Chapter One of Timothy Williamson, *Overfitting and Heuristics in Philosophy* (OUP)
(draft of 23.4.2023)

Heuristics

1. *Counterexamples*

Counterexamples keep theorists honest. It is easy to regard counterexamples as the epistemological gold standard, as Karl Popper did. But just as there is fool's gold as well as genuine gold, so there are fool's counterexamples as well as genuine counterexamples. And just as all of us can be fooled if we trust our first impressions of apparent gold, so all of us can be fooled if we trust our first impressions of apparent counterexamples.

To check whether something is genuine gold, you can ask to have it tested at your nearest assay office. To check whether something is a genuine counterexample to a philosophical generalization, you can ask to have it tested at your nearest philosophy department, though somehow that sounds less reassuring. You may be disappointed to find that the philosophers' tests hardly go deeper than their first impressions.

Of course, if your local philosophy department is slightly old-fashioned, it may claim to possess a philosopher's stone, which turns base metals into pure gold. The likely mechanism is to reclassify the proffered counterexample as an *analytic* or *conceptual* truth, built into the use of the relevant terms. But that will be unsatisfying if the source of the case judgment in question is also the source of other case judgments inconsistent with it. They cannot *all* be pure gold.

In this book, I will argue that many alleged examples and counterexamples in philosophy are the products of *heuristics*, which can produce mutually inconsistent case judgments, so we are indeed in the envisaged predicament. The philosophical questions are not usually *about* the heuristics, and we needn't use the heuristics in *asking* the questions. But we do rely on the heuristics to generate the data on whose basis we *answer* the questions. The trouble is that we tend to rely on those outputs uncritically, treating them as data which our theories must fit. When some of the outputs are false, we are liable to dismiss true theories erroneously, as falsified by counterexamples. The apparent counterexamples may be all too convincing.

The situation is not all bad. If a heuristic produces mutually inconsistent outputs, *no* theory will be consistent with all of them together. False theories as well as true ones will appear falsified. Moreover, we may be able to identify what heuristics produced the outputs, and consequently to understand their strengths and weaknesses. That may enable us to handle our data in a more sophisticated and critical way, as other sciences have learnt to do. The occurrence of heuristic-induced errors is not a generic justification for scepticism. Our capacities for knowledge are hard to extricate from our propensities to error. The same cognitive systems enable us, in good cases, to learn how things are yet, in bad cases, make us misjudge how they are: no risk, no gain. This interdependence of strength and weakness is

crucial to the operation of the heuristics integral to so much cognition—human, animal, and artificial.

2. *What are heuristics?*

Roughly speaking, a heuristic is a rule of thumb for solving problems of some type. The application of the rule may be automatic or deliberate; it may be conscious, unconscious, or somewhere in between. Even if it involves conscious activity, one may or may not know what rule one is applying, and one may or may not think of it as a heuristic. Even on reflection, it may not be obvious to us when we are using a heuristic, still less what heuristic it is.

The function of a heuristic is to provide a way of solving problems of a given type that is fast, easy, efficient, and reliable enough to be useful. The way must be feasible in real time. It can be reliable enough without being *perfectly* reliable. Reliability here is equated with the probability that the way provides a *correct* solution, where the standard of correctness is built into the specification of the problem. For example, sniffing food to check whether it smells bad is a heuristic for determining whether it is still good to eat. Since food can go bad without smelling bad, it is not a fully reliable test, but it is quicker, more convenient, and less expensive than having the food tested in a laboratory. It is more reliable for some foods than for others.

Psychologists have studied many heuristics intensively. Sometimes they characterize heuristics negatively, as 'cheap and dirty', in the tradition of Daniel Kahneman (Kahneman, Slovic, and Tversky 1982), sometimes more positively, as 'fast and frugal', in the tradition of Gerd Gigerenzer (Gigerenzer, Hertwig, and Pachur 2011). At worst, heuristic-based cognition is regarded as a form of *irrationality*, at best, as a form of *bounded rationality*. Presumably, some heuristics are better than others, at least for a given purpose under given conditions. We might be better off avoiding *some* heuristics, but the nature of human cognition—perhaps of finite cognition in general—precludes our avoiding them *all*.

Heuristics, as understood here, can be culturally acquired, or even idiosyncratic. For example, medical experts—communally or individually—develop heuristics for interpreting X-rays. But many important heuristics are virtually universal to humans. For example, visual illusions are probably by-products of such heuristics built into the visual systems of humans and other animals (Fleming 2012, Gigerenzer 2021). The heuristics responsible for such illusions are topics for psychological investigation. When heuristics are virtually universal, they may be innately hardwired, or at least the natural outcome of innate domain-general principles and learning mechanisms. Either way, evolutionary adaptiveness will often play a large role in explaining how we have come to use such heuristics. Still, in principle, checking on Google could become a culturally transmitted virtually universal heuristic, whether or not it is evolutionarily adaptive.

One heuristic which often involves conscious thought is *take-the-best* (Gigerenzer and Goldstein 1996). It is a way to choose between two alternatives for some purpose, given various epistemic cues ranked by 'validity' (how well they indicate optimality for that

purpose). Take-the-best tells you simply to follow the highest-ranked cue that discriminates between the alternatives—as opposed, for instance, to somehow constructing and comparing weighted averages over all the cues. Thus, one might simply decide to shop at the nearest supermarket, without having taken into account price, range, or quality of goods. Of course, even when one consciously applies the heuristic, one rarely thinks of oneself explicitly *as* applying take-the-best.

Often, there is a slower but more accurate alternative to using a given heuristic. For instance, our visual systems routinely treat colour contours as a guide to the shapes of three-dimensional material things. Camouflage succeeds in misleading observers about those shapes by exploiting their reliance on that heuristic. In principle, we can correct such mistakes, for example by using our sense of touch, though that alternative may be unfeasible in the circumstances, as in time of war. Still, heuristics are in principle, and often in practice, *defeasible*.

Sometimes no more reliable alternative is available. With take-the-best, one might expect to do better when time permits by consciously 'weighing up all the pros and cons'. But that may be over-optimistic. One may have only the faintest idea of how to individuate the relevant considerations, what relative weights to assign them, and how to measure performance on one dimension against performance on another. When I try to take a decision by weighing up all the pros and cons, the result is only to make me vividly aware how open the process is to manipulation in favour of whichever alternative I independently prefer. Indeed, experimental studies suggest that take-the-best is surprisingly reliable, compared to more elaborate methods available to the subjects at the time, where the correct answer is known to the experimenter by some method unavailable to the subjects at the time (Gigerenzer and Goldstein 1996). When many complex ramifications of different kinds really must be taken into account in making a difficult decision, my preferred method is to procrastinate until one morning I wake up knowing what I'm going to do. Conscious reflection passes the buck to unconscious processes, which may do a better job of integrating information from many sources (on the limits of reflection see Kornblith 2012). In retrospect, that method has served me fairly well. Many other people seem to do likewise.

When we rely on a heuristic without thinking of it as such, and with no conception of a more reliable way of solving the problem, we may mistakenly regard the heuristic's output as *indefeasible*. For lack of an alternative category to put it in, a philosopher may even call it an 'intuition', an 'analytic truth', a 'conceptual connection', or whatever. That illustrates the poverty of the philosophically current taxonomy, and is all the more reason to make room for the category of heuristics in philosophers' working vocabulary.

In discussing heuristics, I have not specified whether being a heuristic entails being less than perfectly reliable, or being above some moderate level of reliability, or whatever. More generally, I will not stipulate a precise definition for the word 'heuristic'. No such definition is needed for present purposes, and at this early stage of inquiry picking one might even be harmful, by cutting apart from a cognitive joint. We have a range of more or less paradigm cases of heuristics, as already indicated, and by classifying something as a heuristic we draw attention to its similarities to such cases. For present purposes, that is what matters.

Just as heuristics built into the human visual system produce visual illusions in special circumstances, so heuristics built into the human cognitive system may more generally have

the capacity to produce philosophical *paradoxes*, which can be properly diagnosed only once we identify the heuristics at work. Such heuristics may be very general, but even much more specific heuristics may play a role in generating philosophical paradoxes: for example, heuristics for attributing beliefs to people on the basis of what they say, and heuristics for individuating physical objects on the basis of visual perception.

Naturally, postulating a new heuristic does not come for free. For the postulate to be initially plausible, the candidate heuristic should be simple, quick, efficient, and useful. In particular, the problem it solves should crop up often enough to make a solution dedicated to that problem worth our storing it up for future use. Postulating a heuristic is especially plausible when it would be strange if we *didn't* use something like that heuristic.

Philosophers may be tempted to refine a postulated heuristic by adding exception-clauses, restrictions, and qualifications, to rule out counter-instances and so enhance its reliability. One should resist that temptation, for the 'refined' heuristic is likely to be psychologically unrealistic, since it increases computational times and costs of application, typically for a comparatively small gain in reliability, and perhaps even a loss in generality. Those increases will be drastic if they require conscious reflection, which is very slow by neural standards, and liable to create a bottleneck in processing. In the midst of action, a prompt, moderately reliable answer usually does better than a very reliable answer when it is too late, or than no answer at all. When over-reflective creatures pause to reflect, they risk being eaten, or at least beaten to scarce resources, by their less reflective predators or competitors. Even in modern life, indecision can lead to disaster.

