# Where Did It Come From? Where Will It Go?

[draft chapter for Arturs Logins and Jacques-Henri Vollet (eds.) *Putting Knowledge to Work: New Directions for Knowledge-First Epistemology*, Oxford University Press, revised 17 November 2022]

Timothy Williamson

Many lines of thought combined into *Knowledge and its Limits* (KAIL), and many spread out from it—not only in my own head. In this informal account, looking back, I will comment briefly on the historical antecedents for a knowledge-first approach in epistemology, and autobiographically on how I came to take such an approach. Looking forward, I will sketch promising new ways in which the approach is being extended and deepened. To give the big picture, I will use rather schematic formulations, and skip over many details and nuances. I make no attempt to survey the massive body of literature relevant to each topic; the chapter is long enough as it is.

Thematically, the chapter is organized into eleven sections:

1. *Looking back: some history*
2. *Looking back: indiscriminability*
3. *Looking back: knowledge and action*
4. *Looking back: contents and attitudes*
5. *Looking back: assertibility*
6. *Looking back: 'knowledge first'*
7. *Before and after*
8. *Looking forward: mindreading*
9. *Looking forward: knowledge and action*
10. *Looking forward: models of knowledge*
11. *Looking forward: epistemic norms*

The structure is non-linear: topically, the backward-looking sections converge on a point, from which the forward-looking sections spread out.

## 1. *Looking back: some history*

Some people find the knowledge-first approach wildly counterintuitive. Others find it the merest common sense. The divergence was encapsulated for me when I gave a talk at the University of New Mexico in 1994, based on material which later became 'Is knowing a state of mind?' (Williamson 1995) and then grew into the first two chapters of KAIL. In the discussion period, with reference to my claim that 'knowledge' is unanalysable, one professor exclaimed 'If I thought you were right, I would give up philosophy!' Also present was a

professor from the department of English literature department, who had come out of loyalty to my father, having known me as the teenage son of his old doctoral supervisor at Oxford. When we talked afterwards, I could see that he was wondering what all the fuss was about, and why it was worth giving a paper answering so easy a question—of course knowing was a mental state, what else would it be? I felt more sympathy with the scholar of English literature than with the philosopher. That knowing is a mental state, and the knowledge-first approach more generally, are common sense, though that by itself does not prove their correctness. Those who find them counterintuitive have internalised a specific epistemological tradition, which prioritises appearance over reality, and the proximal over the distal.

The situation has been obscured by the legend that the justified true belief analysis of knowledge was standard in Western epistemology from Plato to Gettier. If so, a belief-first approach would have been dominant for well over two thousand years. But that history is wrong. Although Plato toyed with a justified true belief analysis, he did not endorse it, and the popularity of the belief-first approach is a far more recent phenomenon (Dutant 2015). The JTB analysis is pointlessly redundant if justification itself already entails truth. The 'fallibilist' idea that truth-entailing justifications are too much to ask may be associated with the collapse of old certainties—for example, that space is Euclidean—under pressure from modern science. Such episodes are salutary, but do not warrant the panic-stricken moral that no truth-entailing propositional attitude—such as knowledge—is epistemologically fundamental, any more than cases of reference failure—for example, with the word 'phlogiston'—show that the relation of reference is not semantically fundamental. That philosophers and scientists were once certain of knowing that space was Euclidean, even though it was not, just shows that humans can make mistakes in applying the distinction between knowledge and ignorance, not that they cannot have truth-entailing knowledge.

The general knowledge-first approach is not at all original with me, though of course I hope to have taken it further. I suspect that cognitive relations more like knowing than like believing have tended to be central to epistemology whenever and wherever it has been practised. Often, those knowledge-like relations are modelled on the relation of seeing (an object). Even epistemologists who prioritise appearance over reality typically do so on the assumption that we have some sort of immediate awareness of appearances which we lack to the rest of reality. Such immediate awareness is in effect knowledge turned inward and implausibly idealised.

The natural human way of thinking about matters epistemological is, I suggest, knowledge-first (see also Nagel 2014 and forthcoming, and below). For instance, I have been delighted to learn, much classical Indian epistemology takes a knowledge-first approach (for one case see Vaidya 2022, Williamson 2023b; see also the chapter by Nilanjan Das in the present volume). In a wider epistemological perspective, KAIL is no aberration.

Closer to home, there is a long tradition of Oxonian realism, going back at least to John Cook Wilson, my predecessor as Wykeham Professor of Logic at Oxford from 1889 to 1915. Here is a clear articulation of his knowledge-centred conception of the mind, from his posthumously published work *Statement and Inference* (Wilson 1926: I, 38, 1967: 19-20):

> The unity of the activities of consciousness, called forms of thinking, is not a universal which, as a specific form of the genus activity of consciousness, would cover the whole nature of each of them, a species of which thinking would be the name and of which they would be sub-species, but lies in the relation of the forms of thinking which are not knowing to the form which is knowing.

He also explicitly rejected the demand for a definition of knowledge (Wilson 1926: I, 39, 1967: 20).

Cook Wilson was the dominant Oxford philosopher of his time. Although his contemporary, the idealist F.H. Bradley, now has more name-recognition, Bradley was an isolated and reclusive figure. It was Cook Wilson who set the tone—a realist tone, exemplified by the then-popular anti-Kantian slogan 'Knowledge makes no difference to the known' (on early Oxford realism see Marion 2000). It was a healthy attitude to start from, though doubtless too crude; doesn't the knowledge that one is thinking make a difference to the known, one's own thought?

By the time I studied mathematics and philosophy as an undergraduate at Oxford (1973-6), Cook Wilson was almost forgotten. However, the relevant section of *Statement and Inference* was included in a then-standard book of readings on knowledge and belief (Phillips Griffiths 1967), where I read the Cook Wilson excerpt in 1974. I cannot pretend to have been much impressed. It seemed both eccentric and dogmatic. Nor was what I subsequently learned about Cook Wilson encouraging. He bitterly opposed various forms of progress, despising non-Euclidean geometry and non-Aristotelian logic, then represented in Oxford by Lewis Carroll, aka the mathematician Charles Lutwidge Dodgson. Cook Wilson also treated the difference between knowing and not knowing, quite implausibly, as always in principle accessible to the thinker, in stark contrast with KAIL. He affirmed a strong form of the 'positive introspection' thesis that whenever one knows, one knows that one knows: 'The consciousness that the knowing process is a knowing process must be contained within the knowing process itself' (Wilson 1926: I, 107). He did not quite endorse the 'negative introspection' thesis that whenever one doesn't know, one knows that one doesn't know, since in cases of error, he says, someone 'doesn't know and is unaware that he doesn't know' (op. cit.: 106). However, he goes on to describe those as cases of partially 'unawakened consciousness' (op. cit. 110-11) involving 'a certain passivity and helplessness' (op. cit. 113). Presumably, when consciousness is fully 'awakened', negative introspection holds. Cook Wilson never influenced me directly.

Cook Wilson's legacy was mainly indirect. One of his pupils was H. A. Prichard, White's Professor of Moral Philosophy at Oxford from 1928 to 1937. In his work on epistemology, Prichard overtly followed Cook Wilson; KAIL notes my disagreement with his claims that one is always in a position to know whether one knows or merely believes and that knowing entails not believing (Prichard 1950: 86-8). As students at Oxford, both J. L. Austin and Wilfrid Sellars were impressed by Prichard's lectures (Berlin 1973, Sellars 1975). Their realist emphasis on the priority of the predicate 'is F' to the predicate 'seems F' may show Prichard's influence, and through him Cook Wilson's. Austin is also likely to have read *Statement and Inference* itself; it was a major work of Oxford philosophy published shortly before he began his studies. One might connect Austin's emphasis in *Sense and Sensibilia* (Austin 1962) on the inappropriateness of talk of evidence in straightforward cases of

perceptual knowledge with Cook Wilson's attitude that 'In knowing, we can have nothing to do with the so-called "greater strength" of the evidence on which the opinion is grounded; simply because we know that this "greater strength" of evidence of A's being B is compatible with A's not being B after all' (Wilson 1926: 100). Methodologically, Austin's claim 'our common stock of words embodies all the distinctions men have found worth drawing, and the connexions they have found worth marking, in the lifetimes of many generations' (Austin 1956-7: 8) is reminiscent of Cook Wilson's 'Distinctions current in language can never be safely neglected' (1926: 46, 1967: 26).

For several decades after Austin's early death in 1960, the most distinctive manifestation in Oxford of its realist tradition was disjunctivism about perceptual appearances, on which (to put it schematically) 'O looks F' divides into something like a disjunction of the 'good' disjunct 'O is visibly F', which entails 'O is F', and the 'bad' disjunct 'O merely looks F', which entails 'O is not visibly F' (Hinton 1967 and 1973, Snowdon 1980-1 and 1990, McDowell 1982, Martin 2004, Byrne and Logue 2009). When I returned to Oxford in 1988, after eight years as a lecturer at Trinity College Dublin, disjunctive theories of perception had for years formed one of the main topics of debate amongst Oxford philosophers. However, the theory of perception was not one of my special interests, though I had a general interest in epistemology. The disjunctivist who did most to generalize the view in epistemology was McDowell (1982, 1995), but I had always found his Wittgensteinian quietism and gnomic style intellectually alien. He had no direct influence on me, though in retrospect I could see significant similarities in epistemological outlook. Anyway, the possibility of a general disjunctivism about belief was clear; I recall raising it in conversation with my colleague Mike Martin in 1988: 'believes *p*' would divide into the disjunction of the good disjunct 'S knows *p*', which entails that *p* is true, and the bad disjunct 'S merely believes *p*', which entails 'S does not know *p*'. Cook Wilson already held a sort of disjunctivism about cognitive attitudes: 'Beyond then the bare abstraction of conscious activity, there is no general character or quality of which the essential natures of both knowledge and opinion are differentiations, or of which we could say in ordinary language that each was a kind' (1926: 100).

KAIL provides a sympathetic critique of disjunctivism about belief (Williamson 2000: 44-8). I objected that the disjunctive form is just a misleading artefact, because the equivalence of 'S believes *p*' and 'Either S knows *p* or S believes *p* without knowing *p*' is merely a trivial consequence of the accepted entailment from 'S knows *p*' to 'S believes *p*'. The bad disjunct 'S believes *p* without knowing *p*' is gerrymandered and unnatural. As I made clear, the objection speaks more to the letter than the spirit of disjunctivism. What matters most is the insight that 'S knows *p*' cannot be understood as the conjunction of 'S believes *p*' with other factors, all of them somehow prior to 'S knows *p*' itself. Thus disjunctivism about belief is better re-worked as anti-conjunctivism about knowledge.

Analogous points apply to disjunctive theories of perception. The disjunctive form is again just a misleading artefact, because the equivalence of 'O looks F' and 'Either O is visibly F or O looks F without being visibly F' is merely a trivial consequence of the entailment from 'S is visibly F' to 'S looks F'. The bad disjunct 'O looks F without being visibly F' is gerrymandered and unnatural. Again, the objection speaks more to the letter than the spirit of disjunctivism. What matters most is the insight that 'O is visibly F' cannot be

understood as the conjunction of 'O looks F' with other factors, all of them somehow prior to 'O is visibly F' itself. Thus disjunctivism about perceptual appearances is better re-worked as anti-conjunctivism about perception.

For both belief and perceptual appearances, disjunctive theories can take subtler forms, but variants of the same objection still apply.

Already when I was a student, philosophers' search for 'conceptual analyses' corresponding to key terms of natural language such as 'know', 'cause', and 'mean' struck me as a degenerating research programme. There was the familiar depressing pattern of a proposed analysis, a counterexample, and the insertion of another epicycle to give a new proposed analysis. Nor did it feel as though successive analyses were approximating the true analysis more and more closely, if the true analysis was supposed to make explicit our implicit understanding of the term. If such ramshackle definitions were implemented psychologically, by something like complicated lexical entries for the defined terms in our heads, they would render those terms virtually unusable, a point that today's aspiring conceptual engineers would do well to remember. With every added epicycle, the proposed definition seemed to get *further* from psychological reality, not closer to it. Of course, one could give up the demand for psychological reality, and require only metaphysically necessary and sufficient conditions, presumably with some stipulation to avoid circularity. But it was still quite unclear why one should expect words like 'know', 'cause', and 'mean' to *have* analyses even in that less ambitious sense. Thus, from early on, I did not expect knowledge (the state, the concept, or the word) to be analysable. But since unanalysability would be such a common case, that by itself did not make knowledge primary in epistemological theorizing. It did not exclude the internalist view that what matters most in epistemology is a non-factive standard of justification. Other developments were needed too.

## 2. *Looking back: indiscriminability*

From the late 1980s onwards, I started to notice ways in which knowledge, rather than non-factive justification, is exactly what is needed for various philosophical purposes.

I had a consciousness-raising moment at a conference in Cambridge (England) to celebrate the fiftieth volume of the journal *Analysis*, in 1990. Jonathan Bennett gave a paper arguing for the failure of all extant attempts to explain why belief is involuntary (published as Bennett 1990). In the subsequent question-and-answer period, I idly suggested that one could try first explaining why *knowledge* is involuntary, which might be easier, since you cannot come to know at will a given proposition to be true. Perhaps one could then use a connection between belief and knowledge to extend the explanation from knowledge to belief. Bennett's response was that he had been philosophizing all his career without employing the concept of knowledge, and he advised me to do the same. The thought instantly struck me 'That's where you make your big mistake'. Nevertheless, I respected his methodological self-awareness, though not his epistemological judgment. I realized too that many other philosophers were in effect following the same policy as Bennett, perhaps with less self-awareness.

A clue to the importance of knowledge had already come to me through my work on the cognitive relation of *indiscriminability*. To explain its relevance, I must first digress, to

fill in some background. As a first-year undergraduate, I was intrigued by the challenge of finding a criterion of identity or abstraction principle for qualities and quantities as perceived by an observer (I no longer see the issues in such reductive terms). In the writings of Bertrand Russell and A. J. Ayer, I read that identity in a given perceived respect (for a given subject under given conditions) cannot be defined as indiscriminability in that respect, because identity in any respect is transitive, while indiscriminability in a respect is typically non-transitive, as many sorites series bear witness. For instance, imagine a long series of rods ordered by length, where each rod is indiscriminable in length from those next to it (by naked eye under given conditions) but the first rod is easily discriminable in length from the last. We cannot stipulate that, under those conditions, two rods are identical in *perceived* length just in case they are indiscriminable in (real) length. For then each rod is identical in perceived length to those next to it (by the stipulation), so the first rod is identical in length from the last (by the transitivity of identity, not indiscriminability), which is not so (again by the stipulation). Somehow, the associated logical issues gripped me. As an undergraduate, I proved a connection with the Axiom of Choice and wrote up a paper, which I submitted to *The Journal of Philosophy*. It received a revise-and-resubmit—quite reasonably, the editors wanted more on the philosophical significance of my technical results. In my ignorance, I misinterpreted the letter as a rejection with a patronising pat on the head, and angrily put the paper aside. A decade later, long after learning what the letter really meant, I got round to rewriting my paper and cashed in the revise-and-resubmit (Williamson 1986).

