrong to kill the large gentleman on the bridge. Secondly, even if such judgments were the effects of prior emotions, we should not deny on these grounds that there is a heuristic in operation. Unlike many of the heuristics that lead to beliefs, we would expect a moral heuristic to have normative or motivational force, and so it would not be surprising that it operates via an emotional state.

A more substantive problem is that Greene's heuristic G looks, on further examination, not to approximate the utilitarian rule very well. Let R be the correct rule, but one which it is difficult to apply correctly in some cases. If H is a good heuristic for R, it should kick in when R is difficult to apply and not when R is easy to apply. The use of the recognition heuristic in estimating the relative sizes of cities is like this. It kicks in when one city's name is not recognised—a condition where it is very difficult to judge directly that one city is the larger one. But if one knows the size of both cities then the recognition heuristic is silent and one's knowledge makes the accurate judgment for one. G is not a very good heuristic by this standard. Let's call the rule of utilitarianism U. We would expect G to deliver an answer when we find applying U difficult. But G does not do this. Focussing on G does not help us solve the less tractable cases, for example where one might have to weigh numerous different kinds of harm and good against one another. On the other hand it does seem to be in effect in cases such as the trolley cases where U is very clear in its demands. And in such cases it gives the wrong result. G tells us not to push the large person off the bridge, even though U demands that we should. In this case U is very easily applied.[6] Of course, we are more inclined to get the right answer by U's lights in the points/switch case. Which shows that we can use U to get the right result in a case that is computationally the same as the bridge case. The latter shows that we don't need a heuristic for U in cases of this (low) degree of complexity. So why does G kick in here and why does it give a wrong result when it does?

It might be replied that this is a harmful side-effect of the cases where G is effective, like the continuing effect of the representativeness heuristic in the Linda the bank teller case—Gould's homunculus. There are two problems with this reply. First, I think that Gould overstates the effect of the homunculus. He says that it is operative even when he knows the correct answer. But it is noteworthy that the case is presented to subjects in such a way that

[6] Though it could be argued that the utilitarian calculation if done properly is always too difficult. Because, in the trolley problem for example, you would have to assess the effects on happiness of each of people you don't know, plus all the other people you don't know who would be influenced. Of course, you can use a rule of thumb—treat all lives as equally valuable—so save the greater number. But this is itself a heuristic.

the two relevant questions about probability—'how probable is it that Linda is a bank teller?' and 'how probable is it that Linda is a feminist bank teller?' are separated by intervening questions that prevent the subjects from seeing them together. My guess is that the rate of inconsistent answers would fall if these are put side by side, and fall even lower if these are the only two questions. So I don't think that the representativeness heuristic operates—or at least operates with much diminished force—when the correct answer is easy to spot.[7] And secondly, the representativeness heuristic is often a good one to use in many cases when one does not have access to the statistics one needs. But the proposed heuristic G does not have a range of cases when it is reasonably effective while contrasting with the difficulty of applying U correctly.

It is not implausible that the mind does implement something like G. It would be evolutionarily advantageous to care about those with whom one has a personal, geographically proximal relationship. But if we accept that U is the right rule, then the effect of G is to interfere with our judgments and make them worse. It looks as if we would get things right (we would shove the big guy off the bridge) if it were not for the interference of this emotional impulse affecting our judgment. So G does not look like a heuristic at all. It is not helping deliver utilitarian outcomes, quite the opposite. It is instead, therefore, rather more like a bias.

Maybe we have looked in the wrong place for the heuristic. A better heuristic would be: Do not kill people! On most occasions this heuristic will not diverge from U. Since the utilitarian calculus is hard, we can get things wrong if we attempt to implement it. For example, perhaps U seems to suggest that it is a good idea to carve up the passing stranger to distribute his organs to several otherwise terminal patients. But perhaps we need to take into consideration the effects of such behaviour on other people—e.g. people arming themselves against becoming the victims of utilitarian organ harvesters. It may not be clear and we may err. So it is useful to have a heuristic to avoid such problems, at least for the most part. 'Don't kill people' is a maximin heuristic. Using it we can avoid the worst mistakes of misapplying U, even if we miss out on some opportunities to make the world a better place, including missing out on shoving the big man off the bridge. After all, simple trolley cases do not arise very frequently.