In general, what heuristic we use, if any, under given circumstances is a psychological question, open to experimental test. Evolution does not guarantee that our actual heuristics will be the optimally efficient ones. In this chapter, however, the concern will not be with such experimental work, though the need for it in the long run is obvious. The aim here is to clarify our initial theoretical understanding of the potential relevance of specific heuristics to philosophy, rather than to engage 'blind' with the psychological literature. We need to develop theoretical hypotheses properly before we test them, to know what we are looking for.

In the next four sections, I will explain and discuss some plausible candidates for heuristics on which we may be relying, knowingly or unknowingly, when we wrestle with some philosophical problems. In such cases, we risk getting suckered by our own heuristics.

3. *The persistence heuristic*

Here is a short vignette:

Mary was in London when a man wolf-whistled at her. She took a step towards the man, then slapped him.

To check whether a subject has properly understood the vignette, a psychologist might ask this comprehension question:

Where was Mary when she slapped the man?

A natural answer, which the psychologist would presumably accept, is:

She was in London when she slapped him.

However, the vignette only specifies that Mary was in London *when he wolf-whistled at her*. It adds that she took a step towards him before slapping him. Thus, the natural answer in effect assumes that if someone is in London, and takes a step, then they are still in London. But that assumption is not universally correct, for people occasionally walk out of London. In comprehending the vignette, one automatically updates the initial information 'Mary was in London' to the slightly later time when she slapped him, because the change involved in taking a step forward is treated as 'too small to matter'. That treatment is the default, but it is defeasible: if you had previously been told that Mary lived right on the edge of London, or that she had seven-league boots, you might have been wary about updating her supposed location in that way.

The example illustrates a very general cognitive tendency. For instance: you learn today from a trustworthy source that Emomali Rahmon is the President of Tajikistan. Tomorrow, someone asks you 'Who is the President of Tajikistan?' It would be natural for you to answer (complacently): 'Emomali Rahmon'. To answer 'Well, Emomali Rahmon was the President yesterday' would be unnatural and pedantic, even though you know that presidents can die or resign in a day; no president is forever. One day is treated as too small a change to matter. Of course, we have some sense of such information having a use-by date; if you are asked twenty years from now 'Who is the President of Tajikistan?', having heard nothing about Tajik politics in the meanwhile, you may answer 'It used to be Emomali Rahmon'.

When we update information in present-tense form, we often do so by *retaining* the present tense, even though such *present-tense updating* involves going beyond our original information. Much of what we describe as factual 'memory' is the result of present-tense updating ('Do you remember who is the President of Tajikistan?'). By contrast, *past-tense updating* sticks closer to the original information, by putting it in past-tense form, with reference to the time when it was strictly expressed in present-tense form ('Emomali Rahmon was President of Tajikistan on 15th October 2022' or 'The last I heard, Emomali Rahmon was President of Tajikistan'), as we might do when we regard change as plausibly imminent. Past-tense updating is more appropriate for episodic memory of particular events. If one cannot date the event, one may simply use a memory demonstrative such as 'then' or 'that time we were in Barcelona' or 'when I was pick-pocketed'.

Although present-tense updating is not always truth-preserving, it is *usually* truth-preserving. Almost every step that starts in London ends in London; almost every president of a country yesterday is its president today, and so on. Moreover, there is no feasible alternative to present-tense updating, however much sceptics may complain about its fallibility. No one can be constantly rechecking everything. Indeed, even computer data bases use present-tense updating perforce. Once someone's address has been entered into a data base, it cannot be checked every day, let alone every second, to test whether it is still their

current address. Present-tense updating does not reflect some peculiarity of the human brain, but instead far more general features of the problem of information-gathering and retention. Artificial intelligence will have to do present-tense updating, just as natural intelligence does. For example, much of the data on which an AI system was trained up will sooner or later go out of date.

One advantage of present-tense updating over past-tense updating is that the questions to which the former gives direct answers tend to be of more practical significance than the questions to which the latter gives direct answers. For instance, if you want to get food and drink, it is usually more helpful to know where food and drink are *now* than to know where they were *yesterday*. Creatures without episodic memory, as some non-human animals are alleged to be, may well be unable to do past-tense updating; for many of their purposes, present-tense updating will suffice. Even for humans, although we can sometimes make inferences from the outputs of past-tense updating to the information we need for decision-making—from where food and drink were yesterday to where they are now—conscious inference is psychologically costly. In the heat of action, it is more useful to have the required information already available directly—at one's fingertips—than to spend time and attention inferring it. That consideration favours present-tense updating.

The underlying heuristic is more general than the phrase 'present-tense updating' may suggest. The heuristic provides much of our understanding of physical things as persisting through change over time. Seeing a tree, I think 'This tree is here', using 'this tree' and 'there' as perceptual demonstratives. The next day, somewhere else, I remember the tree as so located, thinking 'That tree is there'—not just 'That tree *was* there'—using 'that tree' and 'there' as memory demonstratives anaphorically linked respectively to the original perception, even if I am sure that it lost some leaves over the intervening windy day. I unreflectively treat such changes as too small to matter to the tree's identity. The same underlying principle applies modally as well as temporally, to variation across counterfactual possibilities as well as to variation across times: just as we allow that this ship will soon have another plank in place of this rotten one, we allow that it *could have been originally made* with another plank instead of this one with which it was originally made: a difference of one plank is too small to matter.

The underlying heuristic can be summed up in the generic slogan 'Small changes don't matter'. We may call it the *persistence heuristic*. It plays a major if largely passive role in solving the problem of adapting what we know or believe to new situations as efficiently as possible.

In the slogan 'Small changes don't matter', 'changes' should be understood loosely, even metaphorically. In particular, for present purposes, zero change counts as the smallest change. By the heuristic, things persist when they remain unchanged. Furthermore, the difference from one possibility to a counterfactual alternative, or from one object to a similar object, also counts as a change for these purposes, as will be illustrated below.

Examples of the persistence heuristic and its inhibitors are easily multiplied. Normally, one need not keep rechecking someone's scalp to retain knowledge that they are not bald, even though they lose a few hairs every day. But if you tell me that John, though not yet bald, is rapidly going bald, I may keep glancing at his scalp. If you have borrowed a book, you need not keep asking yourself whether you still have that book every time you dislodge a

few molecules off a page with your fingers. But if the book is a priceless, crumbling medieval manuscript, you may worry more about its survival. ‘I wish this table had been made slightly longer’ is much less likely than ‘I wish this table had been ten times longer’ to prompt the default-breaking thought ‘Would that still have been this table?’ The persistence heuristic explains such patterns, obviating the need to postulate more elaborate forms of proto-metaphysical thinking.

Of course, experience and testimony can modify our sense of what counts as a small change for a specific kind of object, and so raise or lower the threshold for inhibiting the persistence heuristic. But tweaks in how we implement the heuristic do not replace it by something else.

We also use the persistence heuristic to transfer information about one thing to another. I pick an apple from a tree and bite it. The apple tastes sour. I expect it to taste sour at the next bite too, and I expect another similar-looking apple from the same tree to taste sour too. With respect to taste, the difference between the two apples is treated as too small to matter. That is a primitive form of induction.

We use the persistence heuristic *offline* as well as *online*. We use it online when we update on new evidence, perhaps received from sense perception or from testimony. We use the heuristic offline when we adapt what we know or believe to a hypothetical supposition. For example, in deciding whether to eat that other similar-looking apple, I suppose ‘I eat that apple’, and develop its consequences in imagination; as a result, I may decide *not* to eat that apple. You may have been carrying out such offline processing, using your imagination, when reading this chapter, as you considered the various hypothetical cases presented above.

The persistence heuristic is a crucial labour-saving device. Without it, cognition would be continually restarting from scratch. That would be hopelessly inefficient. The heuristic’s utility is manifest. As already emphasized, it is defeasible. Persistence is only the default, and we can often identify its failures. When a large change is in the offing, or we know or strongly suspect that a boundary is nearby, the operation of the heuristic is inhibited. But normally we need not actively exclude such defeating conditions, for that would undermine the heuristic’s utility, which is exactly to avoid such testing. We rely on persistence unless something sets off a mental alarm.

One corollary of the persistence heuristic’s inhibiting conditions is that the heuristic is more easily inhibited for precise terms than for vague ones. For a precise term, we are more clearly aware of its boundaries, and where they lie. Our awareness of their proximity sounds an alarm; the heuristic’s operation is inhibited. By contrast, for a vague term, we have no such clear awareness of its boundaries, and usually no alarm is sounded; the heuristic’s operation is not inhibited—though we may feel growing unease as we slide down a slippery slope. But the heuristic itself is applicable equally to precise and vague terms. For example, in the vignette about Mary and the wolf-whistler, the heuristic delivers the verdict that she is still in London after taking a step, irrespective of whether one envisages the boundaries associated with the name ‘London’ as vaguely or precisely defined. When one reads the vignette, that question does not naturally arise. Checking whether the terms in play are vague or precise is no part of the persistence heuristic: such checking would use up valuable time and energy for no commensurate benefit. The heuristic itself applies equally in vague and

precise cases, but is more liable to be psychologically defeated in the latter than in the former because the boundary is psychologically salient.