My renewed work on indiscriminability set me thinking more deeply about its nature, in ways which led to my first book, *Identity and Discrimination* (Williamson 1990). To continue the example, let the (real) length of $x$ be $\mathrm{l}(x)$. To discriminate $x$ and $y$ in length is to recognize that $x$ and $y$ differ in length, in other words, to come to know that $\mathrm{l}(x) \neq \mathrm{l}(y)$. Thus $x$ and $y$ are discriminable in length just in case, given the available information, it is epistemically necessary that $\mathrm{l}(x) \neq \mathrm{l}(y)$. Consequently, $x$ and $y$ are *in*discriminable in length just in case, given the available information, it is *not* epistemically necessary that $\mathrm{l}(x) \neq \mathrm{l}(y)$, or equivalently (by the duality of necessity and possibility) it is epistemically *possible* that $\mathrm{l}(x) = \mathrm{l}(y)$, which we can write as $\Diamond[\mathrm{l}(x) = \mathrm{l}(y)]$. That line of thought generalizes to any respect: indiscriminability in a respect is the epistemic possibility of identity in that respect.

This account nicely predicts the logical properties of indiscriminability, given the familiar logical properties of identity and of epistemic possibility, as follows.

Indiscriminability in a respect is reflexive on whatever the respect applies to: $x$ is indiscriminable in length from itself because $\mathrm{l}(x) = \mathrm{l}(x)$ is a logical truth on the domain (since identity is reflexive), so $\Diamond[\mathrm{l}(x) = \mathrm{l}(x)]$ is also a logical truth on that domain (if $\alpha$ is a logical truth, so is $\Diamond\alpha$).

Indiscriminability in a respect is also symmetric: if $x$ is indiscriminable in length from $y$ then $y$ is indiscriminable in length from $x$, for $\mathrm{l}(x) = \mathrm{l}(y)$ is logically equivalent to $\mathrm{l}(y) = \mathrm{l}(x)$ (since identity is symmetric), so $\Diamond[\mathrm{l}(x) = \mathrm{l}(y)]$ is logically equivalent to $\Diamond[\mathrm{l}(y) = \mathrm{l}(x)]$ (if $\alpha$ is logically equivalent to $\beta$, then $\Diamond\alpha$ is logically equivalent to $\Diamond\beta$).

But indiscriminability in a respect is *not* in general transitive, even though identity is transitive. For if $x$ is indiscriminable in length from $y$, and $y$ from $z$, then we have $\Diamond[\mathrm{l}(x) = \mathrm{l}(y)]$ and $\Diamond[\mathrm{l}(y) = \mathrm{l}(z)]$, but $\Diamond[\mathrm{l}(x) = \mathrm{l}(z)]$ does *not* follow, even though $\mathrm{l}(x) = \mathrm{l}(y)$ and $\mathrm{l}(y) = \mathrm{l}(z)$ together entail $\mathrm{l}(x) = \mathrm{l}(z)$ (since identity is transitive). The epistemic possibilities in which $\mathrm{l}(x)$

= l($y$) may not overlap the epistemic possibilities in which l($y$) = l($z$): crucially, expressions of the form 'l($v$)' are non-rigid designators; they designate different lengths in different epistemic possibilities, to capture the ignorance of length.

As yet, we have not exploited the specifically *epistemic* nature of the possibility expressed by ◊, compatibility with what is known. We could have made the same arguments while using ◊ to express rational *doxastic* possibility, compatibility with what is rationally believed. We need the 'rational' for reflexivity, so that the logical truth of α implies the logical truth of ◊α, for if what is believed is inconsistent, *nothing* is compatible with it.

But indiscriminability has a further logical feature, a kind of strengthened reflexivity: if things are identical in a respect, they are indiscriminable in that respect. If things are identical in length, you cannot discriminate them in length—though you may be under the illusion that you can. Thus l($x$) = l($y$) implies ◊[l($x$) = l($y$)]. That works for epistemic possibility: if α is true, α is compatible with what is known, since only truths are known, and all truths are mutually compatible. But it does not work for rational doxastic possibility, understood so that consistent false beliefs can be rational: if one rationally but falsely believes l($x$) ≠ l($y$), l($x$) = l($y$) is true while ◊[l($x$) = l($y$)] is false on the rational doxastic reading of ◊, since what is rationally believed contradicts l($x$) = l($y$). Thus, to capture the full logic of indiscriminability, one must invoke specifically *epistemic* possibility; rational *doxastic* possibility will not do, because the relevant body of information must contain only truths.

*Identity and Discrimination* discusses these formal matters in greater depth and detail (Williamson 1990: 10-42). When I wrote it, the need for knowledge rather than mere rational belief in understanding indiscriminability struck me as significant, because discrimination is so cognitively fundamental. If what is needed there is knowledge, that is some indication that it comes first. Of course, one could add in the truth requirement by hand, interpreting ◊ as rational *true* belief, but I already regarded such gerrymandering as an alarm signal, warning of a degenerating research programme.

In analysing the logic of indiscriminability, I used a framework of Kripke models for modal logic, with the modalities interpreted epistemically in the tradition of Hintikka (1962). For simplicity, I excluded quantifiers from the object-language, but included atomic formulas of the form l($x$) = l($y$). To interpret the modality, the model has an *accessibility* relation. Informally, a world $x$ is accessible from $w$, or epistemically possible at $w$, if and only if, for all one knows in $w$, one is in $x$. Then ◊$p$ is true at a world $w$ ($p$ is epistemically possible at $w$) if and only if $p$ is true at some world accessible from $w$. Dually, □$p$ is true at $w$ ($p$ is epistemically necessary at $w$) if and only if $p$ is true at every world accessible from $w$. Informally, epistemic necessity is identified with knowledge: □$p$ is true at $w$ if and only if one knows $p$ at $w$.

Even if the object-language lacks atomic formulas of the form l($x$) = l($y$), we can understand the models in terms of indiscriminability more directly, by regarding accessibility as itself a relation of indiscriminability between worlds. Thus the non-transitivity of indiscriminability becomes the non-transitivity of accessibility. As is familiar, the non-transitivity of accessibility in this framework corresponds to the invalidity of the formula □$p$ ⊃ □□$p$, which is the 'KK' or 'positive introspection' principle that if one knows $p$, one

knows that one knows *p*. Thus one may know *p* without knowing that one knows *p*, a central theme in KAIL.

KAIL's cover alludes to the non-transitivity of accessibility. It shows Nicolas Poussin's 'Landscape with a Man Killed by a Snake', a painting I have loved since I first saw it as a student in the National Gallery, London. On one common interpretation, it illustrates different levels of cognition: the man who sees the snake-entwined corpse, the woman who sees only the man who sees it, the boatmen in the background who have no idea what is going on. It thus shows the non-transitivity of seeing: the woman sees the man, the man sees the corpse, which she does not see. Seeing is often used as a metaphor for accessibility: a world *w* 'sees' a world *x* when *x* is accessible from *w*. Metaphorically, therefore, the painting illustrates the non-transitivity of epistemic accessibility.

Admittedly, the argument from the non-transitivity of indiscriminability to the non-transitivity of accessibility oversimplifies the situation. If it worked, the symmetry of indiscriminability would similarly imply the symmetry of accessibility. As is also familiar, the symmetry of accessibility in this framework corresponds to the validity of the formula $\neg p \supset \Box\neg\Box p$: if *p* is false, one knows that one does not know *p*. But that principle is notoriously counterexample-prone. In some sceptical scenarios, it is false that one has hands, but one is in no position to know that one does not know that one has hands, because for all one knows everything is normal and one knows that one has hands. Thus, although the normal good case is accessible from the sceptical bad case, the bad case is not accessible from the good case. In the good case, one knows that one is not in the bad case. In the bad case, one does not know that one is not in the good case.

The reason for this mismatch between indiscriminability and accessibility is that accessibility involves an asymmetry between two ways in which a world can be presented to a thinker. When '*x* is accessible from *w*' is unpacked as 'for all one knows in *w*, one is in *x*', it is envisaged that the world *w* is presented to one 'by acquaintance' as the world one is in, or 'the actual world', whereas the world *x* is presented to one 'by description' in some way. The good case is accessible from the bad case because, when one is in the bad case, one does not know 'The case I am in is not the good case'. The bad case is inaccessible from the good case because, when one is in the good case, one knows (or is in a position to know) 'The case I am in is not the bad case'. One knows more in the good case than in the bad case. Since the modes of presentation are not held fixed, the argument from the symmetry of indiscriminability to the symmetry of accessibility does not go through.

For similar reasons, the non-transitivity of accessibility cannot simply be subsumed under the non-transitivity of indiscriminability. In saying that world *x* is accessible from world *w*, we envisage *x* as presented 'by description', whereas in saying that world *y* is accessible from *x*, we envisage *x* as presented 'by acquaintance'. Thus the mode of presentation of *x* is not held fixed. Consequently, arguing from the non-transitivity of indiscriminability to the failure of the KK principle is not straightforward. Indeed, *Identity and Discrimination* uses models where indiscriminability is non-transitive but epistemic accessibility is transitive. Nevertheless, one can find cases where non-transitive indiscriminability makes epistemic accessibility non-transitive too. My article 'Inexact Knowledge' (Williamson 1992) does it in effect, by recycling the non-transitivity of the perceptual indiscriminability of quantities as the non-transitivity of the epistemic accessibility

of worlds. The trick was not to argue directly in terms of indiscriminability but to use limits on powers of discrimination to motivate a margin for error principle for knowledge, and then apply that. The argument shows that one can know *p* without even being in a position to know that one knows *p*, so that not even the watering down with the qualifier 'in a position to' saves the KK principle. In that respect, my version of the knowledge-first approach is very different from those of Cook Wilson and Prichard.

Later, I noticed that the argument of 'Inexact Knowledge' can be generalized to a very wide range of conditions C, with the conclusion that C can hold even though one is not in a position to know that C holds. The generalization appeared in my article 'Cognitive Homelessness' (Williamson 1996b), and later as the anti-luminosity argument in KAIL. The generalized argument plays a key role in my epistemological externalism. Internalism is often motivated by appeal to an access constraint on norms of rationality and the like: rational agents should always be in a position to know whether they are complying with a norm. Externalist norms—such as a knowledge norm on assertion or on knowledge—typically violate that constraint. For instance, one is not always in a position to know whether one has kept a promise. By applying the anti-luminosity considerations, I could argue that *no* non-trivial norm meets the internalist constraint, making externalism 'the only game in town'.

## 3. *Looking back: knowledge and action*

The knowledge-first approach was not just a development in epistemology. It also grew out of developments in the philosophy of mind and action.

In late twentieth-century analytic philosophy, the dominant framework for understanding intentional action was neo-Humean belief-desire psychology. Rational agents did things because they believed that doing so would get them what they wanted. The formal version of the view was *decision theory*, which graded the beliefs into credences (represented by subjective probabilities of states of the world) and the desires into preferences (represented by subjective utilities of states of the world). Why pair belief and desire? An attractive answer was that they represented opposite *directions of fit*: the point of belief was to fit mind to world; the point of desire was to fit world to mind. A potential warning sign was that the contrast in direction of fit came from Elizabeth Anscombe, who was of course far from being a neo-Humean herself (Anscombe 1957).

Once one considers the relation between knowledge and belief in the context of belief-desire psychology, another question naturally presents itself: what stands to desire as knowledge stands to belief? In direction of fit, knowledge and belief are on the same side. Belief *should* be fitted to the world: when the belief *does* fit the world (truth), not by merely happening to match it but because all went well in the process upstream from the belief, there is knowledge. Analogously, if X stands to desire as knowledge stands to belief, X must have the same direction of fit as desire. Desire *should* fit the world to it: when the world *does* fit the desire (satisfaction), not by merely happening to match it but because all goes well in the process downstream from the desire, there is action—intentionally realizing the desire. On this view, *action* is what stands to desire as knowledge stands to belief.

That analogy strongly encouraged my incipient knowledge-first tendencies, from about 1990 on. Just as it would be perverse to sideline action in the philosophy of action, it would be perverse to sideline knowledge in the theory of knowledge. That was not a pedantic insistence on reading the phrases 'philosophy of action' and 'theory of knowledge' literally. Rather, the point was that one cannot properly understand desiring things, intending to bring them about, and trying to bring them about except in relation to the good case of intentionally bringing them about. Similarly, one cannot properly understand states like believing except in relation to the good case of knowing. Since the centrality of action to the philosophy of action was more widely accepted, I could use it as an entering wedge for the centrality of knowledge to epistemology.

## *4. Looking back: contents and attitudes*

Belief-desire psychology was typically treated as the core of a more general propositional attitude psychology. Such intentional states and acts varied on two dimensions: the content (believing *p* versus believing *q*) and the attitude to that content (believing *p* versus desiring *p*). Comparison of the two dimensions gave further support to the knowledge-first approach. I will explain some more background before discussing the connection with knowledge.

The 1970s saw the beginnings of a major intellectual switch from *content internalism* to *content externalism*. According to content internalism, the content of a thinker's propositional attitude at a time supervenes on what is internal to that thinker at that time: no difference in the attitude without an internal difference. Content externalism is the negation of content internalism; it denies supervenience. 'Internal' there could be understood in various ways: physically, in terms of brain states or bodily states, or phenomenally, in terms of qualia or the like. Such variations do not matter for present purposes; we can use the terms 'internal' and 'external' schematically.

Content externalism was originally driven by developments in the philosophy of language, concerning the semantics of natural kind terms and singular terms. The reference of such words does not supervene on what is internal to the speaker. In Hilary Putnam's famous example, the reference of 'water' as used in a given community depends on the chemical constitution of the samples to which the community has applied the word, even if the community has no inkling of the chemistry (Putnam 1973). Even pre-scientific speakers can understand that something may *apparently* belong to a natural kind without *really* belonging. Initially, Putnam took the moral of his example to be just that meaning is not in the head, while still treating psychology as in the head, so that psychology does not determine meaning. However, Tyler Burge soon pointed out that similar arguments show that propositional attitude psychology is also not in the head, a conclusion Putnam accepted (Burge 1979).

To vary the example: the term 'tiger' in English refers to members of a species T. We say 'There are tigers', speaking truly. We thereby express our belief that there are tigers, and our belief is true, for the belief that there are tigers is true if and only if there are tigers, and there are indeed tigers. Imagine a counterfactual possibility where the species T never evolved, so there are no tigers, but people just like us use a natural kind term just like 'tiger'

to refer to members of another species T* with which they interact. Superficially, members of T are indistinguishable from members of T*, but they share no common evolutionary ancestry and, owing to subtle genetic differences, would be incapable of interbreeding. Those people say something just like 'There are tigers', speaking truly. However, they do not express a belief that there are tigers, for they lack that belief. In their circumstances, a belief that there are tigers would be *false*, since by hypothesis there are no tigers; but those people are no more in error than we are. In saying 'There are tigers', they express a true belief which stands to T* just as our true belief that there are tigers stands to T, but what they believe is not what we believe. Those people differ from us in the content of their beliefs, but not internally. Thus content does not supervene on what is internal to the thinker.

One can make analogous arguments with perceptual demonstratives in place of natural kind terms. My counterfactual counterpart and I are exactly alike internally. I see a wasp; he sees an exactly similar wasp. My wasp never lived in his circumstances; his wasp never lived in mine. I say 'This wasp is alive', speaking truly. I thereby express my belief that this wasp is alive, and my belief is true, for the belief that this wasp is alive is true if and only if this wasp is alive, and this wasp is indeed alive. Counterfactually, my counterpart says 'This wasp is alive', speaking truly. However, he does not express the belief that this wasp is alive, for he lacks that belief. In his circumstances, a belief that this wasp is alive would be *false*, since by hypothesis this wasp is not alive; but he is no more in error than I am. In saying 'This wasp is alive', he expresses a true belief which stands to his wasp just as my true belief that this wasp is alive stands to my wasp, but what he believes is not what I believe. He differs from me in the content of his beliefs, but not internally. The conclusion is the same as before: content does not supervene on what is internal to the thinker.

