HEURISTICS AND PARADOXES

TIMOTHY WILLIAMSON

Abstract. The paper proposes that philosophical paradoxes are the result of our reliance on efficient but fallible humanly universal heuristics. This is illustrated in relation to paradoxes concerning vagueness and conditionals. The paper is based on parts of chapter 1 of the author's forthcoming book *Heuristics and Overfitting in Philosophy* (New York, Oxford University Press).

Keywords: heuristics; paradoxes; conditionals; vagueness; semantics; metasemantics.

WHAT ARE HEURISTICS?

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.

Timothy Williamson ✉
University of Oxford

Rev. Roum. Philosophie, **67**, *2*, p. 313–332, Bucureşti, 2023

Psychologists have studied many heuristics intensively. Sometimes they characterize heuristics negatively, as "cheap and dirty", in the tradition of Daniel Kahneman[1], sometimes more positively, as "fast and frugal", in the tradition of Gerd Gigerenzer[2]. 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[3]. 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*[4]. 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,

[1] Daniel Kahneman, Paul Slovic, Amos Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge, Cambridge University Press, 1982.

[2] Gerd Gigerenzer, Ralph Hertwig, Thorsten Pachur (eds.), *Heuristics: The Foundations of Adaptive Behavior*, New York, Oxford University Press, 2011.

[3] Roland Fleming, "Human perception: visual heuristics in the perception of glossiness", in *Current Biology*, 22: R865-R866, 2012; Gerd Gigerenzer, "Embodied heuristics", in *Frontiers in Psychology*, 12:711289, 2021, doi: 10.3389/fpsyg.2021.711289.

[4] Gerd Gigerenzer, Daniel Goldstein, "Reasoning the fast and frugal way: models of bounded rationality", in *Psychological Review*, 103, 1996, pp. 650–669.

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 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[5]. 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[6]. 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.

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.

[5] *Ibidem*.
[6] Hilary Kornblith, *On Reflection*, Oxford, Oxford University Press, Chicago Press, 2012.

Philosophers may be tempted to postulate that what we *really* use is not the first-proposed crude heuristic but some complex refinement of it, constructed 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. Of course, philosophers may use the refined heuristic themselves in their consciously controlled theorizing, but they should not attribute it to ordinary pre-reflective human cognition.

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 paper, 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 two sections, I will explain and discuss two 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.

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". To stamp each piece of present-tense information with an expiry date for its validity as it goes into memory would involve significant expenditure of time and energy, for questionable benefits – inefficient, and probably infeasible. Naturally, most memories fade away, at different rates, but that does not mean that the timetable for their doing so has to be written into their content.

What we treat as too small to matter is sensitive to our vague, general sense of realistic timescales for different states and activities: "He is thin" or "He is asleep", "She is writing a novel" or "She is writing an email". How all this works is a topic for detailed psychological investigation. For present purposes, what counts is the general form of the phenomenon, not the specifics of its implementation.

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 content rather than form of the 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.

Predictive processing models of perception may also rely on present-tense updating. For example, Andy Clark writes about the perception of a moving object against a stable background: "most of the background information for the present frame can be assumed to be the same as the previous frame"[7]. Without such assumptions, the task of prediction might become intractably complex.

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

[7] Andy Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, New York, Oxford University Press, 2016, p. 26.

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 made 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 paper, as you considered the various hypothetical cases presented above.

Naturally, what counts as a small change depends on what we are talking about – the table, the house, the city, the country, the planet. A noticeable difference in taste between two bites of the same apple may surprise us more than between two bites of different apples from the same tree. Differential standards for smallness surely have to be calibrated by experience. But most of this happens offstage, without troubling consciousness.

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. We have experienced no relevant analogue of taking a second bite of an apple and suddenly tasting something rotten to make us cautious. 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.

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. 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.

THE SUPPOSITIONAL 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[8]. 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

[8] See Timothy Williamson, *Suppose and Tell: The Semantics and Heuristics of Conditionals*, Oxford, Oxford University Press, 2020, henceforth "S&T", for a book-length discussion of the heuristic.

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[9] (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[10].

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:

[9] Frank Ramsey, "General propositions and causality", 1929, MS. Reprinted in Hugh Mellor (ed.), *Foundations: Essays in Philosophy, Logic, Mathematics, and Economics*, London, Routledge & Kegan Paul, pp. 133–151, p. 143, to which page numbers refer.

[10] Jonathan Evans, David Over, *If*, Oxford, Oxford University Press, 2004; Igor Douven, *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches*, Cambridge, Cambridge University Press, 2016; Timothy Williamson, *Suppose and Tell: The Semantics and Heuristics of Conditionals*, Oxford, Oxford University Press, 2020.

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[11]. 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

[11] Timothy Williamson, "Knowing by imagining", in Amy Kind, Peter Kung (eds.), *Knowledge through Imagination*, Oxford, Oxford University Press, 2016, pp. 113–123.

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 we do not treat as 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[12]. 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, and A * C formalize "If A, C". Applying the simple version of the suppositional heuristic gives the equation Prob(A * C) = Prob(C | A), the identification of the probability of the conditional with the corresponding conditional probability, as proposed by various authors[13]. For Prob(A * C) is the probabilistic assessment of "If A, C", while the conditional probability Prob(C | A) is the probabilistic assessment of C on the supposition A, that is, with all but the A-possibilities excluded. The same connection holds for the generalized version of the suppositional heuristic. Let B be the conjunction of the background suppositions. Then applying the generalized heuristic to assignments of probability results in the equation Prob(A * C | B) = Prob(C | A ∧ B),

[12] Timothy Williamson, *Suppose and Tell: The Semantics and Heuristics of Conditionals*, pp. 142–146.