In cases of vagueness, the shortage of defeaters for the persistence heuristic makes it prone to *sorites paradoxes*, since it can be applied iteratively—which rarely happens under normal conditions. Many small differences add up to a large difference. Correspondingly, the heuristic validates *tolerance principles* such as 'If *n* grains make a heap, *n*–1 grains make a heap' for arbitrary '*n*' or 'If *x* looks red and *y* is visually indiscriminable from *x* then *y* looks red too'. One assesses the principle by supposing the antecedent '*n* grains make a heap' or '*x* looks red and *y* is visually indiscriminable from *x*' and applying the heuristic under that supposition to verify the consequent '*n*–1 grains make a heap' or '*y* looks red'. Informally, one imagines a heap, imagines one grain being removed, or something looking red, and something else where one can see no difference in colour, and uses the heuristic offline in the imagination to confirm that what remains is still a heap or that the second thing looks red too. There is no psychologically salient boundary for 'heap' or 'looks red' to inhibit the heuristic's operation. By default, the tolerance principle is accepted. Notoriously, it suffices to generate the sorites paradox, which drives one from an obviously true starting-point such as 'Ten thousand grains make a heap' to an obviously false conclusion such as 'One grain makes a heap', or from 'This looks red' said of a prototype of red to 'This looks red' said of a prototype of yellow. The tolerance principle only needs to fail at one step out of many in the sorites series for the sorites argument to be unsound. Our instinctive reliance on the highly but not perfectly reliable persistence heuristic helps explain why we are cognitively vulnerable to paradoxes of this form, why we find them so hard to resist.[1]

Some philosophers have got the impression that tolerance principles for vague expressions are somehow 'analytic' or 'semantic', or that they are 'conceptual connections' built into the corresponding concepts, thereby rendering those concepts defective.[2] That is a misunderstanding of the principles' status, perhaps resulting from the absence of 'heuristic' from the traditional philosopher's impoverished menu of options. Tolerance principles for vague expressions are no more 'analytic' than are the analogous tolerance principles for precise expressions; they are all applications of the same heuristic. The difference is just that some of them are psychologically more easily inhibited than others. Since our susceptibility to sorites paradoxes simply results from our reliance on the persistence heuristic in epistemically non-ideal conditions, it motivates no revision of classical logic or bivalent semantics. Much of the literature on vagueness exhibits one of the harms done by the 'linguistic turn': the tendency to seek linguistic solutions for epistemic problems.

4. *The supppositional heuristic for conditionals*

The persistence heuristic is general-purpose. For contrast, we now consider a heuristic primarily for the assessment of conditionals, expressed by sentences of forms such as 'If A, C', although it can also be applied to the assessment of generic generalizations, as explained below (see Williamson 2020, henceforth 'S&T', for a book-length discussion of the heuristic). Arguably, it is humans' primary way of assessing conditionals, though not our only one. It is not a new discovery: for example, it is closely related to the Ramsey Test,

originally described by Frank Ramsey, which uses a form of hypothetical updating. But its role has been misunderstood, because its heuristic status went unrecognized.

Here is Ramsey's concise description, in a footnote (1929: 143, with change of lettering):

If two people are arguing 'If A will C?' and are both in doubt as to A, they are adding A hypothetically to their stock of knowledge and arguing on that basis about C

A simple, schematic version of the suppositional heuristic is this:

Assess 'If A, C' outright as you assess 'C' on the supposition 'A'.

We can see how this works with some examples. Mary has bought a ticket in a lottery. The prize is a million pounds. Here are three conditionals about it:

(1) If Mary's ticket wins, she will get lots of money.

(2) If Mary's ticket wins, it will lose.

(3) If Mary's ticket wins, she will buy a new house.

To assess (1)-(3), we first suppose their shared antecedent, 'Mary's ticket wins', and then assess their consequents on that supposition.

Since the prize is lots of money, we accept (1)'s consequent 'She will get lots of money' on the supposition of (1)'s antecedent 'Mary's ticket wins'. Using the suppositional heuristic, we therefore accept (1) outright.

Since Mary's ticket winning is inconsistent with its losing, we reject (2)'s consequent 'It will lose' on the supposition of (2)'s antecedent 'Mary's ticket wins'. Using the suppositional heuristic, we therefore reject (2) outright.

Since we have no idea of Mary's priorities, we suspend judgment on (3)'s consequent 'She will buy a new house' on the supposition of (3)'s antecedent 'Mary's ticket wins'. Using the suppositional heuristic, we therefore suspend outright judgment on (3).

These predictions fit natural reactions to (1)-(3). Similarly, as we learn more about Mary's priorities, her buying a new house will look more or less likely conditional on her ticket's winning, and (3) will come to seem correspondingly more or less likely outright. There is extensive evidence that speakers' assessments tend to conform to the suppositional heuristic (Evans and Over 2004, Douven 2016, S&T).

Often, we need to assess conditionals not outright but on a further set of background suppositions, Γ. Strictly speaking, that was already happening with our assessments of (1)-(3), since 'Mary has bought a ticket in a lottery' and 'The prize is a million pounds' really played the role of background suppositions; we did not believe them outright. For these purposes, we need a more general version of the suppositional heuristic:

Assess 'If A, C' on the suppositions Γ as you assess 'C' on the suppositions Γ∪{'A'}.

The original, simpler version corresponds to the special case where Γ is the empty set. In more complex reasoning, we often find ourselves making suppositions within suppositions. For example, when we are devising a strategy with multiple choice-points as we confront different contingencies at different stages, we need to consider a tree of branching possibilities. In constructing or following a tricky mathematical proof, one typically has to make hypotheses in the scope of hypotheses already made. Without the generalized suppositional hypothesis, one would be stymied in one's natural attempts to assess conditionals in such situations, but that does not happen. In effect, in the outright version of the heuristic, the final verdict on the conditional is online, whereas the generalized version extends the heuristic to offline cases too.

How does such hypothetical thinking help us? Many of our dispositions to form expectations have been calibrated by experience, our own or our ancestors', and so encode information about the world so experienced. We may need to apply such information to a prospective new situation, in advance of encountering it. Is it a danger to be avoided or an opportunity to be sought? How can we prepare ourselves to encounter it? We imaginatively suppose that the situation obtains, and use our expectation-forming dispositions 'offline' to assess what it may be like, and what it may lead to. We can then store such information in the convenient form of a declarative sentence, as a conditional: 'If the situation obtains, such-and-such will happen'. Such reality-oriented cognitive uses of the imagination are plausibly central to its evolutionary function (Williamson 2016e). In short, the suppositional heuristic enables us to use connections implicit in our cognitive system to make them explicit in a conditional.

One advantage of suppositional thinking is that it is often feasible when truth-functional thinking is not, because we cannot assess the antecedent or consequent separately. I may know that *if* John drops the vase, it will smash, even though I have no idea how likely he is to drop the vase, and so no idea how likely it is to survive. This is an epistemological point, not a semantic one. It does not show that 'if' is not truth-functional. After all, we may verify the truth-functional disjunction 'Either he will not drop the vase or it will smash' or falsify the truth-functional conjunction 'He will drop the vase and it will not smash' by supposing 'He drops the vase' and on that basis verifying 'It will smash'. Just as we can verify a disjunction without verifying either disjunct, and we can falsify a conjunction without falsifying either conjunct, we can verify a conditional without either falsifying its antecedent or verifying its consequent. But conditionals *invite* hypothetical thinking in a way that disjunctions and conjunctions do not; conditionals as it were *ask* to be so assessed. To put it another way, hypothetical thinking feels like a *direct* way of assessing a conditional, but an *indirect* way of assessing a conjunction or disjunction. That difference manifests the suppositional heuristic's naturalness for conditionals.

The suppositional heuristic can also be applied to generic generalizations, such as 'Tigers are striped', which is not refuted by an occasional albino tiger. For 'Ns are F' can be paraphrased as 'If it's an N, it's F' ('If it's a tiger, it's striped'), where 'it' is treated as if it referred to an arbitrarily chosen item. One assesses 'It's striped' on the supposition 'It's a tiger', which gives the appropriate result. Even when the generic is not expressed in

conditional form, the suppositional heuristic is still applicable (S&T: 142-6). Much of humans' general knowledge is most naturally expressed in such generics.

Of course, many of our general biases and prejudices are also most naturally expressed in generics. But that is not the suppositional heuristic's fault, for it prompts one to accept 'Ns are F' only if one *already* has the bias or prejudice, disposing one to accept 'It's F' on the supposition 'It's an N'. What the heuristic does is to enable one to make one's implicit bias or prejudice explicit in a conditional or a generic generalization. The heuristic can hardly be expected to do *better* than the underlying cognitive dispositions—its role is to use them, not to filter the good ones from the bad. Although well-intentioned proposals have occasionally been made to ban the utterance of generics, the likely effect of such a ban would be to force the biases and prejudices underground, while doing the same to most of ordinary humans' general knowledge of the natural and social world, very little of which consists in exceptionless universal generalizations.

Despite all its virtues and benefits, the suppositional heuristic is *inconsistent*, both in itself and with uncontroversial background knowledge. This can be shown in various ways.