In Oxford, content externalism became an increasingly prominent theme from the 1970s on. In the hands of John McDowell (1977) and Gareth Evans (1982), it took a quite distinctive form. Whereas Putnam and Burge focused on kind terms, Evans and McDowell focused on singular terms, especially perceptual demonstratives like 'this wasp'. Unlike Putnam and Burge, they concentrated on the contrast between the good case where reference succeeds and the bad case where it fails, for example, when the thinker is hallucinating, with no internal difference between the good and bad cases. They argued that in cases of reference failure there is no content, so the external difference is between content and no content, rather than between two internally indistinguishable contents. They framed the issues in neo-Fregean terms, positing 'object-involving' senses, rather than highlighting the failure of supervenience as such. Nevertheless, for present purposes, the similarities are more significant than the differences. One can easily rework the tiger and wasp examples with hallucination instead of another reference in the bad case.

Of course, content internalists did not give up without a struggle. Initially, their resistance tended to involve denying the descriptions of the cases. These denials typically depended on misunderstanding how attitude ascriptions work in natural language, and even on confusing use and mention. They forgot that normally when one uses indirect speech to ascribe a propositional attitude to another, the words in the content clause still mean what the speaker means by them, not what the other does. Later internalist resistance often took the more sophisticated form of conceding that the cases show that attitude ascriptions in natural language work in an externalist way, with 'broad' contents, while arguing on those very

grounds that such ascriptions do not cut at the underlying psychological joints. Furthermore, according to such internalists, the attitudes which *do* cut at the underlying psychological joints have 'narrow' contents, and appropriate ascriptions of them work in an internalist way. On this internalist picture, the core of the mental is purely internal, while attitude ascriptions in natural language are hybrids of the internal and the external.

The motivations for content internalism are various—some causal, some epistemic. As a rough generalization, content internalists who self-describe as naturalists tend to have a causal motivation, while content internalists who self-describe as anti-naturalists tend to have an epistemic motivation.

On the causal side, the main fear is that content externalism will imply some kind of magical action at a distance. Beliefs, desires, and other attitudes have causes (for instance, in perception) and effects (for instance, on action), but many reductionists view all the genuine causal work as entirely mediated by underlying brain states, and so best characterized in internalist terms, making distinctively externalist aspects of content causally irrelevant. They conclude that any causally relevant attitudes must have narrow contents.

On the epistemic side, the main fear is that content externalism will undermine one's privileged access to one's own present mental states. Many anti-reductionists hold that, as a rational subject, one has special non-observational conscious access to one's present beliefs, desires, and other attitudes, of a kind which no one has to the attitudes of anyone else. Since one has no such special access to distinctively externalist aspects of content, they conclude that any attitudes epistemically accessible in the special way must have narrow contents.

The trouble is that, although content internalists (of both types) need narrow content, they have no idea how to get it. Of course, they can start with two agents in exactly the same total internal state, and say that they share all their narrow contents. But what is hopelessly unclear is what it takes for two agents who are *not* in exactly the same total internal state to share a narrow content. You and I both believe that there are tigers. We share that broad content, individuated externally in terms of the natural kind: tigers. Yet our total internal states will differ in all sorts of ways. What does it take for someone to share the narrow content supposedly underlying your belief that there are tigers? Indeed, since one's total internal state is changing all the time, what does it take for you to retain that narrow content for five minutes? In brief, how are narrow contents to be abstracted from the flux of the internal? There are no pre-theoretic 'intuitive' answers to those questions, for narrow contents are just theoretical posits, promissory notes which have never been redeemed. By contrast, we have at least some understanding of how broad contents work, because we are continually attributing attitudes to them to each other.

Anyway, the causal and epistemic motivations for content internalism are far from convincing. Normally, behaviour to be causally explained is specified in general and external terms. Why did she flip the wasp out of the window? Why did Napoleon invade Russia? Good explanations may invoke mental states specified in similarly general and external terms. She believed that the wasp could sting her. He believed that Russia was militarily vulnerable. Attitudes to narrow contents may be much less explanatory. As for privileged access, content externalism does not preclude it. I know that I believe that this wasp is alive; my second-order knowledge reuses the same perceptual demonstration as my first-order

belief. In any case, privileged access has its limits. It is often hard to know what one believes or desires.

Although I have been writing in the present tense, that account of content internalism's troubles would already have elicited my assent in the early 1990s; such a negative assessment of content internalism was then already widespread, though of course far from universal. Indeed, content internalism's lack of progress over the intervening decades supports the negative verdict. There was ample time to find a solution to its problems, if one existed.

When I considered why epistemologists and philosophers of mind might resist the classification of knowing as a mental state, I realized that their objections would be like those to content externalism. They were often the very same objections, for what was supposed to play the causal role or to be accessible in the special way was the whole intentional state, which comprised both a content and an attitude to that content. Just as externalism about the content could generate externalism about the state, so could externalism about the attitude; the individuation of the whole state is worldly. Just as examples in favour of content externalism hold the attitude fixed while varying the content externally but not internally, so examples in favour of attitude externalism hold the content fixed while varying the attitude, again externally but not internally. Factive attitudes such as knowing-that, seeing-that, and remembering-that are blatantly externalist. For instance, you are in exactly the same total internal state in the good case and the bad case. In the good case, you know that it is raining. In the bad case, you still believe that it is raining, but you do not know that it is raining, for it is not raining. Consequently, the same general strategies could be used against the causal and epistemic objections to attitude externalism as had been successful against the causal and epistemic objections to content externalism. The details varied from contents to attitudes, but it was clear what to look for. I worked the analogy very hard in defending the claim that knowing is a mental state.

None of this means that content externalism entails attitude externalism, or *vice versa*. Neither the conjunction of content externalism with attitude internalism nor the conjunction of content internalism with attitude externalism is logically inconsistent. But those mixed views have no natural motivation. The natural motivating view both for content internalism and for attitude internalism is a general internalism about the mental, which rules out both mixed views. Conversely, once content externalism has ruled out general internalism, it makes sense to go for attitude externalism too, to complete a unified picture. I regarded content externalism as the first wave of the externalist revolution, and my attitude externalism as the second wave. Admittedly, leaders of the first wave were often unwilling to go along with the second wave—for them, it was an externalism too far. That is a common pattern in revolutions.

## 5. *Looking back: assertibility*

When I was an undergraduate and then doctoral student at Oxford (1973-80), the most influential senior figure in the philosophy of logic and language at Oxford was Michael Dummett. He supervised me for the final year of my doctoral studies, just after taking up the

Wykeham Chair of Logic. In stark contrast to his distant predecessor Cook Wilson, he tended towards anti-realism. He was sympathetic to a theory of meaning on which the meaning of a declarative sentence is given by the condition for it to be *assertible*, rather than by the condition for it to be true. He intended this as a generalization to all language of the intuitionistic treatment of mathematical language, on which the meaning of a mathematical sentence is given by the condition for something to be a proof of it. The idea was that realist truth-conditions problematically transcend speakers' *use* of the language, whereas assertibility-conditions are immanent in use. By contrast, Gareth Evans, John McDowell, Christopher Peacocke, and others then at Oxford, under the influence of Donald Davidson, preferred a realist theory of meaning in terms of truth-conditions. Consequently, assertibility-conditions were much discussed by Oxford philosophers at the time.

Dummett never presented a full assertibility-conditional theory of meaning for a non-trivial fragment of non-mathematical language. Unfortunately, his discussions tended to remain programmatic, apart from occasional intriguing suggestions. However, an assertibility-conditional theory of meaning was supposed to be more or less compositional: the meaning of a complex sentence is determined by the meanings of its constituents and the way in which they are put together. For example, the natural assertibility-conditional semantic clause for disjunction is this: 'A or B' is assertible when and only when either 'A' is assertible or 'B' is assertible. Dummett was well aware that one can legitimately assert that the number of people in the stadium is odd or even, without having counted them to find out which. He handled such cases by reading 'assertible' in the semantic clause as '*canonically* assertible' and ruling that one is entitled to assert a sentence when one knows a procedure for making its canonical assertibility-condition obtain. Such a theory of meaning is liable to invalidate the law of excluded middle, for it makes 'A or not A' assertible only if either 'A' is assertible or 'Not A' is assertible. Thus, if one knows no procedure for making either 'A' assertible or 'Not A' assertible, one is not entitled to assert even 'A or not A'. For example, 'A' might be 'The number of black holes in the past, present, and future of the universe is even'. For a committed proponent of classical logic, as I had been ever since I found out what classical logic is, such results were good evidence that something had gone wrong with the semantic theory.

Anyway, Dummett never gave a plausible account of canonical assertibility-conditions even for simple non-mathematical sentences. There are many ways of knowing that Jo is at home; any of them entitles one to assert 'Jo is at home', and English privileges no subset of them as the 'canonical' ones. What unifies those ways is that they are all ways of knowing *that Jo is at home*. Such a line of thought indicated that truth-conditions were explanatorily prior to assertibility-conditions.

My interest in the relation between assertion and knowledge was further piqued by reading Michael Slote's suggestive article 'Assertion and Belief' (Slote 1979). He was my Head of Department at Trinity College Dublin when I started my first full-time university job there in 1980. His article appeared in a volume of conference proceedings so obscure that I would probably never have seen it otherwise.

Reflection on indiscriminability and margins for error later made it clear to me that not even assertibility-conditions can be as epistemically transparent as Dummett took them to be. In any sense in which truth-conditions transcend use, so do assertibility conditions. Thus

his arguments for a theory of meaning in terms of the latter rather than the former must fail. Whatever norm governs assertion, one is not always in a position to know whether one is complying with it. Hence, although one is not always in a position to know whether one knows, that does not constitute a good objection to a knowledge norm of assertion. I was free to make a full-blooded case for the knowledge norm (Williamson 1996a).

## 6. *Looking back: 'knowledge first'*

Several other lines of research went into KAIL. For instance, in 1990-94 the economist Hyun Song Shin and I overlapped as Fellows of University College Oxford. We were both working on epistemic logic, and wrote two papers on it together (Shin and Williamson 1994, 1996). He introduced me to the rich literature by theoretical economists on epistemic logic, whose influence can be identified at several points in KAIL.

Since my first published article (Williamson 1982), I had from time to time used a framework of bimodal logic, with both alethic modal operators and epistemic operators, to investigate limits to knowability, in response to the so-called paradox of knowability, the proof that if all truths are knowable (as Dummettian anti-realists asserted) then all truths are known (as Dummettian anti-realists were reluctant to assert). The proof was first published by Frederic Fitch, who attributed it to the anonymous referee for an earlier paper he had submitted in 1945 but never published (Fitch 1963). The anonymous referee later turned out to be Alonzo Church (Salerno 2009). My attention was drawn to the result by work of Bill Hart and Colin McGinn (Hart 1979, Hart and McGinn 1976). My original interest was just to show that the full proof did not go through in intuitionistic logic, the preferred logic for Dummettian anti-realists, and to work out how intuitionists might treat the issues it raised. I found it an interesting intellectual exercise, even though my sympathies were all on the side of realism and classical logic. However, Dorothy Edgington proposed an alternative knowability principle for anti-realists that did not collapse into the claim that all truths are known even in the setting of classical logic (Edgington 1985). I argued that Edgington's alternative did not work in the way she needed it to (Williamson 1987a, 1987b; see further Edgington 2010 and Williamson 2021b). I continued to investigate knowability more generally in various classical settings (Williamson 1993).

By the time I came to write KAIL, Dummettian anti-realism no longer felt like a live option. I pointed out essential problems for his conception of assertibility, but only in passing. My interest was rather in the ubiquitous way in which our inherently limited powers of discrimination enable knowledge in many cases while disabling it in slightly different cases. Since I regarded classical logic as under no serious threat, I excluded my work on knowability in a non-classical setting from the book. I included only classical arguments for the existence of unknowable truths. Together with the anti-luminosity arguments, they are the limits of knowledge to which the book's title alludes—though of course there may be others.

Even in the mid-1990s, I was still not clear how the different papers I was writing fitted together, because I had come at them from different angles, with different interests. I did not write any of them as a mere application of a knowledge-first programme. Still, the role of knowledge in all of them was hard to miss, and the equation E = K of one's total

evidence with one's total evidence was an easy extension (Williamson 1997). Combining that equation with epistemic logic and a prior probability distribution gave a treatment of evidential probability as probability conditional on what one knows (Williamson 1998). It was obviously time to pull all this work, and a bit more, together in a book.

While finishing KAIL, I worried that readers might not see how to fit my position into their mental geography of possible epistemological theories. I was very conscious that my second book, *Vagueness* (Williamson 1994), had had a vastly greater impact than my first, *Identity and Discrimination*, in part because *Vagueness* had a much simpler and clearer take-home message: vagueness is ignorance. Up close, KAIL seemed far more intricate in structure than *Vagueness*. Stepping back, however, the unifying theme was obvious. I gritted my teeth, and put the slogan 'knowledge first' into the first sentence of the Preface. Even readers who got only that far would have some idea what the book was about.

## *7. Before and after*

When KAIL was published in 2000, I feared that epistemologists were too set in their ways of thinking to grasp, let alone adopt, the knowledge-first approach. Some reactions were indeed just as I had predicted, by authors who clearly had no idea how much of their accustomed framework I was rejecting—otherwise they would presumably not have taken it for granted without comment in their objections (some of the essays in Greenough and Pritchard 2009 are examples). I also knew that many epistemologists lacked the formal background in logic and mathematics to be comfortable with the more technical parts of the book.

Nevertheless, KAIL had far more impact than I expected. I was pleasantly surprised at how many epistemologists, especially younger ones, were willing and able to engage seriously with the knowledge-first approach. Epistemology was clearly ready for a change, even though old habits die hard. Some of that readiness came from a more general dissatisfaction with the model of philosophy as 'conceptual analysis', exemplified by the post-Gettier tradition of attempts to analyse 'the concept of knowledge'. One intended methodological moral of KAIL was that logical rigour in philosophy does not depend on conceptual analysis. It is better achieved by formal model-building and explicit theoretical hypotheses which make no claim to be 'analytic truths' or 'conceptual connections'.

KAIL's impact was much greater than the combined impacts of the separate articles out of which it was largely composed. Although analytic philosophy is often observed to be an outlier of the humanities in its publication patterns, and closer to the social sciences, with far more emphasis on articles in ranked journals and far less on monographs, books still play a key role. KAIL displayed the various components as working parts of a single theory, and the response showed that many people were looking for such a unified approach to epistemology.