[13] Richard Jeffrey, "If" (abstract), in *Journal of Philosophy*, 61, 1964, pp. 702–703; Brian Ellis, "An epistemological concept of truth", in Robert Brown, C. D. Rollins (eds.), *Contemporary Philosophy in Australia*, London, Routledge, 1969, pp. 52–72; Robert Stalnaker, "Probability and conditionals", in *Philosophy of Science*, 37, 1970, pp. 64–80.

which is in effect the previous equation conditionalized on B. This is the generalized version of the identification of the probability of a conditional with the corresponding conditional probability. For Prob(A * C | B) is the probabilistic assessment of "If A, C" on the supposition B, while Prob(C | A ∧ B) is the probabilistic assessment of C on the suppositions A and B. The generalized equation 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[14]. 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. Of course, philosophers can seek consistent semantic approximations to the generalized probabilistic identity, but the identity is just one manifestation of a more general heuristic, which has non-probabilistic manifestations too. Treating the probabilistic case in isolation is arbitrary.

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[15].

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 Γ ⊢ A * C with Γ∪{A} ⊢ 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 Γ ⊢ A * C to Γ∪{A} ⊢ C is in effect modus ponens (the conditional elimination rule), while the implication from Γ∪{A} ⊢ C to Γ ⊢ 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".

[14] David Lewis, "Probabilities of conditionals and conditional probabilities", in *Philosophical Review*, 95, 1976, pp. 581–589; T. Williamson, *Suppose and Tell: The Semantics and Heuristics of Conditionals*, pp. 42–43.

[15] *Suppose and Tell: The Semantics and Heuristics of Conditionals*, pp. 37–42 presents the proof in more detail.

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 Prob(A * C | B) = 0 if and only if Prob(C | A ∧ 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 ¬(A * C) equivalent to the negated conjunction ¬(A ∧ 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"[16]. 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 → 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. In effect, the suppositional heuristic is responsible for the "paradoxes" of material implication. 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 purely 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

[16] Benson Mates, "Diodorean implication", in *Philosophical Review*, 58, 1949, pp. 234–242, p. 234.

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 identified with provability of the negation, rather than being treated as an independent form of assessment. In effect, mathematical proofs work with acceptance as the only operative mode of assessment. Near enough the *only* way an unembedded sentence occurs in a mathematical proof is as proved from – deductively entailed by – the set of relevant suppositions. In limit cases, that set is either empty or just the singleton of the sentence itself (in the speech act of supposing it). To that extent, the standard logical framework of mathematical is just like that of a natural deduction system. In such a setting, a material reading of "if" is the only one to validate the suppositional heuristic.

The primacy of acceptance over rejection in mathematical practice may be rooted in a more general pattern of human thought: to register rejection of "A" by accepting "Not A", replacing a negative attitude to a positive sentence by a positive attitude to its negation. "Not A" may then in turn be fleshed out in more positive terms[17]. If the default attitude to a sentence occurring in inner speech is acceptance, this would tend to avoid mental clutter, by reducing the need for special attitude-markers. Such a cognitive tendency would be efficient for both outright attitudes and attitudes under suppositions. It would set one up to apply the suppositional heuristic to acceptance, for which it gives good results. That would help explain why the heuristic's inconsistency causes so little trouble in practice, in mathematics or elsewhere, without any special training. Although it would not strictly resolve the inconsistencies lurking in the heuristic, especially as applied to probabilistic assessments, it would help limit the damage.

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

[17] On the psychology of negation, see Barbara Kaup, Rolf Zwaan, Jana Lüdtke, "The experiential view of language comprehension: how is negation represented?", in Franz Schmalhofer, Charles Perfetti (eds.), *Higher Level Language Processes in the Brain: Inference and Comprehension Processes*, Mahwah, Lawrence Erlbaum Associates, 2007, pp. 255–288.

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[18].

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 $\rightarrow$ 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[19].

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

[18] *Suppose and Tell: The Semantics and Heuristics of Conditionals*, pp. 89–102, discusses such cases in detail.

[19] *Ibidem*, pp. 103–110.

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.

IMPLICATIONS FOR PHILOSOPHICAL METHODOLOGY

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

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.

We cannot understand all this by treating the heuristic as the major premise of a deductive argument, an unrestricted universal generalization which will inevitably be false and so no basis for knowledge, just as we cannot understand perceptual knowledge by treating it as based on deductions whose major premise is that perception is perfectly reliable. No such premise is in play; it is neither assumed nor needed. Most cognition is not deductive. Like other biological processes, it often functions properly even though it is capable of functioning improperly.

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 elsewhere. 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.

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 objective probability of true outputs conditional on true inputs. 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.

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[20]. 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.

[20] Timothy Williamson, *The Philosophy of Philosophy*, Oxford, Wiley-Blackwell, 2007, chapter 8.

Some of those may be eliminated because they violate other natural constraints[21]. 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.

The heuristics on which we often rely in philosophy may be very rough indeed.

[21] For more discussion, see Peter Gärdenfors, *Conceptual Spaces: The Geometry of Thought*, Cambridge, MIT Press, 2000; and Igor Douven, Peter Gärdenfors,"What are natural concepts? A design Perspective", in *Mind and Language*, 35, 2020, pp. 313–334.