One route to inconsistency goes via graded attitudes. Let Prob(X | Y) be the probability (in any relevant sense) of X conditional on Y, B be the conjunction of the background suppositions, and A * C formalize 'If A, C'. Applying the generalized heuristic to assignments of probability results in the equation Prob(A * C | B) = Prob(C | A ∧ B). This is the generalized version of the identification of the probability of a conditional with the corresponding conditional probability proposed by various authors (Jeffrey 1964, Ellis 1969, Stalnaker 1970). It feels very natural, thanks to the suppositional heuristic, but a version of an argument originally devised by David Lewis shows the equation to imply that no three mutually exclusive possibilities have nonzero probability (Lewis 1976, S&T: 42-3). That is an absurdly restrictive constraint: when a die is thrown, there are six mutually exclusive outcomes, each with probability 1/6. Attempts to find a loophole in Lewis's argument all founder when applied to the corresponding argument for the generalized suppositional heuristic; it is simply a mathematical result.

Much ingenuity has been spent on finding subtle restrictions or complications of the equation to get around Lewis's result. For a heuristic, that is exactly the wrong reaction. The heuristic's utility depends on its unrestricted simplicity. No subtle restrictions or complications are baked in.

Another proof of the heuristic's inconsistency does not even require the assumption of three mutually exclusive possibilities. It is worth sketching to give an idea of what is going on (S&T: 37-42 presents the proof in more detail).

First, we apply the generalized heuristic to assessments of *deductive entailment*. This is like the special case of the probabilistic equation for probability 1, the principle that Prob(A * C | B) = 1 if and only if Prob(C | B ∧ A) = 1, but without the mathematical complications that arise for probabilities conditional on a hypothesis whose probability is 0 (when the standard ratio definition of the conditional probability, Prob(X | Y) as Prob(X)/Prob(X ∧ Y), involves division by 0). The result can be formalized as the equivalence of $\Gamma \vdash A * C$ with $\Gamma \cup \{A\} \vdash C$, where ⊢ is interpreted as deductive entailment. That equivalence amounts to the combined rules for a standard conditional in a standard system of natural deduction: the implication from $\Gamma \vdash A * C$ to $\Gamma \cup \{A\} \vdash C$ is in effect

modus ponens (the conditional elimination rule), while the implication from $\Gamma \cup \{A\} \vdash C$ to $\Gamma \vdash A * C$ is just conditional proof (the conditional introduction rule). These rules can be shown to make * equivalent to the material (truth-functional) conditional. So far so good, at least for friends of the material reading of 'if'.

The trouble is that we can also apply the generalized heuristic to assessments of *deductive incompatibility*. This is like the special case of the probabilistic equation for probability 0, the principle that $\text{Prob}(A * C \mid B) = 0$ if and only if $\text{Prob}(C \mid A \wedge B) = 0$, but again without the complications arising for probabilities conditional on a hypothesis of probability 0. The result can be formalized as the equivalence of $\Gamma \vdash^{\neg} A * C$ with $\Gamma \cup \{A\} \vdash^{\neg} C$, where $\vdash^{\neg}$ is interpreted as deductive incompatibility. Since being deductively incompatible with something is equivalent to deductively entailing its negation, in effect $\Gamma \vdash \neg(A * C)$ is equivalent to $\Gamma \cup \{A\} \vdash \neg C$. That can be shown to make $\neg(A * C)$ equivalent to the negated conjunction $\neg(A \wedge C)$, which in turn makes * equivalent to *conjunction*. But * cannot be simultaneously equivalent to *both* the material conditional *and* conjunction, since any material conditional with a false antecedent is true, whereas any conjunction with a false conjunct is false. In brief, two legitimate special cases of the heuristic force mutually incompatible readings on natural language conditionals.

Human reliance on the inconsistent suppositional heuristic in assessing conditionals helps explain why their semantics has puzzled logicians for over two millennia, on and off. The issue was so controversial in Alexandria during the third century BCE that the poet Callimachus wrote 'Even the crows on the roof-tops are cawing about which conditionals are true' (Mates 1949: 234). Although some applications of the heuristic require the material reading, using the heuristic we reject (2) above ('If Mary's ticket wins, it will lose'), even though it is almost certainly true on the material reading, since its antecedent is almost certainly false. More generally, when A is highly improbable or C highly probable, and therefore the material conditional $A \rightarrow C$ is also highly probable, C can still be highly improbable conditional on A, so by applying the suppositional heuristic one judges 'If A, C' highly improbable. Since the heuristic is inconsistent, it will generate apparent counterexamples to *any* proposed interpretation of a natural language conditional.

How can the suppositional heuristic be useful, given its inconsistency? How has it survived the pressures of evolution? The answer is much less straightforward than for the persistence heuristic.

An illuminating case to start with is the practice of mathematical proof. Mathematicians write their proofs in a framework of natural language, afforced with lots of mathematical notation and diagrams, not in some formal language—as one can see by glancing at the pages of mathematical journals. In particular, mathematicians reason with natural language conditionals such as 'if'; they receive no special training in how to use them mathematically, no special explanations or warnings. Nevertheless, to a good approximation, their reasoning with 'if' fits standard natural deduction rules for the material conditional—modus ponens and conditional proof—just as in the special case of the heuristic for deductive entailment above. That is why, as often noted, 'if' can be seamlessly read in mathematical texts as a material conditional. Still, since mathematics seems to press our deductive capacity to the utmost, why does the inconsistency between applying the heuristic to deductive

entailment and applying it to deductive incompatibility never surface in mathematics? For example, let A be an implicitly inconsistent mathematical hypothesis. Since A deductively entails any mathematical conclusion C, one can use the heuristic to establish 'If A, C' outright. Since C is also deductively incompatible with A, one can also use the heuristic to refute 'If A, C' outright. That would make mathematics itself inconsistent. Obviously, no such paradox arises in mathematical practice. The reason is that refutability is simply treated as provability of the negation, and in mathematical practice there is no need to apply negation directly to a conditional. A refutation of C from A is recorded as a proof of 'If A, not C'; the status of 'Not(if A, C)' is not directly addressed. Negated universally quantified conditionals are more common, but the suppositional test does not apply to them directly, since their overall form is different.

The preference for negating the consequent over negating the conditional is observable outside mathematics too. 'If he drops the vase, it will not break' is much clearer and more natural than 'It is not the case that if he drops the vase, it will break'. The latter feels clunky: not ungrammatical, but like a flatfooted response to someone who has just asserted 'If he drops the vase, it will break'. Prefixing 'not' to the conditional sentence sounds ill-formed, and is not a way of negating it. Although you could respond to 'If he drops the vase, it will break' with 'Not if he drops it into the pool', that is naturally heard as elliptical for 'If he drops the vase into the pool, it will not break'. Psychologically, 'If A, C' invites assessment by the suppositional heuristic, whereas we find it much less obvious how to assess 'It is not the case that if A, C'.

Underlying these linguistic effects may be a more general pattern in thought: to register rejection of 'A' by accepting 'Not A', replacing a negative attitude by a positive attitude to the negation. 'Not A' may then in turn be fleshed out in more positive terms (on the psychology of negation see Kaup, Zwaan, and Lüdtke 2007). If the default attitude to a sentence in inner speech is acceptance, this would tend to ease the burden of processing, by obviating the need to apply some controlling mechanism to inhibit the default. Such a cognitive tendency would be efficient for both outright attitudes and attitudes under suppositions. In the latter case, it would set one up then to apply the suppositional heuristic to the positive attitude, for which it gives better results. If there is a general human cognitive tendency along such lines, it would help explain why the heuristic's inconsistency causes so little trouble in practice, both in mathematics and elsewhere, in the absence of any special training. Although it would not strictly resolve the inconsistencies lurking in the heuristic, it would tend to minimize their effects.

The effect of the suppositional heuristic is also modified by the generic practice of accepting conditionals preserved by memory or communicated by testimony, without reapplying the suppositional test in the new epistemic context. For example, when I assess the opposite conditionals 'If A, C' and 'If A, not C' by the suppositional heuristic, I do not accept both, because I do not accept both the contradictories 'C' and 'Not C' on the supposition 'A' (when 'A' is consistent). But sometimes I may rationally accept 'If A, C' from one trustworthy source while also accepting 'If A, not C' from another trustworthy source; I then conclude 'Not A'. Perhaps each trustworthy source has direct access to information to which neither I nor the other trustworthy source has direct access, and both trustworthy sources used the suppositional heuristic (S&T: 89-102 discusses such cases in detail).

Once one takes into account the overall practice of using conditionals to encode and transfer information, one can argue that the information stably associated with a conditional is simply that of the material reading, outside mathematics as well as inside.

The point is not obvious, for the suppositional heuristic often grossly underestimates the probability of a conditional on its material reading. For example, the heuristic assigns probability zero to the conditional (1) above, 'If Mary's ticket wins, it will lose', since the consequent is inconsistent with the antecedent and so has probability zero conditional on the latter. That fits the strong unreflective impression that the conditional is idiotic, and the strong unreflective inclination, when asked 'What is the chance that if Mary's ticket wins, it will lose?', to answer 'None'. But the material reading makes the conditional almost certainly true, since its antecedent is almost certainly false, and a material conditional with a false antecedent is true. In isolation, such cases look like decisive counterexamples to the material reading of 'if'. But that attitude is no longer adequate once one realizes that the unreflective judgments are the outputs of an inconsistent heuristic. In those circumstances, we cannot rely on the standard methodology of requiring a semantics for the conditional to vindicate all normal patterns of speakers' unreflective judgments.