Some developments and applications of KAIL took me by surprise. For instance, I was approached for a meeting in Oxford to discuss KAIL by a Christian missionary writing a doctoral dissertation based on his work in central Africa. He had initially tried to apply the literature on formal models of dialogue to analyse his conversations with non-Christians, but

had found it quite unhelpful. Instead, the knowledge norm of assertion turned out to give him the traction he needed. As a straightforward atheist, I had not had such applications in mind when working on assertion. Still, I could hardly object. Part of my case for taking assertibility to require knowledge rather than the dialectical ability to supply reasoned justification was that the latter gives too much weight to skill in smooth-talking confabulation—a professional skill of philosophers, which they are correspondingly liable to over-value. But the knowledge norm has also found application to cases where dialogue manifestly does have a highly formal, dialectical structure. In jurisprudence, Michael Blome-Tillmann has argued persuasively that courts' reluctance to admit strong but merely statistical evidence of individual involvement in wrong-doing is best explained by the hypothesis that the operative norm of evidence is implicitly interpreted as a knowledge norm (Blome-Tillmann 2017).

A dimension of generality that I did anticipate in KAIL is the range of potential knowers. I recognized that young children, non-human animals, and perhaps robots with AI can know truths. I deliberately left it open that social entities such as 'science' can possess knowledge, literally and non-derivatively (for social epistemology friendly to a knowledge-first approach see Bird 2010 and forthcoming, Carter, Kelp, and Simion forthcoming, and Kelp and Simion 2021).

I will discuss in more detail some current developments of the knowledge-first approach, in my work and that of others. The survey is by no means intended to be exhaustive. For example, space does not permit me to discuss the close links between knowing and having and acting for *reasons* (Hyman 2015, Hawthorne and Magidor 2018), or refinements of a *safety* constraint on knowledge (Williamson 2009), or the knowledge-first treatment of evidence (Williamson forthcoming[e], forthcoming[f]) or the complications raised by the semantics of attitude ascriptions—including knowledge ascriptions—in cases of co-reference (Williamson 2021d).

## *8. Looking forward: mindreading*

In KAIL, the case for knowing as a core mental state is based mainly on general philosophical considerations about externalism, causal explanation, self-knowledge, the logical form of attitude ascriptions, and so on. One footnote cites a discussion by the psychologist Josef Perner (1993) of evidence that children understand knowledge and ignorance *before* they understand belief and error, and so do not understand knowledge in terms of belief. I found that encouraging, but did not build on it. However, as Jennifer Nagel later noted (2013), psychologists routinely classify knowing as a mental state. That is not just a terminological point; it draws substance from how they treat the attribution of knowledge as just as central and basic an application of the human mindreading capacity as the attribution of beliefs or desires (see also Nagel 2017). In effect, the human cognitive system thrives on treating knowledge as a mental state.

There is increasingly strong evidence that the capacity to distinguish knowledge from ignorance is cognitively more basic than the capacity to distinguish true belief from error (for an introduction to the recent literature see Phillips, Buckwalter, Cushman, Friedman, Martin, Turri, Santos, and Knobe 2020 and associated discussion). Humans attribute knowledge and

ignorance before they can attribute true belief and error, and they tend to do it faster and more automatically. Nonhuman primates attribute knowledge and ignorance to each other, but not true belief or error. Indeed, that combination may extend much more widely across the animal kingdom. The best available explanations of much animal behaviour interpret them as making such distinctions. Reductive attempts to re-explain the behaviour in terms of mere reflexes become ever more *ad hoc* when faced with the complexity and flexibility of the behaviour. Claims to have found belief attribution at much earlier stages have not proven robust (see Nagel forthcoming, chapter 5, for discussion).

What very young children and nonhuman primates attribute is clearly knowledge-like, not some doxastic *ersatz* such as true belief. It is even sensitive to Gettier cases. For example, here is the experimenters' summary of two experiments with rhesus macaques (Horschler, Santos, and MacLean 2019):

> In Experiment 1, monkeys watched an agent observe a piece of fruit (the target object) being hidden in one of two boxes. While the agent's view was occluded, either the fruit moved out of its box and directly back into it, or the box containing the fruit opened and immediately closed. We found that monkeys looked significantly longer when the agent reached incorrectly rather than correctly after the box's movement, but not after the fruit's movement. This result suggests that monkeys did not expect the agent to know the fruit's location when it briefly and arbitrarily moved while the agent could not see it, but did expect the agent to know the fruit's location when only the box moved while the agent could not see it. In Experiment 2, we replicated and extended both findings with a larger sample, a different target object, and opposite directions of motion in the test trials.

In the background is a generic presumption of persistence: the default is that if the fruit is somewhere, it continues to be there, and that if the agent knows that it is there, the agent continues to know that it is there. In effect, when the monkeys see the agent see the fruit put there, they treat the agent as coming to know that it is there. They continue to treat the agent as knowing that it is there when the agent's view is temporarily occluded but the fruit remains there. Thus they are surprised if the agent reaches for the wrong box, presumably in order to get the fruit. But when the fruit is removed from the box, the monkeys cease to treat the agent as knowing that the fruit is there, since they can see that it isn't. When the fruit is directly put back into the box, they do not treat the agent as again coming to know that the fruit is there, since they can see that the agent did not see it being put back. Thus they are not surprised if the agent reaches for the wrong box. Had the monkeys been thinking in doxastic terms, they would have treated the agent in both conditions as believing throughout that the fruit is there (this is simply a point about belief; it does not depend on the assumption that knowledge entails belief). Thus there would be no difference in surprise between the two conditions if the agent reaches for the wrong box. Indeed, the belief that the fruit is there is true in the final stage of both conditions, and even justified, given the presumption of persistence. Since the monkeys reasonably treat the agent as not knowing that the fruit is there after it has been removed and replaced, that is in effect a Gettier case—although of course they do not think of it as a case of justified true belief.

The experimenters themselves interpret the monkeys as attributing only an 'awareness relation' rather than knowledge to the agent. However, their distinction between knowledge and awareness is unclear, and seems to depend on an unnecessarily doxastic conception of knowledge. The results of the experiments make just as good sense on the assumption that the

monkeys are distinguishing between knowledge and ignorance (see Nagel forthcoming, chapters 4 and 5, for more discussion, including of similar results for young children).

Many philosophers have found the idea that attributing knowledge is easier than attributing belief 'counterintuitive'. They assume that attributing knowledge *must* be harder, and require more sophistication, than attributing belief. Sometimes, the assumption comes from a vision of attributing knowledge as attributing some post-Gettier multi-clause *analysans* of knowledge in terms of belief, truth, and one or more other factors, which would of course be much harder, and require much more sophistication, than attributing belief alone. But even philosophers who do not envisage knowledge as having such an analysis often seem to assume that attributing knowledge must require *more* than attributing belief, simply because knowledge itself requires *more* than belief. After all, even in KAIL, knowledge entails belief, while belief does not entail knowledge. At a more general level, a similar thought may influence many internalists: broad mental states must be harder to identify than narrow mental states because identifying broad states requires monitoring *both* the internal *and* the external, whereas monitoring narrow states only requires monitoring the internal.

Such preconceptions are not surprising for *self*-attributions of mental states. But if, as is likely, mindreading capacities evolved through *social* life, their primary role is in attributing mental states to *others*. For that task, states purely internal to the other may be harder to determine than states involving the mutually observable environment. A simple initial case is the *absence* of factive states. Just from knowing that you didn't eat the banana, you can conclude that I don't know that you ate the banana—but you may still wonder whether I *believe* that you ate the banana.

Of course, attributions of positive mental states are more interesting. A good place to start is with *seeing an object*. When you see an apple and I see it too, typically, each of us can also see that the other sees it. We can check open eyes, direction of gaze, potential occlusions. That will not satisfy sceptics about other minds, but their sceptical scenarios were scarce in our evolutionary history. Similarly, when two people walking together both hear a loud noise, typically, each of them also knows that the other heard it. On the negative side, one may know that the alpha male can't see the apple, because a bush is in the way, or that he is too far away to hear one's breathing. Such knowledge about what others do or don't perceive plays a large role in communication, for example in the use of perceptual demonstratives. When young children interact with other children or adults, mutual gaze at an object is often crucial to communication.

The internal analogue of object-seeing is as-if object-seeing, being in a mental state internally the same as (really) seeing an object. Attributing as-if object-seeing is much more laborious. When I see that you see the apple, I can reason that since every mental state is internally the same as itself, you also as-if see an apple, but that is an artificial intellectual exercise. To consider cases where really seeing and as-if seeing come apart, we can suppose that dreaming that one sees an object involves as-if seeing an object without really doing so. If you see me when I'm asleep, gently snoring with my eyes shut, you know that I am not really seeing an apple, but you cannot tell whether I am as-if seeing an apple.

Although object-seeing is not itself a propositional attitude, it is closely related to propositional attitudes. One can see an apple without seeing *that* it is an apple, because it has an unusual shape, or one thinks it might be a wax replica, or one has been brought up in

ignorance of apples. Still, normally, when one sees an apple, one also sees that it is an apple. Conversely, when one doesn't see an apple, one also doesn't see that it is an apple. Thus it is unsurprisingly typical that when we see an apple together, each of us is in a position to know that the other sees that it is an apple. Seeing-that, 'fact-seeing', *is* a propositional attitude.

Psychologically, perhaps we model seeing that P on seeing an object, treating the state of affairs that P like an object. Just as you can't (really) see what isn't there, you can't (really) see what isn't the case. On this analogy, we treat the non-obtaining of the state of affairs that P like the absence of an object. Just as an object *o* must be there for you to see *o*, it must be that P for you to see that P. Moreover, both object-seeing and seeing-that normally require a suitable causal connection to what is seen: a merely accidental match of your visual image to something external, E, does not constitute seeing E.

In KAIL, I argue that seeing that P is a specific form of knowing that P. Thus, when we see the apple together, typically, each of us is in a position to know that the other knows that it is an apple. There would be little point in my judging merely that you *believe* that it is an apple, for why should I make that judgment if I doubt that you see that it is an apple?

Psychologically, seeing-that seems to be treated as a paradigm of knowing-that. 'See' is often used in an extended sense for a wide range of cases of knowing or recognizing (coming to know): 'I see your point'; 'I don't see how that follows'. What drives the generalization from literal seeing and other forms of sense perception to knowing? A crucial factor is *memory*. When you turn away, you no longer *see* that there are apples on the tree, but you still *remember* that there are (or at least were), and many of the effects on action are similar—you may still go to the tree when hungry. Remembering that P is another form of knowing that P. Having seen the agent see the fruit put somewhere, the rhesus macaques continue attributing knowledge that it is there to the agent even when they can see that the agent can no longer see the fruit. A large part of the excess of knowledge over sense perception is simply what remains when sense perception ceases.

In light of these considerations, knowledge attribution looks rather easier and more natural than philosophers' preoccupations can make it seem. We should not be surprised that the level of cognitive sophistication required for attributing knowledge turns out to be *lower* than the level of cognitive sophistication required for attributing belief—just as it can take less to recognize whether someone knows that P than to recognize whether they have an attitude internally similar to knowing that P.

Knowing also takes primacy when we learn from others about the world (Phillips et al. 2020). If you want to know whether P, but are not in a position to perceive whether P, it matters to you whether *I* know whether P. If I do, you can learn from me (whether I happen to have a *belief* as to whether P is not the issue). Imagine us facing each other. You can see things behind my back that I can't see; I can see things behind your back that you can't see. We may wish to share our knowledge: one of us sees signs of a predator and sounds the alarm. Or we may wish *not* to share our knowledge: one of us sees some delicious food and tries not to react, hoping to eat it once the other has gone. The other can benefit by spotting tell-tale signs that the first has spotted something. In such cases, to focus on the other's internal states is to miss the point.

A converging line of argument comes from considerations of cognitive efficiency, as Robert Gordon has observed (Gordon 2000, 2021; see also his chapter in the present volume).

Creatures with minds put huge effort into learning about their environment and what is happening in it, and keeping their information up to date—it can literally be a matter of life and death. If they are capable of mindreading, they use it to keep informed of similar cognitive states and processes in others. Imagine that whenever they represent something they must also separately represent how each of the others represent it (for instance, whether they believe, disbelieve, suspend judgment, or have some degree of a credence). That is a massive multiplication of effort. Indeed, it threatens to be infinite: I represent X, you represent how I represent X, I represent how you represent how I represent X, you represent how I represent how you represent how I represent X, …. For example, each creature maintains something like a map of its environment. But it also needs to track how each of the others maps the environment, so for each of the others it maintains another map of the environment, representing the other's map. That already threatens to be computationally infeasible, even before we start worrying about the infinite regress of maps of maps.

A much more efficient method would be to maintain just one map, but to try to mark the location of the other knowers on it. That already captures something of their different perspectives on the world. For example, it encodes information about what you can see but others can't, because their view is obstructed by an intervening obstacle. Similarly, it also encodes information about places they can see but you can't. Of course, that is only a start. The rhesus macaques already go further by tracking which present states of affairs another can still view through memory though no longer through sight. The child who can attribute false beliefs is doing something much more complex. Still, the underlying principle may be the same: in mindreading, the default is to treat the other as knowing; the work goes into tracking deviations from that. In Gordon's terms, the default is 'the shared world'. (Harvey Lederman, crediting Taylor Carman for the observation, pointed me to a passage in Merleau-Ponty 1945: 407-8 about the cognitive attitude of young children where he seems to endorse a similar idea.) By contrast, on the mistaken but widespread alternative, the default is to treat the other as a *tabula rasa*, so that attributing any positive mental state requires work.

Watering down the default from knowledge to true belief would make no sense. By default, everything lies open to everyone's view; in those circumstances, there is knowledge, not just true belief.

To make knowledge the default is not to assume that most agents know most truths. Even when that assumption is restricted to simple truths about the environment, it is surely false: think of all the truths about what insects are under what stones, and so on. In practice, mindreading is typically used for matters of actual or potential interest to the agents concerned. The point is that, on such matters, it is typically easier to work down from an initial hypothesis of total knowledge than to work up from an initial hypothesis of total ignorance.

The shared world default may well have been ecologically valid in the conditions under which mindreading evolved: small groups of conspecifics in a local environment, interactions between a predator and prey, and so on. One might worry, though, how much sense it makes in the modern world of highly complex, diverse societies. But that worry may underestimate the epistemic diversity already present under those evolutionary conditions. Even in a small group of hunter-gatherers, there are obvious epistemic asymmetries between adults and children. Children know that they know less than adults, and adults know that they

know more than children. Mindreading in both directions guides how children learn from adults. Within a group, differences in life history, recent experience, skills, and abilities, can all make for significant differences in knowledge and belief. When one group of hunter-gatherers encountered a new group, perhaps with alien customs, how each group mindread the strangers in those sensitive circumstances could make the difference between things going very well and things going very badly—crudely, between sex and death. Human history is not a simple narrative of increasing diversity; notoriously, imperialism and globalization work in the opposite direction. Even in the modern world, people of very different cultures and mindsets do manage to communicate, using a robust capacity for mindreading that evolved under radically different conditions. For that to happen, the shared world is a more effective default than the *tabula rasa*.

The shared world default may also help solve a long-standing problem in game theory and theoretical economics. Many results depend on the hypothesis that various background conditions such as rationality are *common knowledge* amongst the relevant agents, so everyone knows that everyone is rational, everyone knows that everyone knows that everyone is rational, everyone knows that everyone knows that everyone knows that everyone is rational, and so on *ad infinitum*. Demanding such common knowledge of normal humans seems unrealistic. One might expect that, in practice, a finite approximation to common knowledge would do instead, but that is not always so. Some apparently realistic forms of coordinated action can be achieved under common knowledge but not under 'almost common knowledge' (Rubinstein 1989). Moreover, even a few iterations of 'everyone knows' can be unachievable for epistemological reasons explained in KAIL, since each iteration requires a further margin for error (see also Hawthorne and Magidor 2009, 2010; for a different approach to the problem see Lederman 2018a, 2018b). Yet an announcement over a loudspeaker can surely *seem* to create common knowledge amongst the people in a room. What is going on?