This heuristic, however, does not explain our greater willingness to pull the switch to divert the trolley even though that does involve killing a person. But the utilitarian may reply that it does correctly imply that we should feel some degree of reluctance. Furthermore, the utilitarian may say, the structure of the problem, with there being near symmetry in the track layout and switch and the fact that the trolley is going to kill people either way make it easier for the subject to regard the switch case as at least less of a killing by the subject and more of a

[7] In the case of Jim and the Indians I said that the heuristic does continue to exert some force, even when he can easily see that it is better to shoot Pedro. I think that the difference is explained if the heuristic in the latter case operates through the emotions, in which case it will be less easily dismissed. Whereas in the Linda case we have two sources delivering judgments.

choice of which group the *trolley* is going to kill. Clearly "Don't kill" cannot be the only heuristic in the service of U. For killing or not is not the only moral question to face us. There will have to be many heuristics working to implement an approximation to U in multiple respects.

Maybe then, it is possible to find some combination of heuristics and biases that is a reasonable approximation to U. Clearly we have only scratched the surface of the possibilities. What this discussion does reveal, nonetheless, is that the heuristics approach does not deliver a quick victory for utilitarianism. Indeed, one prominent utilitarian, Peter Singer, holds that Greene's work on the nature and origins of our moral intuitions shows that our moral intuitions should not be relied on at all. Indeed Singer's view is that utilitarianism is very demanding indeed—we should give most of our money to the poor, we must give equal consideration to future generations, and so on. If that is what U demands, then clearly common sense morality is no approximation at all to U and so not a heuristic. The heuristics and overfitting approach to utilitarianism will at best be suitable only for a less demanding version of utilitarianism.

Singer's and Greene's work raises another point that deserves consideration. In their view common sense morality is the result of evolutionary forces with a social overlay. In effect they accept an evolutionary debunking argument about common sense morality. Let's say we are utilitarians of a more moderate kind than Singer. Then the question should arise, where does common sense morality come from, such that it can be a heuristic for utilitarianism? The evolutionary argument suggests that there is no reason why common sense morality should track the correct utilitarian code even approximately. For H to be a heuristic for R, our use of H needs to be explained by the fact that it tracks R. But that is unavailable or at least unexplained in this case.

This section has considered the case of utilitarianism because, prima facie, it seemed ripe for the heuristics and overfitting approach. Furthermore utilitarians have historically suggested something similar, that common sense morality is an approximation to the utilitarian calculus that we cognitively limited creatures find easier and more effective to use.

To make this defence of utilitarianism stick, various things need to be done:

(i) The heuristics employed in common sense morality need to be identified—a breezy reference to 'common sense morality' will not do.[8]
(ii) It needs to be shown that these proposed heuristics do correctly predict our judgments in the cases where those judgments conflict with utilitarianism.
(iii) It needs to be shown that we use the heuristics *because* they are usually a good approximation to the utilitarian calculus.

[8] That said, it might be that our ethical judgments are generated in a rather unsystematic way by perceived similarities to salient exemplars or by a rag-bag of heuristics whose application is contextually sensitive.

(iv) It needs to be shown that these proposed heuristics are operating as heuristics, that is to say, they generate our ethical judgments because applying the utilitarian calculus is too difficult; if it is easy to apply the utilitarian calculus in a case, then it is unclear what the heuristics hypothesis is explaining.

Our discussion shows that this is not straightforward. That need not be devastating news for the heuristics approach in general. Sometimes a prima facie good case will not pan out well on inspection. Nonetheless, by generalizing (i)–(iv) we can learn lessons for what a successful implementation of the heuristics and overfitting programme will look like.