We may have to be content with a less direct connection between semantics and heuristics. For example, when we treat the conditional probability Prob(C | A) as an estimate of the probability of the conditional on its material reading, Prob(A → C), it is often too low, but never too high: in that sense, the heuristic may make us trust too little, but will not make us trust too much. More demanding truth-conditions for the conditional lose that advantage, by sometimes making the heuristic overestimate its probability; less demanding truth-conditions make the conditional unnecessarily uninformative, given the heuristic. Thus the material truth-conditions make conditionals as informative as they can be, compatibly with preventing the heuristic from overestimating their probability. Such a useful connection between the heuristic and the truth-conditions provides further confirmation of the overall picture (S&T: 103-10).

Being too cautious with conditionals may be less costly than not being cautious enough. After all, on the present view, the point of conditionals is not to provide access to a special kind of information but rather to provide a special kind of access to information. For example, on the material reading, 'If Mary's ticket wins, it will lose' has the same truth-condition as 'Mary's ticket will either lose or not win'; although we cannot access the high probability of that condition's obtaining via the suppositional heuristic, we can access it via the known high probability of Mary's ticket losing. As already noted, suppositional thinking comes into its own with conditionals like 'If the vase is dropped, it will break'. Even though it has the same truth condition as 'The vase will either break or not be dropped', we may be unable to access the high probability of the condition's obtaining via the separate probabilities of the disjuncts, because we have no idea how to estimate the latter probabilities. Instead, we can apply the suppositional heuristic, since we can access the high probability of the vase's breaking conditional on its being dropped, through an imaginative exercise constrained by our background knowledge. The suppositional heuristic's limitations are a small price to pay for its distinctive benefits.

5. *Disquotation and heuristics for belief ascription*

Here is an elementary speech exchange between two children:

John: I'm taller than you.
Janet: That's not true! I'm taller than you.

We might articulate Janet's underlying thought process as an inner monologue like this:

Janet: John said 'I'm taller than you'. He said that he's taller than me. But I'm taller than him, so he's not taller than me. So what he said is not true.

In passing from the internal direct speech report 'John said "I'm taller than you"' to the internal indirect speech report 'He said that he's taller than me', Janet unreflectively replaces John's pronouns 'I' (first-person) and 'you' (second-person) by her 'he' (third-person) and 'me' (first-person); she also replaces the name 'John' by 'he'. In passing from the thought 'I'm taller than him' to the speech addressed to John, 'I'm taller than you', she unreflectively replaces the third-person pronoun 'him' by the second-person pronoun 'you'. All these effortless replacements are to preserve reference and conversational appropriateness—though Janet's use in inner speech of the third-person rather than the second-person in referring to John suggests that she is keeping her psychological distance from him. By contrast, the words 'taller than' are preserved verbatim from the direct speech report to the indirect speech report, as is the present tense of the verb from 'I'm' (= 'I am') to 'he's' (= 'he is') rather than 'he was', in effect a case of the persistence heuristic, since the speech reports themselves are past tense ('said', not 'says').

In arriving at the indirect speech report, Janet's default is to repeat John's words (homophonic disquotation), while fluently adjusting to the context-sensitivity of pronouns. Reasonably enough, she does not even consider the possibility that John means something different by 'taller' from what she means. Counterfactually, if John had a notorious habit of using words as if they meant the opposite of what they in fact do, her knowledge of his bad habit might have inhibited the default's operation, and she might have reacted differently. Homophonic disquotation is the standard *heuristic* for indirect speech reports, but it is modified more or less automatically in familiar cases of context-sensitivity, and it can also be modified more reflectively in light of special circumstances.

We often need the indirect speech report in order to assess others' statements. For instance, Janet obviously cannot just assess the sentence type 'I'm taller than you', since she addresses that very sentence back to John in rejecting his use of it. Rather, she assesses *what John said*. In doing so, her implicit reasoning is something like this:

(1) John said that he's taller than me

(2) What John said = that he's taller than me

(3) He's not taller than me

(4) That he's taller than me is true if and only if he's taller than me

(5) What John said is true if and only if he's taller than me

(6) What John said is not true

Here (1) is just the indirect speech report, which (2) reworks in a context where nothing else John said is relevant. Line (3) states something Janet knows or believes about John's height compared to hers. Line (4) is just an instance of a standard logical schema for propositional truth, which does not involve disquotation, since 'that' is not a device for quotation:

(T) That P is true if and only if P

Principles not unlike (T) can already be found in Plato's *Sophist* and Aristotle's *Metaphysics*. Line (5) follows from (2) and (4) by the logic of identity (Leibniz's law), since (2) licenses substituting 'what John said' for 'that he's taller than me' in (4). The conclusion (6) follows from (3) and (5) by modus tollens for the biconditional (using its left-to-right direction), a standard principle of propositional logic.

Plato and Aristotle pair their principles about truth with corresponding principles about falsity not unlike (F):

(F) That P is false if and only if not-P

The instance of (F) corresponding to (4) is (4*):

(4*) That he's taller than me is false if and only if he's not taller than me

Just as Janet can derive (5) from (2) and (4), she can derive (5*) from (2) and (4*):

(5*) What John said is false if and only if he's not taller than me

The conclusion (6*) follows from (3) and (5*) by modus ponens for the biconditional (using its right-to-left direction), another standard principle of propositional logic:

(6*) What John said is false

Unless Janet suspects that John is lying, she may well conclude that what he *thinks*, as well as what he *said*, is false, and not true. She may go straight from the indirect speech report 'He said that he's taller than me' to the belief ascription 'He thinks that he's taller than me' ('think' is the usual term in ordinary English where philosophers say 'believe'; they are near-synonyms in this context). In effect, Janet uses what someone *says* as a heuristic for what they *believe*. The default assumption is *sincerity*: if someone says that P, they believe that P. Call that the *sincerity heuristic*.

What about the converse principle, a default assumption of *non-reticence*, that if someone believes that P, they say that P (when the question arises)? If they say that not-P, by the default assumption of sincerity, they believe that not-P, and so do not also believe that P, unless they are inconsistent. But if they say neither that P nor that not-P, can we assume by default that they have no belief either way? Obviously not, when the question whether P did not even arise in the conversation. But if they positively *refuse* to say that P, when the question does arise, a reasonable default assumption is that they lack the belief that P. As usual, the default can be inhibited: for instance, when the matter is confidential, or the speaker did not understand the question, or was unable to speak. Call that the *non-reticence heuristic*.

One can get from a *direct* speech report to a belief ascription by first applying the (suitably modified) homophonic disquotation heuristic and then applying the sincerity heuristic to the result. This can lead to Frege puzzles about co-referential terms such as 'Hesperus' and 'Phosphorus'.

For example, imagine this speech:

NN: Some people confuse Mike Brearley, the former captain of the England cricket team, with J.M. Brearley, the former lecturer at Newcastle University. They are not the same person. J.M. Brearley was once a professional philosopher. Mike Brearley was never a professional philosopher.

NN is mistaken. Mike Brearley, the former captain of the England cricket team, *is* J.M. Brearley, the former lecturer in philosophy at Newcastle University.

Imagine Brearley overhearing NN's speech.

When NN says 'J.M. Brearley was once a professional philosopher', Brearley can use the homophonic disquotational heuristic to make the indirect speech report 'NN said that J.M. Brearley was once a professional philosopher', but normal conversational standards for the use of pronouns *also* allow him to report 'NN said that I was once a professional philosopher'. Since NN's sincerity is not in question, Brearley then applies the sincerity heuristic to infer 'NN believes that I was once a professional philosopher'.

When NN says 'Mike Brearley was never a professional philosopher', Brearley can use the same heuristic to report 'NN said that Mike Brearley was never a professional philosopher', but normal conversational standards for the use of pronouns *also* allow him to report 'NN said that I was never a professional philosopher'. By the sincerity heuristic again, Brearley infers 'NN believes that I was never a professional philosopher'.

Putting the pieces together, Brearley ends up concluding 'NN believes both that I was never a professional philosopher and that I was once a professional philosopher', thereby accusing NN of having mutually contradictory beliefs. Yet NN may be a leading classical logician, with a militant aversion to inconsistency.

In that respect, the threatened contradiction is in NN's beliefs. But contradiction also threatens Brearley's own beliefs, by the non-reticence heuristic. For NN is far from reticent, and he is clearly unwilling to say 'Mike Brearley was once a professional philosopher'. Thus, by the homophonic disquotational heuristic, Brearley can report 'NN is unwilling to say that Mike Brearley was once a professional philosopher', but normal conversational standards for the use of pronouns *also* allow him to report 'NN is unwilling to say that I was once a professional philosopher'. Brearley then applies the non-reticence heuristic to conclude 'NN does not believe that I was once a professional philosopher'. But, as seen above, Brearley has already concluded 'NN believes that I was once a professional philosopher'. The threatened contradiction is now in Brearley's own beliefs (about NN's beliefs), not just in NN's beliefs.