The knowledge default is implicitly a *common* knowledge default. For substituting 'everyone knows that P' for 'P' in the default schema 'If P, everyone knows that P' gives 'If everyone knows that P, everyone knows that everyone knows that P', and so on, which gives arbitrarily many iterations of 'everyone knows that'. Of course, such a default does not mean that there really is common knowledge. It just means that, when nothing inhibits the default, everyone acts as if everything were common knowledge. But that may suffice for coordination to be achieved. It may even be achieved, just as it often seems, with no iteration of epistemic operators, indeed with no epistemic operators at all: the phenomenology is just that of a world open to view. Since the coordination is the predictable result of deeply rooted forms of human thinking, it may even be safe enough for those involved to know in advance that they will coordinate. Naturally, all this needs to be worked out in much more detail. But it promises to be a far more psychologically realistic picture of the cognitive processes underlying apparent common knowledge than any elaborate reconstruction in epistemic logic.

Gordon (2021) connects his arguments about the shared world and cognitive efficiency to the predictive coding model of perception; Daniel Munro (forthcoming) makes a similar connection, arguing that the predictive mind hypothesis is best developed in knowledge-first terms. Incidentally, Gordon (1969, 1987) took a knowledge-first approach to factive emotions long before KAIL.

The distinction between knowledge and ignorance is fundamental to the mindreading capacity of humans and other animals, and so to 'folk epistemology'. By itself, that does not establish that the distinction is also fundamental to epistemology of some more scientific kind, which might in principle use a very different taxonomy of epistemic states. However, the case for psychological fundamentality does hint in that direction. For it does not present the fundamentality as a mere quirk of evolutionary history. Rather, the knowledge-ignorance distinction has primacy because it is cognitively efficient to think in such terms: the distinction is more inter-personally accessible than the 'internal' alternatives, better adapted to sharing information about the world, and much less costly to encode. Those considerations would apply to a vast range of other possible finite thinkers of quite unfamiliar forms. Thus the envisaged 'scientific' alternative to knowledge-first epistemology would mostly be understanding these actual and possible thinkers in terms quite alien to those in which they most fundamentally understood themselves and others, where that understanding is itself part of the subject matter of epistemology. That still does not make the alternative scientific enterprise hopeless, but it starts off at a disadvantage.

The crunch comes when a theorist presents a specific alternative to the knowledge-first approach—for example, an informal, ungraded belief-first approach, or a formal, graded credence-first variant. One key question is how alternative the proposed alternative really succeeds in being. For the theorist is also a human animal whose default is the shared world, an implicitly knowledge-first way of thinking which can easily slip in under the theorist's radar.

For example, I argued in KAIL that subjective Bayesianism constituted a serious epistemology only by helping itself to a category of 'evidence' for the subject to update on. When such updating is described as 'learning', the mask slips and the tacit reliance on a knowledge-first way of thinking reveals itself, for learning is coming to *know*.

A similar example is the literature on the 'washing-out of priors', with theorems to the effect that, under various conditions, when different prior probabilities are successively updated on new evidence, the results converge in the limit, so that the idiosyncrasies of the priors do not matter (in the infinite limit!). The results concern cases where the priors are all updated on the *same* evidence, otherwise there is no reason for convergence. The only natural rationale for the assumption of common evidence is the picture that the evidence comprises facts open to view for all subjects: in effect, a shared world.

A more recent and bizarre case is the large debate on *disagreement*: when one finds oneself differing from one's epistemic peer who assigns a different credence to the same proposition on the same evidence, should one remain steadfast or conciliate by splitting the difference? In effect, on the standard modelling of such situations in the literature, the agents' relevant evidence is exhausted by their levels of credence in the proposition at issue and their epistemic peerhood (in some sense). Their original evidence, on which their current credence was based, is treated as having no further relevance—even when it is mathematically inconsistent with the proposition at issue or with its negation (as in the popular example of disagreement in adding up a restaurant bill). Yet someone else's degree of credence and whether they are one's epistemic peer are typically much harder to know than the fact at issue (what the total bill is). The model treats an arbitrary slice of the world—credences and

peerhood—as shared and open to view, the rest as hidden (contrast Hawthorne and Srinivasan 2013).

Many internalist epistemologies play variations on the same theme, treating just one's internal world—the bubble of consciousness—as open to one's view, making what KAIL calls one's 'cognitive home'. In addition to the anti-luminosity argument, serious work on introspection hardly supports attributing such epistemic privilege to an internal world (Carruthers 2011, Schwitzgebel 2008). Obviously, internalists do not treat what is open to one's view as a *shared* world, since it is not open to another's view, but they still seem to be relying on the cognitively efficient folk technique of modelling one's knowledge of some facts simply by modelling those facts themselves, here with a restriction to 'internal' facts.

None of those epistemological strategies succeeds in fully eliminating the folk epistemological knowledge default. Instead, they restrict it to a privileged class of facts. In the process, they lose much of its flexibility by no longer treating it as a mere default: they bake the privilege into the structure of the theory. In that way, they are more naïve than folk epistemology, because they cannot handle ignorance of the privileged facts.

In the limit, there are accounts on which *no* propositions play the role of evidence. KAIL considers such an account: an extreme form of subjective Bayesianism that permits all updates by Jeffrey conditionalization. Such views reduce epistemic normativity to purely formal probabilistic coherence. They find no epistemic fault with the wildest conspiracy theories and the most bigoted prejudices, no matter how out of touch with reality, as long as they conform to standard axioms of the probability calculus and a lax mathematical constraint on updating. For any random finite set of $n$ possible worlds, they permit one to give probability $1/n$ to each of those worlds, and probability 0 to every other world, irrespective of one's sense experience and memories. Epistemology would hardly be worth bothering with, if that were the best it could do.

The track record of attempts to start epistemological theorizing somewhere quite independent of knowledge-first folk epistemology is discouraging. It is also worth noting that folk epistemology is continually tested in practice, by our use of it to assess our own epistemic position and that of others, and to guide our inquiries. It is surely far from perfect, but it more or less works. By contrast, most epistemological theories are never applied in practice: it is not even obvious how they could be. They are tested against our pre-theoretic verdicts on a few benchmark thought experiments—but those verdicts are themselves likely to be products of folk epistemology.

None of this means that we cannot go beyond folk epistemology. For example, the theory of probability—understood not as something like *plausibility* but as mathematically constrained by the standard Kolmogorov axioms—is presumably no part of folk epistemology; that explains why it did not develop until the seventeenth century. Clearly, probability theory has been extensively applied in scientific practice and has amply proved its worth. But that does not vindicate subjective Bayesian epistemology as an alternative to folk epistemology, since its specifically subjective aspect has not been under test: the prior probability distributions in use have been selected by scientists as reasonable for the case at hand, in ways implicitly constrained by their background knowledge. Many unreasonable but probabilistically coherent priors would yield quite unreasonable results in practice. As in KAIL, evidential probability theory is an enhancement of folk epistemology, not an

alternative to it, just as microscopes and telescopes enhance sense perception rather than enabling scientists to do without it.

Evidently, the research programme on what might be called the psychological reality of knowledge-first epistemology—in both humans and other animals—is in its early stages. We have the barest outline of the picture, but most of the details remain to be filled in: just how are obstacles to knowledge tracked and registered, and just how are false beliefs finally acknowledged? Epistemologists will have much to learn from future developments in the cognitive psychology of humans and other animals, and will have something of their own to contribute to the joint inquiry. For example, it took epistemologists to recognize the significance of Gettier cases. More generally, the epistemic significance of the relations to the world that cognitive systems have the function of implementing is best understood in a setting informed by systematic epistemological theory. But epistemology can only make that contribution properly if it abandons internalist preconceptions.

The knowledge-first approach to mindreading also casts new light on the role of *charity* in interpretation. For Quine and Davidson, a key constraint on interpreting others is, as far as possible, to make what they assent to come out *true*. That applies to both their thought and their talk, and is treated as *constitutive*, not just *instrumental*: what it is for an interpretation to be correct is in part for it to be charitable in that sense. Charity is not meant to be merely an effective means to an independently defined end. In *The Philosophy of Philosophy* (Williamson 2007), I argued that Quine and Davidson's principle of charity, by maximizing *true belief*, inherited the problems of epistemology when it concentrates on true belief, a psychologically unnatural category. If an interpretation maximizes true belief by attributing many beliefs that just happen to be true, though the subject is in no position to know their truth, that should not make the interpretation correct, as I illustrated in detail. I argued that a better principle of charity maximizes *knowledge* rather than true belief.

What the arguments in *The Philosophy of Philosophy* do not fully bring out is the *centrality* of knowledge-attribution to mindreading. One might get the impression that knowledge-maximization acted merely as a tiebreaker, deciding between already given candidate interpretations. But once we understand mindreading as working down from a completely knowing subject, not up from a completely blank slate, it is clear that, without attributing knowledge, interpretation cannot even get started.

## *9. Looking forward: knowledge and action*

One aspect of KAIL never fully satisfied me: the analogy between knowledge and (intentional) action, as developed in KAIL's Introduction and recalled above. I worked there with the formula: knowledge stands to belief as action stands to desire. Belief and desire were contrasted as having opposite directions of fit. When all goes well with fitting world to mind, there is action. When all goes well with fitting mind to world, there is knowledge. However, when I tried to work out the details, things never went quite as smoothly as I had hoped. In particular, the formula presented action as more closely related to desire than to belief. But, even from the perspective of belief-desire psychology, belief and desire were more or less symmetrically related to action, equally necessary inputs to decision-making.

I gradually realized that in starting with the question 'What stands to desire as knowledge stands to belief?', I had already conceded too much to belief-desire psychology. I needed to start from a better place, taking the knowledge-action pair as given, and ask: what stands to action as belief stands to knowledge? The natural answer to that question is not desire but *intention*. Consider the global process centred on decision-making, including the origins of the input premises to practical reasoning, the practical reasoning itself, and the results of the output conclusion, and compare it to the local process of the reasoning itself, narrowly understood. When all goes well in the global process, knowledge is the input to the practical reasoning and action is the output. When something goes wrong on the input side, there may be mere belief rather than knowledge, but still playing the same local role as knowledge, of input to the reasoning. When something goes wrong on the output side, there may be mere intention rather than action, but still playing the same local role as action, of output from the reasoning. When the defect is only partial on the input side, the agent may still *reasonably* believe that P, short of knowing that P. When the defect is only partial on the output side, the agent may still *try* to do A, short of intentionally doing A. Richard Holton (2014) had already drawn a related comparison of belief and intention. One can make up one's mind *that P*; one can make up one's mind *to φ*.

Just as there was a degenerating research programme of trying to analyse 'S knows that P' in terms of 'S believes that P', '(it is true that) P', and other factors not presupposing the category of knowledge, so there was a degenerating research programme of trying to analyse 'S intentionally φs' in terms of 'S intends to φ', 'S φs', and other factors not presupposing the category of intentional action. The two research programmes faced analogous obstacles, in counterexamples to necessity or sufficiency and in implicit circularity. For instance, just as a causal link from its being true that P to S's believing that P is not the missing ingredient, nor is a causal link from S's intending to φ to S's φing. Both proposals face the problem of deviant causal chains and other difficulties (Williamson 2017a, 2018).

Just as the distinction between S's knowing that P and S's not knowing that P is easier to track by observing S than the distinction between S's believing that P and S's not believing that P (see the previous section), so the distinction between S's intentionally φing and S's not intentionally φing is easier to track by observing S than the distinction between S's intending to φ and S's not intending to φ. Just as we should expect attributions of knowledge to precede attributions of belief, so we should expect attributions of action to precede attributions of intention.

This reworking of the analogy between knowledge and action raises an obvious question: where does desire fit into the new picture? Since desire belongs on the input side, which, when all goes well, is the knowledge side, the natural answer to the question is that desire is a form of belief. For belief is what constitutes knowledge when all goes well. Thus one arrives at the ultra-controversial view of desire as belief.

Of course, if desires consist in beliefs, it does not follow that the desire that P consists in the belief *that P*. Rather, the desire that P consists in the belief that Φ(P), for some suitable operator Φ. For instance, the desire that P might consist in the belief that it would be good if P, more exactly, in the belief that if P it would be good that P, in a not specifically moral sense of 'good' ('if P' is needed because 'it is good that P' implies 'P').

Such unpacking already defuses the objection that desires cannot be beliefs because inconsistent beliefs are always irrational while inconsistent desires are sometimes rational. For example, suppose that two of your friends, Mary and John, are among the hundreds of candidates for the same job. Quite rationally, you may both want Mary to get the job and want John to get the job, even though you know that they cannot both get it. For the first desire may consist in the belief that if John got the job, it would be good that he got it, while the second consists in the belief that if Mary got the job, it would be good that she got it. Those two beliefs may be true together, for both outcomes may be good: perhaps Mary and John are equally good candidates, and better than all the rest. You desire one outcome and desire another, knowing that the two outcomes are mutually incompatible, but what you believe about the first outcome and what you believe about the second are mutually compatible. If all desires consist in beliefs, and it is irrational to believe incompatibles, it just does not follow that it is irrational to desire incompatibles, because *what* you believe when you desire is not what you desire. Of course, you still have *an* attitude—desire—to incompatible propositions, but that is not in itself irrational. When I wonder whether a pear is ripe, I also wonder whether it is *not* ripe, without thereby contradicting myself. When the glass is half-full, I rationally take the attitude of rejection to both the proposition that it is empty and the proposition that it is full, despite their mutual incompatibility.

In the setting of knowledge-first epistemology, assimilating desire to belief involves subjecting desire to a knowledge norm. On the view developed in KAIL, if you believe that P without knowing that P, your believing is defective. Thus, if you desire that P, and your desire consists in the belief that it would be good if P, but you do not know that it would be good if P, then your desiring is defective. Some philosophers will deny that desire is subject to any such norm. They may contrast belief and desire in just that respect: belief is subject to an epistemic norm, but desire is not. That view often goes with the claim that desire has no function: desire is simply a matter of individual preference. From an evolutionary perspective, that claim is quite implausible. Desires serve an obvious evolutionary function: to motivate creatures to get things that are good for them, either individually or collectively: food, drink, sex, warmth, safety, …. Iris, the flowering plant, is poisonous for sheep, but they often want to eat it when they see it. In a natural sense, their desire for it is *mistaken*. In effect, their desire consists in a false belief that it would be good to eat. Similarly, if they want to eat a plant which in fact has the same appearance as iris but is good for them, in an area with an abundance of iris, their desire for it is not mistaken but is still epistemically defective. They believe that it would be good to eat, and their belief is true, but fails to constitute knowledge.

Using the word ‘good’ to articulate a sheep’s propositional attitudes may sound like over-intellectualization. But it is not strange to characterize dumb animals as treating some options as *better* than others, and ‘better’ is just the comparative of ‘good’. What such general evaluative terms express here need not be an intrinsically motivating quality, but simply a common measure for weighing different goods (food, drink, sex, warmth, safety, ….) against each other, which even dumb animals have to do—and sometimes get wrong.