The methodology of the heuristics and overfitting

We have sketched how the utilitarian might implement the heuristics and overfitting approach of Williamson in ethics. A plausible psychological disposition, G, turns out to be less of a heuristic and rather more like a bias. We may be able to find a suite of dispositions distilled out of common sense morality that might fare better as a heuristic for U, but this is not straightforward. Indeed, the question of whether a disposition is a heuristic or a bias does not itself always have a straightforward answer. In effect, Singer holds that all our moral dispositions are biases. Mill thought that there were two areas where common sense morality got things badly wrong—the desert (or retribution) theory of punishment and the place of women in society. Mill explained our disposition to see punishment as the right and required response to wrongdoing in terms of our inherited animal instincts. In Mill's eyes this disposition was very much a bias. Whereas subsequent utilitarians have maintained that the long-run effects of desert-based punishment regimes do tend to lead to a high degree of utility. For the latter, the disposition is not a bias but a heuristic.

We should note that it is not straightforward to determine whether or what heuristic is in operation. The recognition heuristic is the paradigm heuristic. It is assumed by Goldstein and Gigerenzer to apply immediately in cases where its conditions are met (one object is recognized and the other is not): '[the] search terminates as soon as recognition has been assessed for both objects. The decision is consequently based on only one piece of information, recognition.' (Goldstein and Gigerenzer, 1999: 57). In fact it turns out that this is rarely the case. If the recognition heuristic were at work as Goldstein and Gigerenzer suppose, then we would expect subjects to choose the one recognised city as frequently when it is in fact the larger city as when it is not in fact the larger city. But researchers find that this is typically not the case and that subjects reject the one recognised city more frequently when it is in fact the smaller one than when it is the bigger one. When additional information is available, subjects do use that information. When one recognizes a city one may well possess further information about it. Furthermore, I suggest, there may well be cues also about the unrecognized city. And this information and these cues may well be important in determining a subject's choice. Let us say you are asked to decide which is bigger: The Vatican City or Zhengzhou. You recognize only the first. You may (and should) answer "Zhengzhou" nonetheless. First, you know that the Vatican City is tiny. Secondly, the name

"Zhengzhou" looks Chinese to you and you know that China has many very populous cities. If people do use additional information and cues as this suggests, then it is questionable whether the recognition heuristic really exists as the heuristic that Goldstein and Gigerenzer propose. An alternative hypothesis is that there is just unconscious reasoning of a second-order kind 'I recognize this one but not that one; that's evidence that it is more significant; so I'll guess that it is larger'. That second order reasoning can be integrated with (and over-ruled by) reasoning with any other information that is available. Evidence for this hypothesis comes from considering how we use a heuristic. Clearly we do not have a precise heuristic just for the size of cities: 'if asked to judge the size of two cities, and recognise the name of just one, judge that this one is the larger'. Rather this is an instance of a more general heuristic, which Goldstein and Gigerenzer (1999: 41) articulate thus: 'if one of the two objects is recognized and the other is not, then infer that the recognized object has the higher value [and choose it]'. The term 'value' here is vague and is not even especially appropriate: one might use this heuristic in deciding which of two named people is more powerful or wealthy. So 'significant' might be more appropriate. Whatever the general term used is, the relevance of the heuristic will be context specific: it can help in deciding which of two cities is the bigger, but one would not use it to decide which of two people is the bigger. So in applying the more general heuristic, one must have a sense of what counts as relevantly significant ('valuable') given the question asked. Some reasoning must be going on that tells one that recognition is likely to be correlated with personal wealth or power but less likely to be correlated with body mass, or with the population of a city but not its annual rainfall. If reasoning (conscious or unconscious) is happening in this respect, why not throughout?