Of course, there is a long history of trying all sorts of ingenious strategies to resolve the inconsistencies, from Frege's distinction between sense and reference to contemporary contextualist accounts of the implicit constraints on the guises or modes of presentation of the relevant objects under which the subject must conceive them in taking the putative attitude, for the attitude ascription to count as true. But when Janet complains to a friend 'John thinks that he's taller than me', she does not seem to be implying that, in so doing, John thinks of

her in some way relevantly similar to the way in which she thinks of herself, or anything of the kind; the issue of the guise under which John thinks of her seems not to arise at all. Naturally, one can imagine deviant cases where John thinks of her in some surprisingly convoluted way, but most things we say can be true in surprising ways. Rather than assume that some elaborate semantic apparatus is needed to explain the puzzle cases, we should explore the hypothesis that they are just predictable outcomes of our fallible heuristics for attitude ascriptions, as the Brearley example illustrates. That may be the right moral to draw from Saul Kripke's article 'A Puzzle about Belief' (1979), even though it is probably not the one he intended—what he seems to treat as an incoherence in the very concept of belief may be better understood as manifesting the inevitable limits of some of our ordinary, useful heuristics for ascribing belief (see chapter 4 for more discussion).

The sincerity and non-reticence heuristics are obviously specific to belief, and do not generalize in any straightforward way to other propositional attitudes, such as hope, fear, and intention. One would expect the human capacity for what psychologists call 'mindreading' to comprise heuristics for many different such attitudes. Furthermore, the sincerity and non-reticence heuristics are specifically based on *speech* behaviour. Yet we also apply our mindreading capacity to ascribe propositional attitudes, including beliefs, to pre-linguistic and non-linguistic creatures, such as very young children and non-human animals, often thereby explaining their behaviour much better than we could if we refrained from ascribing such attitudes to them. We may therefore need other mindreading heuristics to operate on non-linguistic behaviour.

How far can all these mindreading heuristics be unified? After all, linguistic and non-linguistic behaviour are not totally independent of each other, and propositional attitudes are interrelated in various ways: hopes and fears are connected to beliefs about the probabilities of good and bad outcomes, and intentions to beliefs about what one will do. To what extent different mindreading heuristics can all be understood as applications of one more general mindreading heuristic is an open question.

One should not assume that the default is always *not* to ascribe an attitude, in the absence of positive behavioural evidence—such as speech—for ascribing it. In particular, for the central attitude of *knowledge*, the default may be the other way round, to ascribe knowledge of truths unless there is some specific reason not to. For the most efficient cognitive policy may be to treat the world as by default open to view for all potential knowers, and then track specific obstacles to cognitive access. Metaphorically, if each of us carries around a mental map of the world in our head, I don't want to carry around mental maps of everyone else's mental maps, and so on *ad infinitum*. It would be easier just to carry around one mental map, mark on it where others are, and make further requisite adjustments on that basis in more or less systematic ways, or at worst *ad hoc*, rather than treating other minds as by default blank slates. With such an open-world heuristic, we will ascribe plenty of knowledge to creatures who exhibit no speech-like behaviour at all (Williamson 2024b, section 8). Given that knowledge is treated as entailing belief, we will ascribe plenty of beliefs to them too—at least when the occasion arises, since there is most point in attributing belief when we are not willing to attribute knowledge. Thus several more or less independent heuristics or sub-heuristics can combine, or even compete, in ascribing the presence or absence of the same attitude to the same subject at the same time. The result is not

‘conceptual incoherence’ but just what one might expect when several methods or sources of evidence are available for answering the same question.

We have seen how homophonic disquotational principles for the ascription of belief generate paradoxes of belief. Notoriously, and for related reasons, homophonic disquotational principles for the ascription of truth and falsity generate Liar-like semantic paradoxes. From the present perspective, such paradoxes are evidence that an underlying heuristic is at work. Although (T) and (F) are not strictly disquotational themselves, they are still associated with versions of the Liar paradox.

For example, I say ‘What I’m saying is not true’. The corresponding first-person present-tense indirect speech report is (7) (which I can think rather than say):

(7) I’m saying that what I’m saying is not true

In a context where nothing else I say is relevant, I can rework (7) as (8), just as Janet could rework her indirect speech report (1) as (2) above:

(8) What I’m saying = that what I’m saying is not true

The relevant instance of (T) is (9):

(9) That what I’m saying is not true is true if and only if what I’m saying is not true

Just as Janet could derive (5) from (2) and (4) above by the logic of identity, so I can derive (10) from (8) and (9), substituting ‘what I’m saying’ for ‘that what I’m saying is not true’ in (9):

(10) What I’m saying is true if and only if what I’m saying is not true

But (10) is a contradiction, since it is of the form ‘P if and only if not-P’, and so cannot be true, given classical logic.

In the analogous paradox for (F), I say ‘What I’m saying is false’. The relevant indirect speech report is (7*):

(7*) I’m saying that what I’m saying is false

In a context where nothing else I say is relevant, I can rework (7*) as (8*):

(8*) What I’m saying = that what I’m saying is false

The relevant instance of (F) is (9*):

(9*) That what I’m saying is false is false if and only if what I’m saying is not false

In the same way as before, I can derive (10*) from (8*) and (9*), substituting ‘what I’m saying’ for ‘that what I’m saying is false’ in (9*):

(10*) What I’m saying is false if and only if what I’m saying is not false

But (10*) is another contradiction, since it too is of the form ‘P if and only if not-P’.

These paradoxes have been taken to warrant revision of classical logic, in particular by accepting some instances of ‘P if and only if not-P’. From the present perspective, such drastic reactions look methodologically perverse. There is a far more obvious suspect: the

homophonic disquotational heuristic for speech reports. We already know that it is only a heuristic, as the elementary case of pronouns and other indexicals makes clear. With the indirect speech reports (7) and (7*), the problem is not with the personal pronoun 'I'. Rather, the natural explanation is that in uttering the sentence 'What I'm saying is not true' or 'What I'm saying is false' in the envisaged contexts, I fail altogether to say that something is the case. No positive indirect speech report at all is appropriate. Such failures may be initially surprising, but they violate no law of logic. Since (7) and (7*) are to be rejected, the paradoxical arguments do not even get started.

A natural objection is that the underlying problem does not really depend on indirect speech reports, because it still manifests in direct speech reports such as (7D) and (7*D):

(7D) I'm uttering 'What I'm uttering is not true'

(7*D) I'm uttering 'What I'm uttering is false'

Here 'utter' is used in place of 'say' to indicate that a relation to sentences rather than to propositions is in play. What is uttered is a sentence. So understood, (7D) and (7*D) are much harder to deny than (7) and (7*). Where no other utterances are relevant, we then have the required equations:

(8D) What I'm uttering = 'What I'm uttering is not true'

(8*D) What I'm uttering = 'What I'm uttering is false'

Since the problem now concerns the truth or falsity of sentences, it requires appropriately modified analogues of (T) and (F). The closest analogues are these familiar disquotational schemata:

(TD) 'P' is true if and only if P

(FD) 'P' is false if and only if not-P

The paradoxical arguments can then proceed much as before, with quotation marks in place of 'that' and 'utter' in place of 'say'.

However, a reason for restricting homophonic disquotational indirect speech is also a reason for restricting (TD) and (FD). To put it schematically, when in uttering 'P' you fail to say that P, you cannot be expected to have said something that is true if and only if P, or false if and only if not-P. For instance, when you utter the sentence 'I'm hungry', you do not say that *I*'m hungry, so I do not expect the sentence as uttered by you to be true if and only if *I*'m hungry, or false if and only if *I*'m not hungry. More generally, (TD) and (FD) should be restricted to contexts where the homophonic disquotational schema (D) also holds:

(D) In uttering 'P', one says that P.

A gloss is needed, for an actor can utter a declarative sentence on stage without asserting that anything is the case, and so in a sense without really saying that anything is the case. For purposes of disquotation, we can understand 'say' more liberally than that. Such non-assertive utterances will form another case where the sincerity heuristic for belief ascription is inhibited.

Of course, when you utter 'I'm hungry', you say something that is true if and only if *you* are hungry, and false if and only if *you* are not hungry, for you say that *you* are hungry. Thus, the proper generalizations are something like these non-homophonic principles (where *s* is a sentence):

(TG) In contexts where, in uttering *s*, one says that P, *s* is true if and only if P

(FG) In contexts where, in uttering *s*, one says that P, *s* is false if and only if not-P.

From (TG) and (FG), one can recover the homophonic principles (TD) and (FD) respectively for contexts where (D) holds. The paradoxes are resolved because one cannot recover the relevant instances of (TD) and (FD) in the relevant contexts, since (D) fails there (see Williamson 1998 and Andjelković and Williamson 2000 for some relevant discussion). For the sentential as well as the propositional versions of the paradoxes, the culprit is the homophonic disquotational heuristic for indirect reported speech. A similar diagnosis applies to versions of the paradoxes for thought rather than speech.

Although the specific problems for disquotation differ between Frege puzzles and semantic paradoxes, they both manifest its rough-and-ready character. Naturally, much remains to be explored about exactly where and why homophonic disquotational speech breaks down. In particular, we need to understand better the mechanisms of its failure in semantic paradoxes, which may also help explain its failures elsewhere. Since we already have decisive independent evidence that homophonic disquotation has merely heuristic status, postulating failures in unrelated principles—such as those of elementary propositional logic—is gratuitous and methodologically wrong-headed.