Presumably, for evolutionary reasons, animals normally are motivated to pursue their own individual or collective good, at least under evolutionary normal conditions, but not out of any metaphysical or conceptual necessity. As we know from the human case, one can

believe and even know that it would be good to eat (in a sense of 'good' suitably related to evolutionary fitness), yet still not desire to eat. Thus when the desire to eat does consist in a belief that it would be good to eat, which motivates one to eat, that is no automatic consequence of believing that it would be good to eat, but an effect dependent on contingent circumstances. If a belief that Φ(P) *can* constitute a desire that P, it does not follow that a belief that Φ(P) *must* constitute a desire that P. The belief that doing X would annoy their parents motivates some teenagers to do X, other teenagers to avoid doing X, and leaves still others indifferent. In short, desires play the general functional role of belief—one acts on them—but they also play a more specific functional role of their own, by helping characterize the end state of action in practical terms (this more specific functional role may help explain how children manage to attribute desires—'I want it!'—comparatively early).

If desire has its own functional role, what is gained by assimilating it to belief? One advantage emerges when we try to understand the agent's practical reasoning from a first-personal perspective. We seek an argument in the first person present tense that, as far as possible, favours the action taken, attempted, or at least intended. Even for unreflective agents incapable of articulating their reasons themselves, such an argument should capture something of what they were up to from their perspective. For the argument to favour a specific course of action, it will need *premises*. But they must be premises that the agent in effect assumes or endorses; in brief, the agent should *believe* the premises. Since the role of desire in practical reasoning is to provide some of the inputs, it must in effect contribute believed premises, in which the desires consist. Thus assimilating desire to belief helps us understand the agent's practical reasoning from the agent's point of view.

By applying the knowledge norm of belief to agents' beliefs in the premises of their practical reasoning, we can then assess the epistemic standing of their starting-point. This is close to a connection between knowledge and action for which John Hawthorne and Jason Stanley have argued independently (Hawthorne and Stanley 2009).

Perhaps the 'practical reasoning' is sometimes as simple as 'Playing loud music will annoy my parents, so I'll play loud music', with no mediating conception of the wider good to be served by annoying the parents. The recalcitrant teenager may even *know* the truth of the premise. In the unlikely event that there is really no more to it than that, we may just have to settle on the verdict that the premise is in an epistemically fine position, but the argument is a *non sequitur*.

The challenge to articulate the agent's reason for action in first-personal terms poses a significant difficulty for standard decision theory, which works on the agent's credences (subjective probabilities) and preferences (subjective utilities)—the graded analogues of beliefs and desires. The problem may not be obvious at first sight, since the decision theorist can present the usual calculation of the subjectively expected utilities of the various options as a formal representation of the agent's implicit practical reasoning. The trouble is that, in doing so, they represent the agent as reasoning *about* her own mental states.

For example, when a mother searches for her baby, the decision-theoretic calculation has a premise like 'The mother greatly prefers her baby being found to her baby being lost'. Thus, when put into the first person, it represents the mother as implicitly reasoning from a premise like 'I greatly prefer my baby being found to my baby being lost'. But that is quite implausible, even though there is no suggestion that the mother says such words to herself.

For the mother's thoughts are much less likely to be on her own preferences than on her baby's needs. The decision-theoretic representation misrepresents the mother as self-absorbedly reasoning as though it were all about her.

Decision theorists can of course represent the mother's altruism by giving more weight to the baby's welfare than to the mother's in the latter's ranking of possibilities. That is not in doubt. The point is rather that, when the decision-theoretic calculation is treated as something the mother could in principle endorse as giving her reason for action, her reason is represented as consisting in facts about her own subjective psychological states. She is depicted as pathologically self-regarding, of treating the reasons for action as all about her. That gets her psychology hopelessly wrong.

The point applies just as much to the agent's credences as to her preferences. The decision-theoretic calculation has premises such as 'The mother has a much higher credence in her baby being upstairs than in her baby being downstairs'. Thus, when put into the first person, it represents the mother as implicitly reasoning from a premise like 'I have a much higher credence in my baby being upstairs than in my baby being downstairs'. But that too is quite implausible, even though there is no suggestion that the mother says such words to herself. For the mother's thoughts are much less likely to be on her own doxastic states than on where her baby is objectively likely to be. She is trying to be true to the world, not true to herself. Here too the decision-theoretic representation depicts the mother as self-absorbedly reasoning as though it were all about her.

The problem is exacerbated when, as often, decision theory is proposed as a normative theory about the decision-making of ideally rational agents. For, if anyone can articulate their reason for action in the first person, it is an ideally rational agent. But, as just observed, an agent whose reason for action is articulated in the first person as the subjective decision-theoretic calculation is thereby revealed to be pathologically self-regarding, which would be an unfortunate consequence of ideal rationality.

None of this is to deny the value of calculating the expected utility of various options when appropriate. For that value does not depend on interpreting the probabilities and utilities as merely subjective. One can rationally estimate the genuine costs and benefits of different outcomes, and their probabilities on one's evidence, and calculate the expected utility of various options accordingly. But the premises of the calculation will be true or false, depending on how the world is; they will not be a mere expression of one's own subjective psychological state.

The 'shared world' approach to mindreading helps with attributing desires as beliefs. For example, when something is good to eat or drink, the default is for that to be out in the shared world, observable to others too. As usual, the default can be overridden in many ways. When a task is hard (for me), I must not automatically treat it as hard (for you); likewise, when something is good (for me), I must not automatically treat it as good (for you). Some adjustments are needed for mindreading other members of the same species, since we may be competing for scarce resources; larger adjustments are needed for mindreading members of a species on the other side of a predator-prey distinction. Nevertheless, such adjustments are a fair price to pay for not having to start by treating others as blank slates with respect to desire.

The analogy between knowledge and action can take one a long way. I will briefly sketch some further connections between knowledge-first epistemology and the philosophy of action.

When I was completing KAIL, I was also working on a paper with Jason Stanley, arguing for the *intellectualist* view that knowing-how is a special case of knowing-that, thereby rejecting the cliché that knowing that and knowing how are mutually exclusive, one theoretical and one practical (as though theory and practice were mutually exclusive). Our article 'Knowing How' appeared the following year (Stanley and Williamson 2001; see also Stanley 2011). Roughly, to know how to φ is to know, of some way *w*, that *w* is a way for one to φ. At least in paradigmatic cases of knowing how, one has the knowledge under a *practical mode of presentation* of *w*, under which *w* is ready for one to implement. That was to be expected from obvious analogies between 'how' (in what way?) and comparable interrogative words such as 'why' and 'wherefore' (for what reason?), 'when' (at what time?), 'where' (in what place?), 'whence' (from what place?), 'whither' (to what place?) 'who' and 'whom' (what person?), 'whose' (of what person?), 'whether' (yes or no?), 'which' (of given alternatives?), and 'what' itself. If 'how' had been spelt and pronounced 'whow', there might have been less fuss.

Considerations of semantic compositionality, combined with the semantics of indirect questions, strongly favour intellectualism: at least the literal reading of 'S knows how to φ' is intellectualist. Of course, 'S knows how to φ' might also have an idiomatic meaning too, but then it would be ambiguous, and so should pass standard tests for ambiguity, a point most anti-intellectualists ignore.

Critics have noted that in some languages one asks the equivalent of 'Can she swim?' when in English one might ask 'Does she know how to swim?', but that hardly shows the questions to be synonymous. After all, it is usually a matter of indifference whether one asks 'Can she speak English?' or 'Does she speak English?', even though the questions are not semantically equivalent—perhaps she can speak English but refuses to do so.

Although both KAIL and 'Knowing How' are centrally about knowledge, the two projects were pursued independently. There was no reason to fear that they might lead to inconsistent conclusions. Gradually, more connections between them have emerged, not least through the work of Carlotta Pavese. In an important series of papers, she has developed a knowledge-first conception of action as radically informed and controlled by knowledge (Pavese 2015, 2016, 2017, 2019, 2020a, 2020b, 2021, and her chapter in the present volume). Stanley and I have argued for a related view of *skill* (contrast *strength*) as a disposition to have the knowledge required for controlling action (Stanley and Williamson 2017). Thus action is not merely the *analogue* of knowledge on the output side; it is itself deeply knowledgeable. More recently, I have explored how desire as belief, a knowledge norm for belief, an action norm for intention, and intellectualism about knowing how combine to transform a traditional belief-desire account of means-end reasoning step by step into a knowledge-first account of the role of knowing how in action (Williamson 2023a).

Perhaps this unification of knowledge and action can be taken still further. In particular, *intentionally bringing it about* that P might be understood as a kind of *active knowing* that it will be that P (maker's knowledge). That also suggests an assimilation of *intending to bring it about* that P to a kind of *believing* that it will be that P.

Some deny that ‘S intentionally brings it about that P’ is sufficient for ‘S knows that it will be the case that P’, on the grounds that luck can play too large a role for knowing without playing too large a role for intentionally bringing it about, though it is unclear whether such claims are robust. Even if there is no entailment, perhaps all central cases of intentionally bringing it about that P are cases of knowing that it will be that P, which would already be a significant result. In this area, as in most others, much remains to be understood.

### *10. Looking forward: models of knowledge*

KAIL makes extensive use of a model-building methodology, in the tradition of epistemic logic. Even very simple models can cast light on epistemic structure. For instance, in assessing sceptical arguments, we can understand the non-symmetry of the accessibility relation by considering a model with just two worlds, corresponding to the good case and the bad case (the sceptical scenario).

By itself, the framework of Kripke models is neutral towards the knowledge-first approach. We can do epistemic logic, requiring accessibility to be reflexive (since knowledge entails truth), but equally we can do doxastic logic, not requiring accessibility to be reflexive (since belief does not entail truth)—though we might still require it to be *serial*, so every world sees at least one world (if inconsistent belief is excluded). However, when one surveys applications of the framework in computer science and economics, one typically finds the models explicitly described in terms of knowledge, not of mere belief. Usually, the models are multi-agent, with one accessibility relation per agent, which is required to be an equivalence relation, so it induces a partition of the worlds in the model into mutually exclusive, jointly exhaustive cells. The agent cannot discriminate between any worlds in the same cell, but can discriminate between any worlds in different cells. For each agent, accessibility is reflexive, since any world is in the same cell as itself. Thus the models are genuinely epistemic rather than doxastic. This is surely not the result of any prior theoretical commitment to a knowledge-first approach on the part of computer scientists or economists. They are just using the simplest non-trivial multi-agent models within the overall framework, which strikingly turn out to be of knowledge rather than just belief.

Partitional models also endow agents with perfect introspection, both positive and negative: they always know whether they know a given proposition. That is often a defensible idealization: in studying communicative obstacles to inter-agent epistemic transparency (common knowledge), it makes sense to stipulate away obstacles to intra-agent epistemic transparency, in order to isolate the phenomena of interest. Unfortunately, that idealization has tended to harden into an ideological dogma, resulting in a kind of self-imposed epistemological *naïveté* in non-philosophical formal epistemologists, who have never properly confronted the sceptical consequences of their assumptions, if treated as more than temporarily useful model-building idealizations.

The arguments in KAIL against positive introspection, the KK principle, were inspired by devising models of agents with limited powers of discrimination. Subsequently, I extended the arguments to reach stronger conclusions. For example, imagine someone with normal but not perfect vision looking at an unmarked clock. On that basis, how much can she

know about where the hand is pointing? If one models the situation, including a probability distribution, in the simplest, most natural way, one can prove that for some proposition *p* in the model (as it were, the proposition that it is pointing between 3:05 and 3:15), the agent knows *p*, even though it is almost certain on her evidence that she does *not* know *p*. Many variations can be played on this theme (Williamson 2011, 2014). Thus KK does not just fail in the model; it fails as drastically as it could.

Of course, a single type of model is never conclusive by itself. With enough *ad hoc* gerrymandering, one can almost always remodel a given phenomenon so as to avoid the feature one dislikes. That applies in particular to the phenomenon of the agent's knowledge of the unmarked clock, and the feature of KK failure. In response, rather than simply pointing out *ad hoc* aspects of particular alternative models, it is more satisfying to prove that *any* model of the given phenomenon which lacks the controversial feature will have a specified *ad hoc* feature. In the case of the unmarked clock, that can be done. For the relevant space of possibilities has natural rotational symmetry, induced by the rotational symmetry of the clock-face itself, without the hand; in any given possibility, the hand breaks the symmetry, because it has a specific position, but there is still symmetry in the overall range of prior possibilities. One can prove that in any epistemic model of the case with rotational symmetry, if the agent learns *something* but not *everything* about the hand's position by looking at the clock, then KK fails for some proposition (Williamson 2021a). Thus to reconcile the case with KK without being hopelessly unrealistic, the model must be *ad hoc* by breaking the symmetry. In that sense, the counterexample to KK is *robust*.

The model-building methodology can play a further role in making epistemological conclusions more robust. For it is not far-fetched to worry that the standard procedure of analytic epistemology relies on a kind of *naïve falsificationism*. Universally general theories in epistemology make predictions about hypothetical scenarios; by doing the relevant thought experiment, we either verify or falsify the prediction. If it is falsified, so is the epistemological theory, since we have a counterexample to the generalization. Thus the refuted theory can be dismissed. One trouble with the naïve falsificationist methodology in natural science is that scientists are not infallible in designing, conducting, and interpreting experiments in real life. For example, they may neglect an interfering factor. That is why experiments need to be repeatable, preferably by different people, under different conditions, by a different method. Not every apparent falsification is genuine. In analytic epistemology, universal generalizations are often treated as refuted by a single thought experiment. The danger is that we dismiss a true theory on the basis of a mistaken verdict on one thought experiment. The mistake will not be merely idiosyncratic, since the epistemological community dismisses a theory only when it reaches consensus on the theory. But the negative verdict in the thought experiment might result from our reliance on a natural but imperfectly reliable *heuristic*, such as an easy, quick and dirty way of judging whether knowledge is present in a given case. Sometimes, we have no pre-theoretic way to second-guess our heuristic; what indicates its merely heuristic status is that its outputs can be mutually inconsistent (for philosophers' and linguists' reliance on such heuristics see Williamson 2020a, 2021d). An output of such a heuristic may well secure inter-personal assent, even when false.

This worry about reliance on thought experiments may remind readers of the 'negative program' in experimental philosophy, initiated by the work of Jonathan Weinberg, Shaun Nichols, and Stephen Stich (2001). Some early results seemed to indicate that received verdicts on standard thought experiments in analytic epistemology vary with subjects' ethnicity or gender, and so are unreliable. A salient case in point was the verdict that the protagonist of a Gettier case lacks knowledge. Proponents of the negative program drew the moral that philosophers should stop relying on 'philosophical intuitions', though they never satisfactorily explained what counts as a 'philosophical intuition' (there is nothing peculiarly philosophical about the judgment that someone doesn't know something). Might the old justified true belief analysis of knowledge be right after all? However, most of the early results proved misleading: they were not repeated when experiments were redone more carefully. In particular, the negative verdict on Gettier cases turned out to be more like a cross-cultural human universal, prompting experimental philosophers to postulate a universal 'folk epistemology' (Machery et al. 2017). I have criticized the negative program elsewhere, and will not repeat those arguments here (Williamson 2021c; see also Nagel 2012). Still, it did highlight the risks involved in treating a single thought experiment as a conclusive counterexample (see Alexander and Weinberg 2014 on 'error fragility'). Indeed, even if the negative verdict on Gettier cases is a human universal, that does not automatically mean that it is *true*. For it may be the output of a humanly universal fallible heuristic. Brian Weatherson has suggested that the word 'know' could refer to justified true belief by virtue of the latter's being the most 'natural' property to approximately fit the use of 'know', despite our reaction to Gettier cases (Weatherson 2003).