From the foregoing we should draw the conclusion that even in the most well-studied cases, the nature, operation, and even existence of a heuristic are not methodologically straightforward to establish. These problems are exacerbated severalfold when considering ethics and philosophy more generally, because we have little in the way of an independent test of whether a subject's judgments are correct or not. In ethics and other areas, one common approach to problems of this sort is to seek a reflective equilibrium. The latter allows some evidential value to intuitions about particular cases, but they are not regarded as infallible. The task then is to find a theory that can accommodate a suitably large proportion of the cases while being plausible in itself. A theory that is overly complex as a result of seeking to accommodate all or even most intuitions will be implausible. A theory that is simple but which captures few intuitions will not be well supported by the evidence. Rawls explicitly considers this approach to be akin to what goes on in science. The heuristics and overfitting approach may be seen as a variant of the reflective equilibrium approach where appeal to heuristics is required to explain away the intuitions that need to be rejected or revised because they conflict with the theory. As such, there is a danger that the heuristics and overfitting programme will share some of the methodological weaknesses of the reflective equilibrium approach. The main worry, in my view, is that we may have no independent reason to locate the fulcrum of this equilibrium in one place or another. If we place more weight on the credibility of our intuitions, then the equilibrium may result in a more complex theory. If we place less weight on them, we get a simpler theory. But there is no general rule

for determining the correct weighting. Different weightings might be justified by different theories of what intuitions are and of what kind of theory, if any, we are looking for. At one extreme, for example, moral particularism denies that there is any correct single unified theory. In which case we would put little weight on the value of such a theory and its ability to overrule a subset of intuitions.

The heuristics approach does, however, add a further constraint to the method of reflective equilibrium. Others have recognized the demand that we not simply ignore some individual judgments but seek to explain why we make them. The heuristics programme makes that demand specific. And the desiderata listed at the end of the previous section are, as we have seen, not trivial to satisfy. Above I suggested that judgments regarding city sizes might be produced by unconscious second-order reasoning. In the light of this we should add a fifth demand:

> (v) It needs to be shown that the judgments in question really are produced by the heuristic, H (for the rule R), as opposed to unconscious (and sometimes fallible) reasoning with R itself.

Curve fitting and overfitting in science

The problem of fitting a curve to data is much discussed in the philosophy of science. One frequently cited discussion is by Forster and Sober (1994), where they aim to give a principled response to the overfitting problem, built around the Akaike information criterion. When Forster and Sober come to applying the lessons of their discussion to cases from the history of science, they mention only one example, Ptolemaic astronomy. The latter is a case of scientists adjusting a theory to fit anomalous data, rendering it more complex in the process. As Forster and Sober say, this is an instance of *adhocness*. Although their results gained from consideration of the Akaike information criterion may well illuminate the problem of adhocness, that is not the same as the problem of overfitting. Overfitting, as we have seen, is a matter of promoting a false theory by fitting it to data points some of which are *false*. Whereas the problem of adhocness involves promoting a false theory by fitting it to data points that are *true*. The problem with Ptolemaic astronomy was not that it sought to accommodate observational errors. Rather, it is a false explanatory theory that sought to avoid falsification from accurate data by making the underlying theory highly complex. The problem of adhocness does arise fairly frequently in the history of science. The problem of overfitting arises rather less often.

When is it that curve fitting (and the risk of overfitting) is a potential problem in science? It can arise when we know that those data points are generated by a combination of signal and noise. It is particularly problematic when we have little idea about what will generate the noise. In machine learning we may be confronted with a large data set and no model of the phenomenon generating the data nor any model of the noise. Often in science, however, we have a good idea about the noise that afflicts observation. Note first that perfect fitting of a

theory to data is sometimes literally impossible, however extreme a theory we are willing to contemplate. For if we are looking for a function $y = f(x)$, there can be no such function if our data include more than one distinct value of y for a given value of x. But that is the position scientists often find themselves in—they repeat the observation of the dependent variable, y, for the same value of the independent variable x (mutatis mutandis for multiple independent variables), and get multiple values. It is now of course meaningless to attempt to fit a curve through all the data points. An important advantage of the process of multiple repeat observations is that the difference between data points must be a result of noise, and so with enough observations we learn something about the noise. In astronomy it has long been understood that the data we collect will be normally distributed around the true location of, say, a planet, and with enough data we may be able to estimate the variance of that distribution. Understanding and accounting for noise can be an important part of the scientific process. The process of identifying gravitational waves, for example, involves detecting vibrations in an interferometer and then eliminating all possible sources of noise as explanations of the observed data.