6. *The weighing heuristic for reasons*

Talk of 'reasons' is central to much contemporary debate in metaethics and, more generally, metanormativity. It promises to unify the practical with the theoretical: there are both reasons for action and reasons for belief. The term 'reasons' is assumed to be intellectually perspicuous enough to serve in the most abstract reasoning, yet also securely enough rooted in pre-philosophical normative thought and talk to ground what we say in concrete cases. There is even a research programme with the slogan 'Reasons first', which proclaims that the category of reasons is explanatorily fundamental (Schroeder 2021).

The use of the word 'reasons' in the plural is a reminder that we need some way of thinking and talking about *combining reasons*, on pain of being left at a loss when more than one reason bears on our decision. For example, in a group debate on whether or not to take a certain course of action, each side may present various considerations for and against taking that course, and the group faces the challenge of combining those considerations and resolving them into a decision one way or the other. As a single individual, one may carry out a similar process in one's own head.

We do indeed have such a way of combining reasons, for we often speak of 'weighing reasons', 'adding up' or 'balancing' the 'pros and cons', the 'reasons for' and the 'reasons against', and of reasons that 'outweigh' other reasons. The metaphor is of a pair of scales,

with reasons-for going into one pan, reasons-against into the other, and the decision for or against depending on which pan goes down, which up. The metaphor is not inert. It structures our thinking about what to do or what is the case, when we think about more than one reason. *Without* this organizing metaphor, our thought about reasons would be in danger of impotence.

The metaphor of weighing reasons is in effect an *additive* model. If you put two lumps of metal into a pan, the added weight is the sum of the weight of one lump and the weight of the other. Likewise, two reasons-for add up to a weightier case-for than either reason-for by itself.

Such an additive model has costs as well as benefits. For sometimes it gives the wrong result. Here is a simple case. A number labelled 'N' has been chosen from the set {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}, by a random draw. You have to guess whether 'N is even' or 'N is odd'; if you are right you win $100, if you are wrong you lose $100. A perfectly trusted and trustworthy informant, X, tells you just 'N is in the set {2, 4, 6, 7}'. X's testimony is a good reason for guessing 'N is even', since its probability on X's testimony is 75%. Another perfectly trusted and trustworthy informant, Y, tells you just 'N is in the set {7, 8, 10, 12}'. By parity of reasoning, Y's testimony is another good reason for guessing 'N is even', since its probability on Y's testimony is again 75%. But X's testimony and Y's testimony together amount to a decisive reason *against* guessing 'N is even', since the conjunction of what you learn from X's testimony and what you learn from Y's testimony entails that N is 7. Thus two good reasons for doing something can together make a decisive reason against doing it, contrary to the additive model of weighing reasons (see Titelbaum 2019 for more extensive discussion of such cases).

Friends of the additive model tend to object to such examples along the following lines. When you have just one of the two testimonies, it is a reason for guessing 'N is even'. But once you have *both* testimonies, each of them is a reason *against* guessing 'N is even', given what else you know. The trouble with such replies is that they effectively abandon the weighing metaphor as a useful way of structuring our thinking about how to combine reasons. If putting a second lump of metal into one pan of the scales may cause both lumps to jump into the other pan, all bets are off. Less metaphorically, such replies on behalf of the weighing model presuppose that we already have some *other* way of thinking about how to combine reasons, so that we can determine the new strength and valence of each reason once it is combined with the other reasons. Of course, in examples with a simple probabilistic structure like that above, we *do* have such an alternative structure, because we can work with conditional probabilities, as the discussion implicitly illustrated. The real work of combining the two testimonies was done in the framework of probability theory, not in the framework of reasons theory (if there is such a thing). A serious defence of the reasons framework must show how to combine reasons *within* that framework, not by abandoning it. Rendering the additive model harmless by rendering it impotent does not constitute such a serious defence.

Friends of the reasons framework can do better by treating the weighing metaphor as a convenient *heuristic* for combining reasons. It assesses the weight of each reason, and which pan it goes into, *separately*, and then combines the results additively. As a result, it will sometimes give the wrong answer, as in the example above. Nevertheless, its friends can plausibly claim, such examples tend to be rather artificial: the additive model may typically

give the right answer in realistic cases. In many such cases, any assignment of numerical probabilities would be highly artificial, while the reasons framework is in much less danger of imposing a false precision, and may be psychologically more realistic as a model of human thinking.

Failures of the additive model are not just intellectual curiosities. They can have practical consequences. The contested term 'intersectionality' may sometimes be used to get at such practically important failures of the additive model. For example, in an assessment of reasons for compensating someone for discrimination, if the weight of their being a black woman is equated with the sum of the weight of their being black and the weight of their being a woman, then in some circumstances a serious injustice will be done (compare Crenshaw 1989, the seminal text on intersectionality).

Whether the category of reasons is really as useful or as fundamental as proponents of the 'Reasons first' programme like to claim is not a question to be conclusively settled here. Still, one may wonder how fundamental the ideology of weighing reasons really is. After all, the metaphor makes sense only in a society familiar with the mechanism of a balanced scale. Although the technology for weighing and balancing is modest, there may not have been much need of it under evolutionary conditions. In any case, the reasons framework seems much better adapted to the regulation of debate than to tracking perception and memory—the acquisition and retention of the knowledge that should inform the debate. It is a strange child who acquires the category of reasons before they acquire the category of knowledge. Indeed, *having* a reason is arguably a matter of knowing the relevant fact, so that the ideology of reasons has to be explained in terms of knowledge, not the other way round (Hawthorne and Magidor 2018). But even if the category of reasons does not go very deep in the human cognitive system, we still cannot use it properly without heuristics to help us determine the results of combining reasons.

7. Implications for philosophical methodology

The last four sections presented various ways in which reliance on unacknowledged heuristics may have distorted our philosophical understanding—in particular, of vagueness, conditionals, belief, truth and falsity, and reasons. Specifically, what look like clear counterexamples to philosophical and logical theories may be the misleading artefacts of fallible heuristics.

This concern should not be confused with the 'negative program' characteristic of the early stages of 'experimental philosophy', which tried to demonstrate by surveys that philosophers' verdicts on hypothetical cases were too sensitive to subjects' ethnicity or gender to be reliable.[3] By contrast, many heuristics like those above are so general and so useful that they may well turn out to be more or less universal features of the human cognitive system, and not susceptible to significant variation with ethnicity, gender, social class, or other such factors. Of course, in the long run, the presence or absence of those heuristics in cognition over various human populations can and should be tested experimentally. However, since none of the heuristics at issue is specifically *philosophical*—each of them is targeted on a general class of cognitive challenges that frequently arise in

ordinary life—they will be best investigated in the broader setting of cognitive psychology. They do not call for a special experimental branch of philosophy, though naturally frequent two-way interaction between philosophically informed psychologists and psychologically informed philosophers is likely to benefit both sides.

How should we react to the discovery that we have been relying on fallible heuristics? Don't panic! After all, sense perception has long been known to rely on heuristics whose limitations result in perceptual illusions, but it would be melodramatic to conclude that we have no perceptual knowledge. Generic sceptical arguments from the occurrence of heuristic-induced errors are no better than generic sceptical arguments from the occurrence of errors of other kinds. Whatever kind of reliability or safety from error knowledge requires, it is local, not global. If a heuristic is humanly universal, or nearly so, it is likely to have survived because it is adaptive; in the most straightforward case, a heuristic is adaptive because it tends to give correct results in normal cases.

In particular, we should be wary of drawing pessimistic methodological conclusions for philosophy from our reliance on fallible heuristics. The heuristics are not themselves specific to philosophy; they underpin much of our thinking in general. Since our reliance on them does not warrant generic scepticism, assuming it to warrant philosophy-specific scepticism would be arbitrary.

Still, such general reflections do not warrant complacency. We should at least ask what improvements on our current philosophical methodology might make it less vulnerable to heuristic-induced illusions. That is work for the following chapters. It is not easy, for if we are heuristic-using creatures, we are probably creatures who *need* to use heuristics. We can sometimes correct their outputs, but in correcting them we may well rely on other heuristics, or even on other applications of the *same* heuristic. Nevertheless, methodological improvements *are* feasible, and they will call into question some currently fashionable ideas.

The role of sense perception in natural science is a helpful precedent here too. Without sense perception, natural science is simply impossible. Although scientists use artificial aids such as microscopes and telescopes, measuring instruments and computers, at some point or other they must be able to see or hear or touch at least some of the results. To put it crudely: if you are hallucinating, you are in no fit state to do science. Yet human sensory systems are riddled with fallible heuristics. In effect, scientists have learnt how to control their reliance on sense perception in ways that minimize the risks and costs of misperception. Incidentally, they have *not* done it as many epistemological internalists do, by treating subjective perceptual appearances as foundational: such appearances are quite unsuitable to play the role of scientific evidence, since they are not open to inter-subjective checking. Rather, they have applied whatever external controls were needed to resolve specific problems of misperception as they were identified. Something analogous may be possible, and necessary, to control the risk of errors induced by the more abstract heuristics prevalent in philosophy, such as those above.