Since real-life cases are just as relevant to epistemology as are counterfactual thought experiments, I will write simply of 'examples', to cover both.

The point is not that judgments on examples are *especially* fallible, but simply that they *are* fallible. The same applies to perceptual judgments. We cannot do science if we never rely on sense perception, but we still have to control for perceptual error. Similarly, we have to control for error in judgments on examples. This is where model-building can help. We can use it as an independent test of conclusions reached through examples. For instance, one can model the JTB analysis of knowledge as justified true belief, to explore its consequences in a formal setting. Contrary to the impression of it as a beautiful theory tragically and unjustly slain by counterexamples, one can show that it reduces the known to awkward disjunctions, where one disjunct secures truth while the other secures justified belief, and that it has various consequences inimical to the views of those who treat justification as epistemologically fundamental (Williamson 2013a, 2013b, 2015). Naturalness works *against* the JTB analysis, not for it. Similarly, one can assess the slogan 'evidence of evidence is evidence' by exploring its unfortunate consequences on various interpretations in a formal setting (Williamson 2019).

In general, we can make our conclusions more robust by testing them with different methodologies. The idea is not that formal models supersede examples, any more than examples supersede formal models. It is just that hypotheses confirmed by *both* examples *and* formal models are better confirmed than those confirmed in only one of the two ways.

Obviously, much work in epistemology is resolutely informal, and its authors might well reject the challenge to model it formally as inappropriate. In doing so, however, they

give up the chance to test and develop their hypotheses in one of the most rigorous and fruitful ways available. Nor is it clear *why* the challenge of formal modelling is supposed to be inappropriate. After all, the task is not to formalize the epistemological theory itself, but just to show exactly how it works in a mathematically precise, simple but not quite trivial case. If theorists cannot manage even that much, one may suspect that something is amiss with their theory.

Formal model-building in epistemology is of course not confined to the knowledge-first approach. It has a much longer history in the Bayesian tradition. A welcome recent development is the increasing interaction between formal and informal epistemology: when pursued in mutual isolation, both sides suffer. In particular, subjective Bayesian epistemology confined itself to a purely formal standard of rationality, on which a consistent Nazi whose credences satisfy the Kolmogorov axioms and who updates only by conditionalizing on Nazi-vetted propositions counts as perfectly rational. Epistemology can do better than that. One place to start is with the observation that the Nazi grossly violates the knowledge norm of belief. One advantage of the knowledge-first approach is that it formally models both knowledge and belief.

### *11. Looking forward: epistemic norms*

Epistemic justification is often treated, more or less definitionally, as what it takes to comply with the basic epistemic norm for belief, whatever that is. On a knowledge-first view, the norm is knowledge. Thus a belief is justified if and only if it constitutes knowledge (I leave tacit the gloss 'epistemically' on 'justified'). Consequently, only true beliefs are justified.

Alas, another constraint is often treated as more or less definitional too: that beliefs are equally justified in corresponding good and bad cases. Thus, since my belief that I have hands is justified in the actual good case, it is also justified in a corresponding bad case, a sceptical scenario where I lack hands. Consequently, not only true beliefs are justified.

To treat *both* identity with the basic epistemic norm for belief *and* equality across the good and bad cases as simultaneously definitional is cheating. For it sneaks in the contentious internalist assumption that compliance with the basic epistemic norm for belief is equal across the good and bad cases, as if it were a matter of stipulation (Williamson forthcoming[a]).

KAIL treats justification as *graded*, and measured by probability on the evidence. Thus a false belief may be justified to some degree. Still, only knowledge is *fully* justified. I became more forthright on that point after KAIL was published (Williamson forthcoming[a]; see also Srinivasan 2020 on externalism about justification).

Of course, a question remains: *why* are so many epistemologists so tempted to judge that beliefs are as justified in the bad case as in the good case? A natural answer is that beliefs in the two cases are envisaged as resulting from the very same cognitive dispositions. In general, dispositions are typically more durable than their manifestations. If you are put in circumstances where it is harder for you to exercise your skills effectively, it does not follow that you have become less skilful. Were we to suspect that the agent changed cognitive dispositions in being switched between the good and bad cases, we might well be much less

tempted to judge that the level of justification remained constant over the switch. But we must not confuse judging the believer with judging the belief.

An analogy: Someone borrows a book, promising to return it. But the book gets stolen, so she cannot return it. She did her best, but she failed to keep her promise. So she violated the norm of promise-keeping on that occasion (promising to φ is not merely promising to do your best to φ). However, she does have an excellent *excuse* for having violated the norm of promise-keeping. Similarly, imagine that it is illegal for a commoner to touch the royal sceptre. You are a law-abiding commoner. A troublemaker throws the sceptre in your face, making you touch it. You did your best, but you broke the law. However, you do have an excellent *excuse* for having broken the law. Likewise, although the brain in a vat does its best, it violates the knowledge norm of belief, and even a truth norm, by believing that it has hands. However, it does have an excellent *excuse* for violating the knowledge norm of belief.

We cannot hope to do justice to such examples if we approach them with a radically impoverished normative vocabulary, consisting only of 'justified' and 'unjustified'. We need at least the category of excuses for cases of fully or partially blameless norm-violation. When I have applied it, I have sometimes been interpreted as doing so on the basis of an *analysis* of excusability into necessary and sufficient conditions. That is a misinterpretation. We need the category of excuses to handle the gap between the inflexible simplicity of many norms and the unpredictable, messy, human complexity of real-life cases. Any attempt to codify excusability in terms of necessary and sufficient conditions misses exactly that point.

Still, we can identify some salient types of excuse. Given a primary norm N on acts, there is a secondary norm on agents: to be disposed to comply with N. Thus the brain in a vat's beliefs violate the knowledge norm, but the brain may still comply with the corresponding secondary norm, by being disposed to form beliefs that constitute knowledge (if we treat its envatment as abnormal), even though its attempts to exercise that disposition misfire in its unfortunate circumstances. There is even a tertiary norm, on acts again: to act as someone disposed to comply with N would. That an act complies with the tertiary norm is typically a good excuse for its failure to comply with the primary norm.

Maria Lasonen-Aarnio (2010, 2014, 2021) has used the distinction between a primary 'occurrent' norm and a corresponding secondary 'dispositional' norm in a converse way, to argue against the alleged phenomenon of knowledge defeat by misleading counter-evidence. Ignoring counter-evidence is a bad cognitive disposition; its normal effect is to block opportunities for agents to learn from their mistakes. However, agents occasionally benefit from their bad habits, and achieve undeserved success. Those who ignore counter-evidence are sometimes lucky enough to be retaining genuine knowledge. They comply with the primary norm of belief but violate the secondary norm—the opposite of the brain in the vat.

Unfortunately, epistemologists often fail to distinguish between primary and tertiary norms of belief. That failure is especially prevalent in discussions of rational belief, partly because we apply the term 'rational' to both acts and agents (Williamson 2017b). For generality, we may suppose just that the primary rational norm (on beliefs) is to *accord with* one's evidence, where 'accord with' is schematic. The knowledge norm is one way of implementing that schema, but not the only one. Thus the secondary norm (on believers) is to be disposed to form beliefs that accord with one's evidence; that is what it is to be a rational

believer. The tertiary norm (on beliefs) is to believe what a rational believer would believe in the circumstances. Epistemologists often confuse the primary and tertiary norms, by treating 'Does it accord with one's evidence?' and 'Would a rational believer believe it?' as interchangeable tests for the rationality of a belief. But the primary and tertiary norms are equivalent only if the beliefs that accord with one's evidence are exactly the beliefs a rational believer would have. That equivalence is far from trivial. For example, if the primary norm is implemented by the knowledge norm, then, as typically envisaged, the brain in a vat's beliefs violate the primary norm (they are not knowledge) but comply with the tertiary norm (they are what a rational believer in the relevant sense would believe in the circumstances); thus the equivalence fails.

More generally, suppose that there can be *illusions of evidence*, where something seems to be evidence but isn't, or seems not to be evidence but is. Then the disposition to form beliefs that accord with one's evidence can misfire, because it generates beliefs that accord with one's *apparent* evidence but not with one's *real* evidence. Such beliefs violate the primary norm but comply with the tertiary norm. Consequently, epistemologists who treat the primary and tertiary norms of rational belief as interchangeable in effect assume that there are no illusions of evidence, in other words, that evidence is *transparent*.

Similarly, if either the presence or the absence of evidence is non-luminous in the sense of KAIL, one would not expect the primary and tertiary norms of rational belief to be equivalent. Thus epistemologists who treat the two norms as interchangeable in effect assume that both the presence and the absence of evidence are luminous.

Given the anti-luminosity argument and similar considerations, the primary and tertiary norms of rational belief are *not* equivalent (see Srinivasan 2013 for a reply to the objections in Berker 2008 to the anti-luminosity argument). Thus conflating the two norms amounts to assuming a false internalist principle.

These considerations are quite general. For any non-trivial norm, complying with it, being disposed to comply with it, and doing as someone disposed to comply with it would do are pairwise non-equivalent. Moreover, as in KAIL, complying with it, being in a position to know that one is complying with it, and not being in a position to know that one is not complying with it, are pairwise non-equivalent too. Thus any non-trivial norm is surrounded by a cloud of closely related but non-equivalent norms, each of which seems to have its own distinctive force. For example, although the dispositional norm derives from the occurrent norm, we may be more concerned with agents' dispositions to infringe than with their actual infringements. Similarly, we may be more concerned with agents' knowledge of their standing with respect to the primary norm than with that standing itself. *None* of these norms provides the agent with fully operational guidance (Williamson 2008, Srinivasan 2015, Hughes 2018).

In such circumstances, when we start out not knowing which norm is primary, the normative landscape is very hard to map. Undifferentiated pre-theoretical reactions to individual cases are a recipe for confusion, since one has no way of holding the operative norm constant, especially when the judgments are supposed to be 'all things considered'. We are pulled in opposite directions by variant norms, which can create something like epistemic dilemmas (Hughes 2019, forthcoming[a], forthcoming[c]).

We do better to proceed abductively rather than inductively, first conjecturing a single primary norm—the simpler and more perspicuous the better—and then trying to explain the relevant normative phenomena on that basis, invoking its derivative norms as necessary.

For belief, truth and knowledge are the two most salient candidates for the primary norm (KAIL argues that a knowledge norm explains more than a truth norm). By contrast, if one starts only with a norm of internal coherence or reflective equilibrium, it is quite unclear how to explain what is wrong with the belief system of a Nazi who consistently sees the world in Nazi terms (Williamson forthcoming[b]). Internalist attempts to refute the possibility of such a person are far from convincing. As we saw, subjective Bayesianism faces the same problem.

To make the explanatory starting-point maximally perspicuous, we should formulate the primary norm simply in terms of a condition for compliance, not in terms of a condition for permissibility. The reason is that, semantically, deontic modals such as 'permissible' depend on a domain of contextually relevant possible worlds; variations in that parameter are extraneous to the underlying norm but easily cause confusion. For example, they have led to many needless complications in the debate on a truth norm for belief (Williamson 2020b).

What *kind* of norm is an epistemic norm for belief? Much of the literature focuses on reflective, responsible, rational persons—perhaps professional philosophers as they might like to be. Moralizing talk of epistemic 'virtues' and 'vices' encourages that impression. Once we appreciate how widespread knowledge and belief are amongst non-human animals, we should be wary of such over-intellectualization.

A broadly functional starting-point is more promising. But we should not confine ourselves to norms on individual beliefs, for one could in principle comply with such norms by just suspending all belief (if that were psychologically possible). Instead, we should consider norms on whole cognitive systems (Williamson forthcoming[c]). After all, why does an animal have a cognitive system in the first place? The natural, evolutionarily plausible answer is: to provide it with knowledge to act on. That function is not served by suspending belief. Nor is it served by non-knowledgeable beliefs; they are *defective*: something went wrong. Only knowledge will do.

There is no obvious internalist alternative to providing knowledge as the function of a cognitive system. Why bother with such a system if its goal is merely to achieve internal coherence or reflective equilibrium? Although Mona Simion (2019) tries to locate a non-factive standard for justified belief on a knowledge-first functionalist approach, the term 'justification' seems out of place in such a setting.

What function would be analogous to that of providing knowledge for a cognitive system with probabilistic credences rather than outright beliefs? Sarah Moss (2018) proposes that such an analogue is available if the contents of attitudes are reconfigured in probabilistic terms, but it is not clear that such radical complications are well-motivated (Williamson forthcoming[d]).

An evolutionary perspective is useful, because it reminds us what a cognitive system is *for*. If it tempts us into an over-specific biological reductionism, we can correct that tendency by considering the generality to which epistemology aspires, over many kinds of cognition and many kinds of cognizer. At that level, the category of knowledge is as natural as the category of nutrition.

Acknowledgments

Thanks for very helpful comments on drafts of this chapter to Robert Gordon, Daniel Kodsi, Harvey Lederman, Jennifer Nagel, participants in a class at Yale given with Jason Stanley, the editors of this volume, and an anonymous referee for it.

References

Alexander, J., and Weinberg, J. (2014): 'The "unreliability" of epistemic intuitions', in E. Machery and E. O'Neill (eds.), *Current Controversies in Experimental Philosophy,* (London: Routledge): 128-145.

Anscombe, G.E.M. (1957): *Intention* (Oxford: Blackwell).

Austin, J. L. (1956-7): 'A plea for excuses', *Proceedings of the Aristotelian Society* 57: 1-30.

Austin, J. L. (1962): *Sense and Sensibilia*, ed. G. J. Warnock. (Oxford: Oxford University Press).

Bennett, J. (1990): 'Why is belief involuntary?', *Analysis* 50: 87-107.

Berker, S. (2008): 'Luminosity regained', *Philosophers' Imprint* 8/2: 1-22.

Berlin, I. (1973): 'Austin and the early beginnings of Oxford philosophy' in I. Berlin and others (eds.), *Essays on J. L. Austin,* (Oxford: Clarendon Press): 1-16.

Bird, A. (2010): 'Social knowing: the social sense of "scientific knowledge"', *Philosophical Perspectives* 24: 23-56.

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

Blome-Tillmann, M. (2017): '"More likely than not"—knowledge first and the role of statistical evidence in courts of law', in Carter, Gordon, and Jarvis (2017): 278-292.

Burge, T. (1979): 'Individualism and the mental', *Midwest Studies in Philosophy* 4: 73-121.

Byrne, A., and Logue, H. (2009): *Disjunctivism: Contemporary Readings* (Cambridge, Mass.: MIT Press).

Carruthers, P. (2011): *The Opacity of Mind: An Integrative Theory of Self-Knowledge* (Oxford: Oxford University Press).

Carter, J.A., Kelp, C., and Simion, M. (forthcoming): 'On behalf of knowledge-first collective epistemology', in P. Silva and L. Oliveira (eds.), *Doxastic and Propositional Warrant* (London: Routledge).

Carter, J.A., Gordon, E.C., and Jarvis, B.W. (eds.), (2017): *Knowledge First: Approaches in Epistemology and Mind* (Oxford: Oxford University Press).

Dutant, J. (2015): 'The legend of the justified true belief analysis', *Philosophical Perspectives* 1: 95-145.