Once we properly understand the source of noise and take it into consideration, we *must* accommodate our observations, even if they lead to a more complicated theory, as in the case of the orbits of Uranus and Mercury. Astronomers were sufficiently confident in their observations of the anomalous data points in the orbit of Uranus that it was simply not possible to regard them as the products of noise that could be ignored. Likewise, the anomalous perihelion of Mercury was not a matter of a few potentially noisy data points. It was a well-established phenomenon that must be accommodated by any true theory.

It is useful to make a distinction between *data* and *evidence* (Bird 2022: 163–4). Data are the records of our measurements and such like. If the thermometer in an experiment to measure the melting point of lead reads 327.7 °C then that is a data point—the actual temperature of the substance may be something different. However, if we collect and analyse enough such data points, we can come to know that the melting point is 327.5±0.2 °C. This then is *evidence*. A theory does not have to be consistent with any data point if it is not also evidence. But if a theory is inconsistent with the evidence then it is false. Evidence *must* be accommodated.

For another example consider the equation that describes the actual behaviour of a gas. The combined gas law states that $PV = kT$. We might prefer to retain this simple equation by ignoring data points that arise when pressure is high. But enough observations deliver evidence that there are cases, when pressure is high, where $PV \neq kT$. So we must accommodate such evidence, and doing so pushes us towards the more complex equation $(P + a(n^2/V^2))(V - nb) = nRT$ which is more accurate than the combined gas equation, accounting for effects of the van der Waals forces that are typically small but genuine. For that matter, the more sophisticated equation (van der Waals equation) is not the last word. For example, the 'a' in that equation represents a constant, but an improvement due to Soave makes a temperature dependent. There are further refinements yet. In this case the more

accurate (more truthful) equations are the more complex ones. The simple version, although a useful and revealing simplification, is false.

Of course it is true that there can be unexpected sources of noise and interference. In some areas we are not able to tell in advance what the sources of noise might be. Sometimes the area is such that we don't know what the sources of noise are. But in others they are very well understood. In philosophy of science, the problem of overfitting is presented as an instance of the problem of underdetermination of theory by data as if it is somehow a very general problem in science. But it is not a general problem, it is typically only a problem in some areas. In many areas we have theoretically very well grounded knowledge of what the sources of noise are such that we can design our experiments and methods of analysis to eliminate it.

Conclusion

Although I have discussed the application of the heuristics and overfitting approach to ethics at length, the significance of so doing is intended to be general. Precisely because the ethics case looked promising, the obstacles we encountered are particularly informative regarding the obstacles to implementing the heuristics and overfitting approach elsewhere.

The lessons learned are these. First, it is valuable to be aware that the curve fitting problem is not a general one in science. It arises in those cases where we have no information about the nature and extent of the noise, so that we do not know for any data point what the relative contributions of signal and noise may be. That does arise in some cases, but typically in science we do have information about the noise and we can take steps to quantify and account for it.

In philosophy we have tended to take the data (intuitive judgments) as perfectly reliable, and hence require that our theories are interpolations—they fit the data exactly. Williamson proposes that we drop this assumption. As we have seen, utilitarian ethicists had already followed this injunction for a century and a half, going back to Mill and Sidgwick. Consideration of the possibility of applying Williamson's approach to ethics, in a sense bringing Mill and Sidgwick up to date, shows that it is not straightforward to implement that programme. Generalizing from that case, we draw the following conclusions about what any implementation must do. Let T be hypothesized to be the correct theory; let H be the heuristic (or, often, heuristics) hypothesized to be the source of our intuitive judgments in many cases. Then in order to confirm this pair of hypotheses:

> (i) Heuristic H needs to be identified with some precision—just hypothesizing 'some heuristic' will not do.
> (ii) It needs to be shown that H does correctly predict our judgments in the relevant cases where those judgments diverge from T.