Before we turn to ways of controlling the risk, its general nature could do with some further clarification. In discussing the *reliability* or *unreliability* of heuristics, one typically presupposes that their outputs are judgments, classifiable as *true* or *false*. The heuristic's degree of reliability may then be identified with the relative frequency of true to false outputs. In practice, reliability is often a more complex matter. If the heuristic is inferential, with

premise-like inputs, then what counts is truth-*preservation* from inputs to output, rather than just the truth of the output, and the degree of reliability may be identified with the relative frequency of true outputs *given true inputs*. If the heuristic's output is an *estimate* rather than a judgment, it may be assessed on a graded scale of accuracy, rather than on the binary distinction between truth and falsity. One may in turn relativize all such standards of reliability to specified conditions under which the heuristic was applied. And so on. Yet, irrespective of all these complications, reliability is still defined in terms of a standard of truth or accuracy given quite independently of the heuristic itself. More specifically, the heuristic has been assigned no role in determining the *content* of the judgments or estimates which it outputs. That may look like a bad picture when the heuristic is central to our practice of making judgments or estimates with those contents. For example, one might take the disquotational heuristics for ascribing belief and truth and falsity to be at least partially *constitutive* of the meanings of the words 'believe', 'true', and 'false'.

At the opposite extreme, a heuristic—probably not so-described—may be treated as an 'analytic' or 'conceptual' connection, quasi-definitional of the terms at issue. That may induce a philosophical crisis when the heuristic turns out to be inconsistent, at least given uncontroversial background knowledge, as with those above: however important to our lives the practices which involve those terms, they suddenly look 'incoherent'. But, as also emerged in those case studies, once the heuristics are properly identified, they are rarely promising candidates for 'analytic' or 'conceptual' status. Not only are the heuristics inconsistent, given our background knowledge: they fail in straightforward, unpuzzling cases—especially once we strip out the *ad hoc* apparatus of qualifications added as afterthoughts to disqualify exceptions, with no 'analytic' or 'conceptual' guarantee that no further qualifications will need to be added as further exceptions turn up.

On a better, intermediate alternative, heuristics lack 'analytic' or 'conceptual' status, but still play a role in determining the meanings of the relevant terms. This is at the level of *metasemantics*, the study of the factors on which the semantics of a language as used by a given community supervenes, or at least constitutively depends. At that level, something like a principle of charity operates, to favour interpretations which maximize the attribution of true beliefs or (as I prefer) knowledge to the community, given whatever other constraints on interpretation are operative (Williamson 2007/2021a, chapter 8). The heuristics used by the community or its members belong to the putative supervenience base for the metasemantics. They form a significant part of what has to be interpreted charitably.

Of course, no community or individual is omniscient, or error-free, and something is wrong with any metasemantic theory that implies otherwise. Inconsistent heuristics merely increase how much ignorance or error must be ascribed. Charitable interpretations still do what they can for a much-used heuristic, making it more rather than less reliable, though not perfectly reliable. For instance, we saw how the material interpretation of 'if' might do that for the suppositional heuristic for assessing conditionals. Despite the persistence heuristic's sorites-susceptibility, it can still exert pressure towards assigning a predicate a *convex* region of the relevant similarity space for its extension. Informally, the convex closure of a shape is the result of filling in all its holes and hollows, and a convex shape is one which is already its own convex closure. More formally, a region is convex just in case any point directly between two points in the region is itself in the region. Violations of convexity tend to

multiply counter-instances to persistence without necessity, so persistence militates in favour of convexity. Of course, the convexity constraint falls far short of uniquely determining predicate extensions; typically, the similarity space can be partitioned into convex regions in many different ways. Some of those may be eliminated because they violate other natural constraints (see Gärdenfors 2000 and Douven and Gärdenfors 2020 for more discussion). Still, we have no grounds to expect natural constraints to achieve uniqueness: a residual element of happenstance is likely to remain in the determination of reference.

One general strategy for charitable interpretation is *contextualist*: by varying the assignment of reference to a term with the context in which it is used, the strategy grants itself the flexibility to count more utterances as knowledgeable, or at least true. Contextualist strategies have been applied to all the kinds of case in which heuristics like those above are used: vagueness, conditionals, ascriptions of belief, truth and falsity, and reasons. However, since the heuristics are applicable even within a single context—which contributes to their power and usefulness—contextualism still cannot make them come out perfectly reliable.

Contextualist strategies have their own drawbacks, often overlooked. They do poorly when information in verbal form is transmitted across contexts through memory and testimony, unless agents keep track of the relevant features of all those contexts (Williamson 2005). For example, on some contextualist theories of belief ascriptions, the truth-condition of the sentence 'John believed that Cicero was a Roman orator' varies with which guises John has to have believed the proposition that Cicero was a Roman orator under for the sentence to be true. Believing the proposition under the guise of the sentence 'Tully was a Roman orator' may count in some contexts but not in others. Thus, if John loses track of the original set of contextually relevant guises, he in effect loses track of the belief ascription's original truth-conditions, and so is ill-placed to use the stored sentence in a new context, for instance, to pass on information to someone else. Thus, contextualist strategies open up myriads of new error-possibilities for speakers who do not carefully store lots of information about the contexts in which they originally acquired linguistically encoded information. Speakers unaware of such contextualist features of the semantics of their language will be especially liable not to do the hard work of storing all that information.

If we store that information about linguistic contexts in linguistic form, an infinite regress threatens. Even if we do not store the information in linguistic form, we are still in danger of having to back up all semantic memory with episodic memory of contexts, which is psychologically quite implausible.

Of course, obviously context-sensitive terms such as pronouns and demonstratives already do impose burdens of adjustment to changing contexts, which speakers and hearers usually manage to handle, often automatically, but contextualist strategies tend to multiply those burdens drastically, with no serious check on whether the benefits really outweigh the costs. That going contextualist conduces to more charitable interpretation is much less clear than it is normally taken to be. In particular, one should not be too optimistic about the prospects of making heuristics like those above come out much more reliable on a contextualist semantics. For their inconsistency was established with respect to the underlying level of content, whereas contextualism is just a doctrine about the mapping of form to content. Although contextualists may hope to limit how far the inconsistency is manifested in actual speech situations, that is likely to involve *ad hoc* complications. If the

contextualist can easily model whatever data come in, scientists would tend to regard that as a warning sign of bad science, for reasons explained in the next chapter.

In short, the heuristics on which we often rely in philosophy may be very rough indeed. The next chapter will consider some methodological consequences of that conclusion.

For the present, we may console ourselves with one reflection. Although the role of heuristics in our pre-theoretical assessments of examples makes our lives harder methodologically, because our data are less reliable than we thought, it also holds out the prospect that true answers to our theoretical questions may often be much simpler than we thought, because true, simple answers have already been wrongly dismissed on the basis of what are really heuristic-generated fool's counterexamples.

Notes

1 The phrase ‘tolerance principle’ goes back to Wright 1976. Williamson 2020: 63-7 treats tolerance principles as heuristics, though not with the generality of the persistence heuristic, in relation to sorites paradoxes, and provides numerical estimates of their reliability in some cases. Williamson 2022b (a review essay on Dorr, Hawthorne, and Yli-Vakkuri 2021) makes the generalization to the persistence heuristic. The latter exchange contributes to a debate about whether the S4 axiom (that what is possible is possibly possible) holds for metaphysical possibility, despite apparent examples of series of cases where each case entails the possibility of the next, but the first case does not entail the possibility of the last because the difference between neighbouring cases (for instance, in the original constitution of a given artefact) is ‘small enough’ but the difference between the first case and the last is ‘too large’. Salmón 1989 argues that the cases are genuine counterexamples to S4, Williamson 1990 that the underlying motivation for his premises is soritical and so unsound, and Salmón 1993 that the motivation is not soritical. Dorr, Hawthorne, and Yli-Vakkuri 2021 argue for a metasemantic approach on which a tolerance principle for constitution as uttered in a given possible world expresses a true proposition, but which proposition it expresses is contingent, while the corresponding necessitated tolerance principle expresses a false proposition. They defend the unnecessitated tolerance principle by non-soritical, epistemological considerations. Williamson 2022b responds that the pre-theoretic appeal of the tolerance appeal extends to the necessitated tolerance principle, because it does not depend on thinking of the cases as actual, and is best explained as deriving from the persistence heuristic. Similarly, the appeal of the crucial premises in Salmón’s anti-S4 reasoning is easily explained as deriving from the persistence heuristic. When an independently attested heuristic validates an assumption, explaining the latter’s pre-theoretic appeal in some other way is typically ill-motivated. Incidentally, although the persistence heuristic does not figure in the case for an epistemicist account of vagueness in Williamson 1994, our reliance on it supports my approach there.

2 See Eklund 2002 for an account of how principles can be ‘analytic’ without being true.

3 The seminal paper for the negative program was Weinberg, Nicols and Stich 2001, of which Nagel 2012 is an effective critique. For further criticism of the negative program see Williamson 2011a and 2016d. Many early results of experimental philosophy have turned out not to be repeatable under more rigorous conditions. For instance, after more extensive experimentation, early claims that the Gettier ‘intuition’ (that the subject of a classic Gettier case lacks knowledge) depends on ethnicity and gender have been replaced by the hypothesis that the Gettier ‘intuition’ is part of a humanly universal folk epistemology (Machery, Stich, Rose, Chatterjee, Karasawa, Struchiner, Sirker, Usui, and Hashimoto 2017). Most contemporary experimental philosophy is not involved in the negative program. Sytsma and Buckwalter 2016 is a wide-ranging recent survey of experimental philosophy.