Edgington, D. (1985): 'The paradox of knowability', *Mind* 94: 557-568.

Edgington, D. (2010): 'Possible knowledge of unknown truths', *Synthese* 173: 41-52.

Evans, G. (1982): *The Varieties of Reference*, ed. J. McDowell (Oxford: Clarendon Press).

Fitch, F. (1963): 'A logical analysis of some value concepts', *Journal of Symbolic Logic* 28: 135-142.

Gordon, R. (1969): 'Emotions and knowledge', *Journal of Philosophy* 66: 408-413.

Gordon, R. (1987): *The Structure of Emotions* (Cambridge: Cambridge University Press).

Gordon, R. (2000): 'Sellars's Ryleans revisited', *Protosociology* 14: 102-114.

Gordon, R. (2021): 'Simulation, predictive coding, and the shared world', in M. Gilead and K. Ochsner (eds.), *The Neural Basis of Mentalizing* (Cham: Springer).

Greenough, P., and Pritchard, D. (eds.) (2009): *Williamson on Knowledge* (Oxford: Oxford University Press).

Hart, W. (1979): 'The epistemology of abstract objects: access and inference', *Proceedings of*

*the Aristotelian Society* sup. 53: 152-165.
Hart, W., and McGinn, C. (1976): 'Knowledge and necessity', *Journal of Philosophical Logic* 5: 205-208.
Hawthorne, J., and Magidor, O. (2009): 'Assertion, context, and epistemic accessibility', *Mind* 118: 377-397.
Hawthorne, J., and Magidor, O. (2010): 'Assertion and epistemic opacity', *Mind* 119: 1087-1105.
Hawthorne, J., and Magidor, O. (2018): 'Reflections on the ideology of reasons', in Daniel Star (ed.), *The Oxford Handbook of Reasons and Normativity* (Oxford: Oxford University Press): 113-140.
Hawthorne, J., and Srinivasan, A. (2013): 'Disagreement without transparency: some bleak thoughts', in D. Christensen and J. Lackey (eds.), *The Epistemology of Disagreement: New Essays* (Oxford: Oxford University Press): 9-30.
Hawthorne, J., and Stanley, J. (2008): 'Knowledge and action', *Journal of Philosophy* 105: 571-590.
Hintikka, J. (1962): *Knowledge and Belief* (Ithaca, NY: Cornell University Press).
Hinton, J. M. (1967): 'Visual experiences', *Mind* 76: 217-227.
Hinton, J. M. (1973): *Experiences* (Oxford: Oxford University Press).
Holton, R. (2014): 'Intention as a model for belief', in M. Vargas and G. Yaffe (eds.), *Rational and Social Agency: Essays on the Philosophy of Michael Bratman* (Oxford: Oxford University Press): 12-37.
Horschler, D., Santos, L., and MacLean, E. (2019): 'Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis', *Cognition* 190: 72 -80.
Hughes, N. (2018): 'Luminosity failure, normative guidance and the principle "ought -implies-can"', *Utilitas* 30: 439-457.
Hughes, N. (2019): 'Dilemmic epistemology', *Synthese* 196: 4059-4090.
Hughes, N. (forthcoming[a]): 'Epistemology without guidance'. *Philosophical Studies*.
Hughes, N. (ed.) (forthcoming[b]): *Epistemic Dilemmas* (Oxford: Oxford University Press).
Hyman, J. (2015): *Action, Knowledge, and Will* (Oxford: Oxford University Press).
Kelp, C., and Simion, M. (2021): *Sharing Knowledge: A Functionalist Account of Assertion* (Cambridge: Cambridge University Press).
Lasonen-Aarnio, M. (2010): 'Unreasonable knowledge', *Philosophical Perspectives* 24: 1-21.
Lasonen-Aarnio, M. (2014): 'Higher-order evidence and the limits of defeat', *Philosophy and Phenomenological Research* 88: 314-345.
Lasonen-Aarnio, M. (2021): 'Dispositional evaluations and defeat', in J. Brown and M. Simion (eds.), *Reasons, Justification, and Defeat* (Oxford: Oxford University Press): 93-115.
Lederman, H. (2018a): 'Uncommon knowledge', *Mind* 127: 1069-1105.
Lederman, H. (2018b): 'Two paradoxes of common knowledge: coordinated attack and electronic mail', *Noûs* 52: 921-945.
Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., and Hashimoto, T. (2017): 'Gettier across cultures', *Noûs* 51: 645-664.
Marion, M. (2000): 'Oxford realism: knowledge and perception', parts I and II, *British*

*Journal for the History of Philosophy* 8: 299-338 and 485-519.
Martin, M. (2004): 'The limits of self-awareness', *Philosophical Studies* 120: 37-89.
McDowell, J. (1977): 'On the sense and reference of a proper name', *Mind* 86: 159-185.
McDowell, J. (1982): 'Criteria, defeasibility, and knowledge', *Proceedings of the British Academy* 68: 455-479.
McDowell, J. (1995): 'Knowledge and the internal', *Philosophy and Phenomenological Research* 55: 877-893.
Merleau-Ponty, M. (1945): *Phénoménologie de la perception*. (Paris: Gallimard.)
Moss, S. (2018): *Probabilistic Knowledge*. Oxford University Press.
Munro, D. (forthcoming): 'Perceiving as knowing in the predictive mind', *Philosophical Studies*.
Nagel, J. (2012): 'Intuitions and experiments: a defense of the case method in epistemology', *Philosophy and Phenomenological Research* 85: 495-527.
Nagel, J. (2013): 'Knowledge as a mental state', *Oxford Studies in Epistemology* 4: 275-310.
Nagel, J. (2014): *Knowledge: A Very Short Introduction* (Oxford: Oxford University Press).
Nagel, J. (2017): 'Factive and nonfactive mental state attribution', *Mind and Language* 32: 525-544.
Nagel, J. (forthcoming): *Recognizing Knowledge: Intuitive and Reflective Epistemology*.
Pavese, C. (2015): 'Practical senses', *Philosophers' Imprint* 15: 1-25.
Pavese, C. (2016): 'Skill in epistemology', Parts I and II, *Philosophy Compass* 11: 642-660.
Pavese, C. (2017): 'Know-how and gradability', *Philosophical Review* 126: 345-383.
Pavese, C. (2019): 'The psychological reality of practical representation', *Philosophical Psychology* 32: 785-822.
Pavese, C. (2020a): 'Practical representation', in Fridland and Pavese (2020): 226-244.
Pavese, C. (2020b): 'Probabilistic knowledge in action', *Analysis* 80: 342-356.
Pavese, C. (2021): 'Knowledge, action, and defeasibility', in J. Brown and M. Simion (eds.), *Reasons, Justification, and Defeaters* (Oxford: Oxford University Press).
Perner, J. (1993): *Understanding the Representational Mind* (Cambridge, Mass.: MIT Press).
Phillips, J., Buckwalter, L., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., and Knobe, J. (2020): 'Knowledge before belief', *Behavioral and Brain Sciences* 44: e140.
Phillips Griffiths, A. (ed.) (1967): *Knowledge and Belief* (Oxford: Oxford University Press).
Prichard, H. (1950): *Knowledge and Perception* (Oxford: Oxford University Press).
Putnam, H. (1973): 'Meaning and reference', *Journal of Philosophy* 70: 699-711.
Rubinstein, A. (1989): 'The electronic mail game: strategic behavior under "almost common knowledge"', *American Economic Review* 79: 385-391.
Salerno, J. (2009): 'Knowability noir 1945-1963', in J. Salerno (ed.), *New Essays on the Knowability Paradox* (Oxford: Oxford University Press): 29-48.
Schwitzgebel, E. (2008): 'The unreliability of naïve introspection', *Philosophical Review* 117: 245-273.
Sellars, W. (1975): 'Autobiographical reflections (February 1973)', in H.-N. Castañeda (ed.), *Action, Knowledge, and Reality: Studies in Honor of Wilfrid Sellars*, (Indianapolis: Bobbs-Merrill): 277-293.
Shin, H., and Williamson, T. (1994): 'Representing the knowledge of Turing machines',

*Theory and Decision* 37: 125-146.
Shin, H., and Williamson, T. (1996): 'How much common belief is necessary for a convention?', *Games and Economic Behavior* 13: 252-268.
Simion, M. (2019): 'Knowledge-first functionalism', *Philosophical Issues* 29: 254-267.
Slote, M. (1979): 'Assertion and belief', in J. Dancy (ed.), *Papers on Language and Logic* (Keele: Keele University Library): 177-190
Snowdon, P. (1980-1): 'Perception, vision and causation', *Proceedings of the Aristotelian Society* 81: 175-192.
Snowdon, P. (1990): 'The objects of perceptual experience', *Proceedings of the Aristotelian Society* sup. 64: 121-150.
Srinivasan, A. (2013): 'Are We Luminous?', *Philosophy and Phenomenological Research* 90: 294–319.
Srinivasan, A. (2015): 'Normativity without Cartesian Privilege', *Philosophical Issues* 25: 273–299.
Srinivasan, A. (2020): 'Radical externalism', *Philosophical Review* 129: 395-431.
Stanley, J. (2011): *Know How*. Oxford: Oxford University Press.
Stanley, J., and Williamson, T. (2001): 'Knowing how', *The Journal of Philosophy* 98: 411-444.
Stanley, J., and Williamson, T. (2017): 'Skill', *Noûs* 51: 713-726.
Vaidya, A. (2022): 'Elements of knowledge-first epistemology in Gaṅgeśa', *Oxford Studies in Epistemology*, forthcoming.
Weatherson, B. (2003): 'What good are counterexamples?', *Philosophical Studies* 115: 1-31.
Weinberg, J., Nichols, S., and Stich, S. (2001): 'Normativity and epistemic intuitions', *Philosophical Topics* 29: 429-460.
Williamson, T. (1982): 'Intuitionism disproved?', *Analysis* 42: 203-207.
Williamson, T. (1986): 'Criteria of identity and the Axiom of Choice', *The Journal of Philosophy* 83: 380-394.
Williamson, T. (1987a): 'On the paradox of knowability', *Mind* 96: 256-261.
Williamson, T. (1987b): 'On knowledge and the unknowable', *Analysis* 47: 154-158.
Williamson, T. (1990): *Identity and Discrimination* (Oxford: Blackwell). 2nd edition 2013.
Williamson, T. (1992): 'Inexact knowledge', *Mind* 101: 217-242.
Williamson, T. (1993): 'Verificationism and non-distributive knowledge', *Australasian Journal of Philosophy* 71: 78-86.
Williamson, T. (1994): *Vagueness* (London: Routledge).
Williamson, T. (1995): 'Is knowing a state of mind?', *Mind* 104: 533-565.
Williamson, T. (1996a): 'Knowing and asserting', *Philosophical Review* 105: 489-523.
Williamson, T. (1996b): 'Cognitive homelessness', *Journal of Philosophy* 93: 554-573.
Williamson, T. (1997): 'Knowledge as evidence', *Mind* 106: 717-741.
Williamson, T. (1998): 'Conditionalizing on knowledge', *British Journal for the Philosophy of Science* 49: 89-121.
Williamson, T. (2000): *Knowledge and its Limits* (Oxford: Oxford University Press).
Williamson, T. (2007): *The Philosophy of Philosophy* (Oxford: Wiley-Blackwell).
Williamson, T. (2008): 'Why epistemology can't be operationalized', in Q. Smith (ed.), *Epistemology: New Essays* (Oxford: Oxford University Press): 277-300.

Williamson, T. (2009): ‘Probability and danger’, *The Amherst Lecture in Philosophy* 4: 1-35.
Williamson, T. (2011): ‘Improbable knowing’, in T. Dougherty (ed.), *Evidentialism and its Discontents* (Oxford: Oxford University Press): 147-164.
Williamson, T. (2013a): ‘Gettier cases in epistemic logic’ *Inquiry* 56: 1-14.
Williamson, T. (2013b): ‘Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier Cases in Epistemic Logic’, *Inquiry* 56: 77-96.
Williamson, T. (2014): ‘Very improbable knowing’, *Erkenntnis* 79: 971-999.
Williamson, T. (2015): ‘A note on Gettier cases in epistemic logic’, *Philosophical Studies* 172: 129-140.
Williamson, T. (2017a): ‘Acting on knowledge’, in Carter, Gordon, and Jarvis (2017): 163 -181.
Williamson, T. (2017b): ‘Ambiguous rationality’, *Episteme* 14: 263-274.
Williamson, T. (2018): ‘Knowledge, action, and the factive turn’, in V. Mitova (ed.), *The Factive Turn in Epistemology* (Cambridge: Cambridge University Press): 125-141.
Williamson, T. (2019): ‘Evidence of evidence in epistemic logic’, in M. Skipper and A. Steglich-Petersen (eds.), *Higher-Order Evidence: New Essays* (Oxford: Oxford University Press): 265-297.
Williamson, T. (2020a): *Suppose and Tell: The Semantics and Pragmatics of Conditionals* (Oxford: Oxford University Press).
Williamson, T. (2020b): ‘Non-modal normativity and norms of belief’, *Acta Philosophica Fennica* 90: 101-125.
Williamson, T. (2021a): ‘The KK principle and rotational symmetry’, *Analytic Philosophy* 62: 107-124.
Williamson, T. (2021b): ‘Edgington on possible knowledge of unknown truth’, in L. Walters and J. Hawthorne (eds.), *Conditionals, Paradox, and Probability: Themes from the Philosophy of Dorothy Edgington* (Oxford: Oxford University Press): 195-211.
Williamson, T. (2021c): *The Philosophy of Philosophy*, enlarged ed. Oxford: Wiley -Blackwell.
Williamson, T. (2021d): ‘Epistemological consequences of Frege puzzles’, *Philosophical Topics* 49: 287-319.
Williamson, T. (2023a): ‘Acting on knowledge-how’, *Synthese* 200.
Williamson, T. (2023b): ‘Vaidya on Nyāya and Knowledge-First Epistemology’, *Oxford Studies in Epistemology* 7: 365-375.
Williamson, T. (forthcoming[a]): ‘Justifications, excuses, and skeptical scenarios’, in J. Dutant and F. Dorsch (eds.), *The New Evil Demon: New Essays on Knowledge, Justification and Rationality* (Oxford: Oxford University Press).
Williamson, T. (forthcoming[b]): ‘Boghossian, Müller-Lyer, the parrot, and the Nazi’, in L. Oliveira (ed.), *Externalism about Knowledge* (Oxford: Oxford University Press).
Williamson, T. (forthcoming[c]): ‘Epistemological ambivalence’, in Hughes forthcoming[d].
Williamson, T. (forthcoming[d]): ‘Knowledge, credence, and strength of belief’, in A.K. Flowerree and B. Reed (eds.), *Expansive Epistemology: Norms, Action, and the Social World* (London: Routledge).
Williamson, T. (forthcoming[e]): ‘Dunn on inferential evidence’, *The Monist*.
Williamson, T. (forthcoming[f]): ‘E = K, but what about R?’, in M. Lasonen-Aarnio and C.

Littlejohn (eds.), *Routledge Handbook of Evidence* (London: Routledge).
Wilson, J.C. (1926): *Statement and Inference*, 2 vols, ed. A.S.L. Farquharson (Oxford: Clarendon Press).
Wilson, J.C. (1967): 'The relation of knowing to thinking', in Phillips Griffiths (1967): 16-27. Reprint of Wilson (1926), vol. I: 34-47.