(iii) It needs to be shown that we use H *because* it is usually a good approximation to T—there must be an explanatory chain from H's approximate reliability to our use of H.

(iv) It needs to be shown that H does operate as a heuristic, that is to say, it delivers judgments because applying T is too difficult; if it is easy to apply T in a case, then it is unclear what is explained by the hypothesis that heuristic H is in play.

(v) It needs to be shown that the judgments in question really are produced by H, as opposed to unconscious (and sometimes fallible) reasoning with T itself.

Bibliography

André, M., L. Borgquist, and S. Mölstad (2003) Use of rules of thumb in the consultation in general practice—an act of balance between the individual and the general perspective, *Family Practice* 20: 514–519.

Bird, A. (2022). *Knowing Science*. Oxford University Press, Oxford.

Bird, A. (2024) Naturalized knowledge-first and the epistemology of groups. *Philosophy and Phenomenological Research*, 109: 856–73.

de Grefte, J. (2023). Knowledge as justified true belief. *Erkenntnis* 88, 531–49.

Forster, M. and E. Sober (1994) How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science* 45: 1–35.

Goldstein, D. G., and G. Gigerenzer, (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & ABC Research Group (Eds.), *Simple Heuristics that Make Us Smart*. Oxford: Oxford University Press, pp. 37–58.

Gould, S. J. (1991) *Bully for Brontosaurus: Reflections in natural history*. W. W. Norton.

Greene, J. (2002) *The Terrible, Horrible, No Good, Very Bad Truth About Morality, and What to Do About It* (Ph.D. thesis, Department of Philosophy, Princeton University).

Haidt, J., F. Bjorklund, and S. Murphy, (2000) Moral dumbfounding: When intuition finds no reason. Unpublished manuscript, University of Virginia. Available at: https://polpsy.ca/wp-content/uploads/2019/05/haidt.bjorklund.pdf

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108: 814–34.

Jackson, F. (1998) *From Metaphysics to Ethics*, Oxford: Oxford University Press.

Kaufman, D. (2007) Family resemblances, relationalism, and the meaning of 'art', *British Journal of Aesthetics* 47: 280–97.

Laskowski, N. and S. Finlay (2017) Conceptual analysis in metaethics, in T. McPherson, and D. Plunkett (eds) *The Routledge Handbook of Metaethics*, New York: Routledge.

Moore, G. E. (1903) *Principia Ethica*, Cambridge: Cambridge University Press.

Sidgwick, H. (1874) *Methods of Ethics*. London: Macmillan. (References to the edition by J. Bennett, 2017, https://www.earlymoderntexts.com/assets/pdfs/sidgwick1874.pdf.)

Singer, P. (2005) Ethics and intuitions, *The Journal of Ethics*, 9: 331–52.

Sinnott-Armstrong, W. (2006) Moral Intuitionism Meets Empirical Psychology, in T. Horgan and M. Timmons (eds.), *Metaethics after Moore*, Oxford: Oxford University Press.
Sunstein, C. R. (2005) Moral heuristics, *Behavioral and Brain Sciences*, 28: 531–42.
Tversky, A. and D. Kahneman, (1974) Judgment under uncertainty: heuristics and biases, *Science* 185: 1124–31.
Williams, B. (1973) A critique of utilitarianism, in J. J. C. Smart and B. Williams, *Utilitarianism: for and against*. Cambridge: Cambridge University Press, pp. 75–150.
Williams, B. (1994) Jim and the Indians, in P. Singer (ed.) *Ethics*. New York, NY: Oxford University Press, pp. 339–45.
Williamson, T. (2020) *Suppose and Tell*. Oxford: Oxford University Press.
Williamson, T. (2024) *Overfitting and Heuristics in Philosophy*. Oxford: Oxford University Press.